ICML 2012 Handbook

International Conference on Machine Learning

June 26 - July 1, 2012

Edinburgh, Scotland, UK

Program at a Glance

Day	COLT	ICML	Time	Venue
Mon, June 25	COLT Sessions		daytime	AT, IF
Tue, June 26	COLT Sessions (page 167)	ICML Tutorials (page 14)	daytime	AT, IF
	ICML Reception		evening	SNG
Wed, June 27	ICML/COLT Joint Sessions (pages 25/169)		daytime	AT, IF
	ICML/COLT Poster Session (pages 31/35)		evening	AT, IF
Thu, June 28		ICML Sessions Day 2 (page 40)	daytime	AT, IF
		ICML Poster Session 2 (page 49)	evening	AT, IF
Fri, June 29		ICML Sessions Day 3 (page 57)	daytime	AT, IF
		ICML Poster Session 3 (page 66)	evening	AT, IF
Sat, June 30		ICML Workshops Day 1 (page 170)	daytime	AT, IF, DHT
		Banquet	evening	NMS
Sun, July 1		ICML Workshops Day 2 (page 170)	daytime	AT, IF, DHT

AT: Appleton Tower
DHT: David Hume Tower
IF: Informatics Forum
NMS: National Museum of Scotland
SNG: Scottish National Gallery

See the end of the book and the back cover for venue maps.

Sponsors

Platinum Sponsor

National Science Foundation

Silver sponsors

eBay Research Labs

GHCQ – Government Communications Headquarters

Google

PASCAL

SICSA – The Scottish Informatics and Computer Science Alliance

Winton Capital

Supporting Organizations

IBM Research

Best paper awards

Machine Learning Journal

Artifical Intelligence Journal

Computing Community Consortium

Contents

1 Welcome to Edinburgh

We are pleased to welcome you to the 29th International Conference on Machine Learning. This is the second time that Scotland has hosted ICML, the first time being the 9th conference in Aberdeen in 1992.

Edinburgh has a distinguished intellectual history, including David Hume, James Clerk Maxwell, and (briefly) Charles Darwin. Perhaps especially interesting to ICML attendees will be the Rev Thomas Bayes, who studied theology at the University of Edinburgh from 1719–1722. It is likely that Bayes studied mathematics here as well, although we do not have direct evidence for this.

If you have spare time, probably best is simply to walk around the city, enjoying the medieval street plan and the 18th and 19th century architecture. A few of the more popular attractions are the Royal Mile, the New Town, Edinburgh Castle, and Arthur's Seat. If the weather is good, a large and beautiful park called the Meadows is immediately south of the conference area. Immediately south of the Meadows is Bruntsfield Links, a free public golf park.

The city boasts many good restaurants, including six with Michelin stars. Those with more humble tastes can consider the traditional Scottish delicacies of haggis, deep fried Mars bars, whisky, and IRN BRU, although perhaps not in one sitting. It is also possible to find excellent coffee and cask-conditioned ale.

However, you may find that you have little time to leave the conference floor, as by many measures this will be the largest ICML ever. This year the conference features **242 papers** selected from **890 submissions**, a 50% increase over previous years. The workshop programme has been significantly expanded, featuring **20 workshops** over 2 days. Additionally the conference is colocated with the Conference on Computational Learning Theory, and a joint ICML/COLT poster session will be held on Wednesday night.

We hope that you enjoy your visit to Scotland.

Charles Sutton, University of Edinburgh
on behalf of the ICML 2012 local arrangements committee

2 Information

2.1 Practical Information

Even if you have attended the conference many times, a few things are different this year. The next few pages explain some of the major things to keep in mind.

Spotlight Talks

This year we have 91 more accepted papers than last year. To fit them all, you will notice that some papers have full length talks (20 minutes) while other papers have spotlight talks (5 minutes). Please be aware of this if you wish to switch sessions midway through.

Coffee Breaks

All coffee breaks will be held jointly between Appleton Tower, the building in which you registered, and the Informatics Forum, the large building directly across the street. If Appleton Tower gets crowded, simply make your way across the street. Tea, coffee, and biscuits will be available in both buildings.

Café Area

There is a breakout area with tables, chairs, and white boards for informal meetings and discussions. This is at the rear of the Informatics Forum on the ground floor.

Poster Sessions

Poster sessions will be held on the evenings of Wednesday, Thursday, and Friday, immediately after the last talk ends. All papers will have an associated poster, on the same day as their talk. The posters will be split between the Informatics Forum and Appleton Tower. See the floor plans later in this book to help you find particular posters.

At registration you received drink tickets to use during the poster sessions, and there will also be food and a cash bar.

Internet Access

You can obtain guest access to the University of Edinburgh's wireless network `central` at the Wifi desk in Appleton Tower, near the catering stations.

However, the University also participates in the `eduroam` worldwide roaming access service. You can use this if your home institution also participates. We recommend that you use this if you can.

COLT

The COLT conference overlaps in time with ICML. The two days of overlap are Tuesday, the tutorials day, and Wednesday, the first day of ICML sessions. All ICML delegates are welcome to attend the COLT sessions during the two days of overlap.

Workshops

This year there will be two days of workshops, on Sat June 30 and Sun July 1. You were asked to select workshops when you registered online, but this choice is not binding. If you have paid the fee for the workshops, you may attend any workshops that you like.

Banquet

The conference banquet will be held on Sat June 30, during the workshop days. Your workshop fee included one ticket to the banquet. To attend the banquet, you must either (a) have registered for the workshops or (b) have purchased a separate ticket to the banquet.

Certificates of Attendance

If you require a certificate confirming your attendance at the conference, please ask at the registration desk.

Twitter

Hashtag is `#ICML2012` Official account is `@IcmlConference`

Conference Book

If you prefer to read the conference abstracts electronically, there is a PDF version of this book available at `http://icml.cc/`

2.2 Local Information

2.2.1 Restaurants

We've tried to give a rough idea of the price of a main course with the following scheme, but we can't promise to pay the difference if you end up paying more!

£	£7 and under
££	£7–£10
£££	£10–£17
££££	more than that

Cheap lunch places around the Informatics Forum

Nile Valley
6 Chapel Street, tel. 0131 667 8200

Takeaway. African wraps, a lot of vegetarian options. In the evening the place turns into a restaurant serving good value delicious African food (BYOB). Seating available inside and downstairs groups of 10 people or probably more could be accommodated. *Price: £-££*

Mother's Kitchen
77 Nicolson Street, tel. N/A

Probably one of the cheapest lunch options around the forum. Huge portions, nice food, low prices and friendly service. Take-away or sit-in. Tables can be put together for larger groups, 10 people or so could fit reasonably easily. *Price: £*

Mosque Kitchen (Old)
West Nicholson Street, tel. N/A

This used to be an Edinburgh institution, cheap curry served directly from the Mosque's kitchen, in the Mosque yard, with pigeons all around. Now with some refurbishments it lost some of its grotty charm, the food is still decent, although more expensive than it used to be... Entrance either through the Mosque courtyard on Potterrow, or a little gate on W Nicolson St. *Price: £*

Mosque Kitchen (New)
Nicholson Square, tel. N/A

A newly opened indoor premises for the Mosque kitchen, same great curry, reasonable prices, huge indoor seating area, suitable for very large groups. *Price: £*

Kebab Mahal
7 Nicolson Square, tel. 0131 667 5214

Excellent Indian/Pakistani food in an extremely unpretentious setting and very reasonably priced. Groups of more than six will have trouble. No alcohol (and no BYOB). *Price: £*

Hong Kong Yum Yum
13 West Richmond Street, tel. 0131 667 8263

Great Chinese food. A place is tiny, so unlikely you will find space for more than 4, also does great take-away, duck soup is especially recommended. *Price: £*

Picnic Basket
West Nicolson Street, tel. N/A

Takeaway. A selection of freshly made sandwiches and baked potatoes. Every day they do a daily special meal for about £3, usually it rice and chicken curry, chilli con carne or potato stovies. *Price: £*

Beetle Juice
33 West Nicolson Street, tel. 0131 668 3824

Takeaway. Home made soups and sandwiches. Very limited seating inside (2 or 3 people). *Price: £*

Saborsito
Chapel Street, tel. N/A

Mexican food (mainly burrito wraps with rice, beans, chicken/beef/vegetables). Takeaway. Groups up to 6 could be seated inside, but there is only one large table, however it is rarely occupied as most people get take-away. *Price: £*

Double Dutch
27-29 Marshall Street, tel. 0131 667 9997

Very nice middle eastern food, decent portions, lots of tables, good for larger groups (10 people or so). *Price: £-££*

Elephant Juice
George sq., opposite Hugh Robson build., tel. N/A

A man in a van selling wonderful home made soups and freshly baked bread to take-away, a bit pricey for soup but absolutely delicious. (Nothing to do with other Elephant named places...) *Price: £*

Elephants and Bagels
37 Marshall Street, tel. 0131 668 4404

A huge selection of bagels. Sit-in or take-away, suitable for groups up to 4-6, but not more. *Price: £*

Red Fort
10 Drummond Street, tel. 0131 557 1999

All you can eat Indian buffet (around £7 for students, £9 non-students). Nice and plentiful food, tables could be put together to accommodate groups of up to 10-12 people. There is almost always space for a big group, so no need to book. *Price: £-£ £*

56 North
2-8 West Crosscauseway, tel. 0131 662 8860

Huge south facing windows catch the sun whenever it is out. A selection of good quality European foods from burgers to healthier options, all very good, but portions are not huge. On a sunny day there is outside seating. Tables for groups up to 6-8 people are usually available. *Price: £-£ £*

Union of Genius
8 Forrest Rd., tel. N/A

Soup only. Good and close to the George square area. *Price: £*

Edinburgh Larder
Blackfriar's St, tel. N/A

A more foodie take on sandwiches and picknicking. Close to Royal Mile. *Price: £*

Kim's Korean meals
5 Buccleuch Street, tel. 0131 629 7951

Wonderful and plentiful authentic Korean food. Mains around £7 at lunchtime, £10 in the evening. A big group could fit, although the place is quite small, so better to phone ahead to make sure there is space for everyone. *Price: £-£ £*

Yocoko Noodle Bar
44-46 South Bridge, tel. 0131 558 3889

Decent food comes in decent portions here. Mainly a mish mash of Chinese food with influences from other oriental countries. Mains around £5-6. Even though the space is quite big, not so suitable for groups, although 6-7 people could be squeezed in. *Price: £*

Newington Traditional Fish Bar
23 South Clerk Street, tel. 0131 667 0203

In the running for Edinburgh's best fish and chips shop. Seating space as well as take-away. *Price: £*

2.2.2 Cafés

Peter's Yard
Quartermile, 27 Simpson Loan

Swedish café/bakery on the Meadows with excellent pastries and great coffee although on the expensive side.
Open 7:00–18:00, 9:00–18:00 weekends.

Kilimanjaro
104 Nicolson Street

Great coffee and sandwiches.
Open 7:45–20:30, 8:30–20:00 weekends.

Press Coffee
30 Buccleuch Street

An offshoot of Kilimanjaro, usually less busy.
Open 8:00–18:30, Sat 9:00–17:00, Sun 11:00–17:00.

Black Medicine Coffee Company
2 Nicolson Street

Smoothies and sandwiches as well as coffee.
Open 8:00–18:00, 10:00–18:00 Sundays.

Cafe Turquaz
119 Nicolson Street

0131 667 66641

Lovely Turkish meze, great Turkish coffee, quite a small places, not really suitable for groups.

The Elephant House
George IV road bridge

One of the few cafés that stay open late. A huge selection of teas and great cakes. Food is also served, however, while not bad, it is not great and a bit overpriced. Nice atmosphere and huge tables, that are often shared with other customers, but at a less busy time a large group (6-8) could easily fit.

Restaurants Around the Informatics Forum and points south and west

Namaste Kathmandu
17-19 Forrest Road, tel. 0131 220 2273

Nepalese and Indian cuisine. Possible seating for large groups. *Price: £-££*

Blonde
75 St Leonard's Street, tel. 0131 668 2917

A classier place in the Southside. *Price: £ £ £*

Home Bistro
41 West Nicholson Street, tel. 0131 667 7010

Fairly small place that does classics well. *Price: £ £*

Kalpna
2-3 St Patrick Square, EH8 9EZ, tel. 0131 667 9890

South Indian whole-food vegetarian *Price: £ £*

Ann Purna
45 St Patrick's Square, EH8 9ET, tel. 0131 662 1807

Similar concept to the Kalpna, with different menu options *Price: £ £*

Mother India
3-5 Infirmary Street, EH1 1LT, tel. 0131 524 9801

First rate Indian "tapas". Big tables so good for groups *Price: £ £*

Spoon
6a Nicolson Street, tel. 0131 557 4567

Upstairs from the Black Medicine Coffee Company, this bistro does 'quirky comfort food' with a seasonal focus. Large space: good for groups. *Price: £ £*

Mezbaan
14 Brougham Street, tel. 0131 229 5578

South Indian in Tollcross (west of the Informatics Forum), down the street from Cloisters pub. Large selection of fish and vegetarian dishes. *Price: £ £*

Tanjore
8 Clerk Street, tel. 0131 478 6518

South Indian food, significantly different than the typical British curry. Spacious, unpretentious atmosphere and excellent dosa. *Price: £ £*

Old Town - Restaurants

Creelers
3 Hunter Square, tel. 0131 220 4447

Seafood restaurant with an emphasis on local provenance. *Price: £ £ £*

Ondine
2 George IV Bridge, tel. 0131 226 1888

Highly rated seafood restaurant, the first in Edinburgh to be accredited by the Marine Stewardship Council. *Price: £££*

Wedgewood
267 Canongate, tel. 0131 558 8737

Imaginative Scottish food with a focus on local/seasonal ingredients. *Price: ££££*

Monteith's
61 High Street, EH1 1SR, tel. 0131 557 0330

A more traditional take on Scottish food, highly rated. *Price: ££££*

David Bann
56 St Mary's Street, tel. 0131 556 5888

Excellent modern vegetarian restaurant. *Price: £££*

La Garrigue
31 Jeffrey St, tel. 0131 557 3032

Very well-regarded Languedoc restaurant.
Another branch at *14 Eyre Place, (0131 558 1608).* *Price: £££*

The Outsider
15-16 George IV Bridge, tel. 0131 226 3131

A popular restaurant with great views of the castle out the back. Inventive food at surprisingly reasonable prices; reservations are recommended. *Price: ££*

Petit Paris
38-40 Grassmarket, tel. 0131 226 2442

Unpretentious French little place, offering good value for money and friendly service. Mains around £15 *Price: £££*

New Town - Restaurants

The Dogs
110 Hanover Street, EH2 1DR, tel. 0131 220 1208

Michelin Bib Gourmand-rated for its excellent value. Excellent Scottish food with a no-fuss attitude to fine dining. It is surrounded by its offspring, also all highly rated: Amore Dogs (Italian, if you go there try the perfect semi-fredo desert), Seadogs (seafood, seafood barley risotto is superb). Reservations recommended. *Price: ££-£££*

Mussel Inn
61-65 Rose Street, tel. 0843 2892 481

The place to go for mussels. They have (non-shell)fish too, but don't press your luck much beyond that. *Price: ££*

The Dining Room
28 Queen Street, EH2 1JX, tel. 0131 557 3500

Downstairs from the Scottish Malt Whisky Society; excellent up-scale dinning and an unparalleled whisky selection. *Price: ££££*

Fishers in the City
58 Thistle Street, EH2 1EN, tel. 0131 225 5109

The city-centre branch of a Leith seafood restaurant that offers a number of vegetarian and steak options as well. Quieter area in the back. *Price: £££*

Edinburgh's Five Michelin Stars

Booking is **highly** recommended for all of these restaurants. Expect to pay £35/person for lunch and £75/person for dinner, excluding wine.

The Kitchin
78 Commercial Quay, Leith, tel. 0131 555 1755

The Kitchin offers an outstanding selection of French-inspired Scottish cuisine. Famed for rolled Pig's Head and renowned for seasonal game and excellent seafood. *Price: £££££*

21212
3 Royal Terrace, tel. 0845 222 1212

Revel in a choice of two starters, a soup, two main courses, a cheese course, and two desserts. Expect dishes that are assembled with up to fifteen different ingredients. Watch a team of chefs prepare each dish through a glass wall into the kitchen. *Price: £££££*

Martin Wishart's
54 The Shore, Leith, tel. 0131 553 3557

Exquisite food on the banks of the Water of Leith at The Shore. A sheer delight from the amuse-bouche to the cheese cart. *Price: £££££*

Castle Terrace
33/35 Castle Terrace, tel. 0131 229 1222

The only lunch option within walking distance of the conference offering French-inspired Scottish cuisine. Expect robust seasonal game alongside a delightful array of desserts. *Price: £££££*

Number One
The Balmoral, 1 Princes Street, tel. 0131 557 6727

Serves only dinner in the luxurious surroundings of The Balmoral hotel. Expect superb seafood and a wonderful wine list. *Price: £ £ £ £ £*

2.2.3 Pubs

Near the Forum

The Peartree
36 West Nicolson Street

Close to the Informatics Forum, one of the few pubs that has seating outside (albeit in the shadow of Appleton Tower).

Doctors
32 Forrest Road

Decent pub very near the Informatics Forum, with a number of guest ales on tap.

Sandy Bells
25 Forrest Road

Famous for its traditional folk music performances most evenings. It is quite small and can get quite cramped. Could be difficult to find even standing space if you come as a large group.

The Cloisters
26 Brougham Street

Probably the best selection of real ales in town; also excellent pub food served until 8:30 pm Tuesday – Thursday as well as lunch every day apart from Monday.

The Brass Monkey
14 Drummond Street

A popular local bar, quite dingy but cosy at the same time. It has a large carpeted room with cushions for sitting, essentially, on the floor. However, this area is often closed as it can hired for functions, and sometimes it has film-screenings on.

Teviot Row House - Library Bar
13 Bristo Square

One of the cheapest bars in town (part of the student union). Has some ales on tap and serves food till late (mainly burgers and nachos, which are fine, however it is not recommended to venture beyond these options).

The Southsider
3-7 West Richmond Street

A cheap local bar with lots of seating space.

Under the stairs
3A Merchant Street

A wonderfully cosy cocktail bar, lot's of comfortable couches and a huge menu of fantastic cocktails.

Auld Hoose
23-25 Saint Leonard's Street

A short walk from the Informatics Forum, one of the few places that has cloudy cider on tap. Also a selection of guest ales, nice atmosphere, large seating area.

Dagda
93 Buccleuch Street

Small, traditional pub with a good selection of Scottish beer. Good for conversations.

Meadow Bar
42-44 Buccleuch Street

More student vibe but spacious upstairs and some good cask ales. Reasonably good burgers.

Old Town

Halfway House
24 Fleshmarket Close

Another tiny pub precariously located on steps leading down to Waverly Station. Lots of ales, whiskies, and charm.

BrewDog
143–154 Cowgate

BrewDog is run by the indie brewers responsible for stuffing high percentage beer into taxidermied rodents. They offer a large selection of their own beers as well as micro-brewery imports. Not far from Royal Mile.

Holyrood 9A
9A Holyrood Road

Bar/restaurant boasting a selection of 20 local beers on tap, an eclectic collection of gourmet burgers and lovely interior, a bit more on the trendy side.

Bow Bar
80 West Bow/Victoria Street

A wee pub in the heart of the old town. Good selection of real ales and whiskies.

New Town and beyond

Cafe Royal and The Guilford Arms
19 West Register Street and 1 West Register Street

Despite the street numbers, these two pubs are just next to each other, hidden off the eastern end of Prince's Street. Cafe Royal is slightly more up-scale but has better food; The Guilford Arms has a larger selection of real ales.

The Cumberland Arms
1–3 Cumberland Street

The Cumberland Arms has leafy outdoor seating and a warren of rooms inside, and a good selection of real ales.

Kay's
39 Jamaica Street

Adorable wee pub with real ales and whiskies.

3 ICML Tutorials, June 26

COLT Sessions will also be held in Appleton Tower on this day. See page 167 for details.

Statistical Learning Theory in Reinforcement Learning and Approximate Dynamic Programming
09:00 - 11:30. Venue: AT LT4.
Presenters: Alessandro Lazaric and Mohammad Ghavamzadeh.

Probabilistic Topic Models
09:00 - 11:30. Venue: AT LT5.
Presenters: David Blei.

Performance Evaluation for Learning Algorithms: Techniques, Application and Issues
09:00 - 11:30. Venue: AT LT1.
Presenters: Nathalie Japkowicz and Mohak Shah.

Mirror Descent and Saddle Point First Order Algorithms Invited COLT Tutorial
12:30 - 13:30. Venue: AT LT4.
Presenters: Arkadi Nemirovski.

Representation Learning
12:30 - 15:00. Venue: AT LT5.
Presenters: Yoshua Bengio.

Prediction, Belief, and Markets
12:30 - 15:00. Venue: AT LT1.
Presenters: Jennifer Wortman Vaughan and Jacob Abernethy.

PAC-Bayesian Analysis in Supervised, Unsupervised, and Reinforcement Learning
15:30 - 18:00. Venue: AT LT1.
Presenters: Yevgeny Seldin, François Laviolette and John Shawe-Taylor.

Causal Inference - Conditional Independency and Beyond
15:30 - 18:00. Venue: AT LT4.
Presenters: Dominik Janzing and Bernhard Schölkopf.

Spectral Approaches to Learning Latent Variable Models
15:30 - 18:00. Venue: AT LT5.
Presenters: Geoffrey J. Gordon, Le Song and Byron Boots.

Statistical Learning Theory in Reinforcement Learning and Approximate Dynamic Programming

09:00 - 11:30. Venue: AT LT4.

Presenters: Alessandro Lazaric and Mohammad Ghavamzadeh (INRIA Lille - Team SequeL, France)

Tuesday, 26th June 2012

Approximate dynamic programming (ADP) and reinforcement learning (RL) algorithms are used to solve sequential decision-making problems in which the environment (i.e., the dynamics and the rewards) is not completely known and/or the size of the state and action spaces is too large. In these scenarios, the convergence and performance guarantees of the standard DP algorithms are no longer valid, and different theoretical analysis have to be developed. Statistical learning theory (SAT LT) has been a fundamental theoretical tool to explain the interaction between the process generating the samples and the hypothesis space used by learning algorithms, and has shown when and how-well classification and regression problems can be solved. In recent years, SAT LT tools have been used to study the performance of batch versions of RL and ADP algorithms with the objective of deriving finite-sample bounds on the performance loss (w.r.t. the optimal policy) of the policy learned by these methods. Such an objective requires to effectively combine SAT LT tools with the ADP algorithms, and to show how the error at each iteration is propagated through the iterations of these iterative algorithms.

The main objective of this tutorial is to strengthen the links between ADP and SAT LT. More specifically, our goal is two-fold: 1) to raise the awareness of the RL and ADP communities with regard to the potential benefits of the theoretical analysis of their algorithms for the advancement of these fields, and 2) to introduce the potential theoretical problems in sequential decision-making to the researchers in the field of SAT LT. An introduction to ADP methods will be given, tools from SAT LT needed in the analysis of these methods will be introduced, followed by an overview of the existing results. The properties of these results and how they may help us in tuning the algorithms will be discussed, the results will be compared with those in supervised learning, and the potential open problems in this area will be highlighted.

Probabilistic Topic Models

09:00 - 11:30. Venue: AT LT5.

Presenters: David Blei (Princeton University)

Tuesday, 26th June 2012

Probabilistic topic modeling provides a suite of tools for analyzing large collections of documents. Topic modeling algorithms can uncover the underlying themes of a collection and decompose its documents according to those themes. This analysis can be used for corpus exploration, document search, and a variety of prediction problems.

This tutorial will describe three aspects of topic modeling: (1) Topic modeling assumptions (2) Algorithms for computing with topic models (3) Applications of topic models

In (1), I will describe latent Dirichlet allocation (LDA), which is one of the simplest topic models, and then describe a variety of ways that we can build on it. These include dynamic topic models, correlated topic models, supervised topic models, author-topic models, bursty topic models, Bayesian nonparametric topic models, and others. I will also discuss some of the fundamental statistical ideas that are used in building topic models, such as distributions on the simplex, hierarchical Bayesian modeling, and models of mixed-membership.

In (2), I will review how we compute with topic models. I will describe approximate posterior inference for directed graphical models using both sampling and variational inference, and I will discuss the practical issues and pitfalls in developing these algorithms for topic models. Finally, I will describe some of our most recent work on building algorithms that can scale to millions of documents and documents arriving in a stream.

In (3), I will discuss applications of topic models. These include applications to images, music, social networks, and other data in which we hope to uncover hidden patterns. I will describe some of our recent work on adapting topic modeling algorithms to collaborative filtering, legislative modeling, and bibliometrics without citations.

Finally, I will discuss some future directions and open research problems in topic modeling.

Performance Evaluation for Learning Algorithms: Techniques, Application and Issues

09:00 - 11:30. Venue: AT LT1.

Presenters: Nathalie Japkowicz (University of Ottawa, Canada) and Mohak Shah (Accenture, Chicago)

Tuesday, 26th June 2012

Machine learning has matured as a field with many sophisticated learning approaches being imported to practical applications. Due to its inherent interdisciplinary nature the field draws researchers from varying backgrounds. It is of critical importance in such a case that the researchers are aware of both the proper methodologies and the respective issues that arise in terms of evaluating novel learning approaches. This tutorial aims at educating as well as getting the machine-learning community to discuss these critical issues in the performance evaluation of learning algorithms. The tutorial will focus on various aspects of the evaluation process with a focus on classification algorithms and highlight the various choice decision vis-à-vis the issues, assumptions and constraints involved in doing so.

The tutorial will span four areas of classifier evaluation: - Performance Measures (evaluation metrics and graphical methods) - Error Estimation/Re-sampling Techniques - Statistical Significance Testing - Issues in data Set Selection and evaluation benchmarks design

It will also briefly cover R and WEKA tools that can be utilized to apply them.

The purpose of the tutorial is to promote an appreciation of the need for rigorous and objective evaluation and an understanding of the available alternatives along with their assumptions, constraints and context of application. Machine learning researchers and practitioners alike will all benefit from the contents of the tutorial, which discusses the need for sound evaluation strategies, practical approaches and tools for evaluation, going well beyond those described in existing machine learning and data mining textbooks, so far.

Mirror Descent and Saddle Point First Order Algorithms

Invited COLT Tutorial. Venue: Presenters: Arkadi Nemirovski (Georgia Institute of Technology).

12:30 - 13:30

AT LT4

Tuesday, 26th June 2012

Mirror Descent is a technique for solving nonsmooth problems with convex structure, primarily, convex minimization and convex-concave saddle point problems. Mirror Descent utilizes first order information on the problem and is a far-reaching extension of the classical Subgradient Descent algorithm (N. Shor, 1967). This technique allows to adjust, to some extent, the algorithms to the geometry of the problem at hand and under favorable circumstances results in nearly dimension-independent and unimprovable in the large scale case convergence rates. As a result, in some important cases (e.g., when solving large-scale deterministic and stochastic convex problems on the domains like Euclidean/ℓ_1/nuclear norm balls), Mirror Descent algorithms become the methods of choice when low and medium accuracy solutions are sought.

In the tutorial, we outline the basic Mirror Descent theory for deterministic and stochastic convex minimization and convex-concave saddle point problems, including recent developments aimed at accelerating MD algorithms by utilizing problem's structure.

Representation Learning

12:30 - 15:00. Venue: AT LT5.

Presenters: Yoshua Bengio (University of Montreal)

Tuesday, 26th June 2012

The success of machine learning algorithms generally depends on data representation, and we hypothesize that this is because different representations can entangle and hide more or less the different explanatory factors of variation behind the data. Although domain knowledge can be used to help design representations, learning can also be used, and the quest for AI is motivating the design of more powerful representation-learning algorithms. We view the ultimate goal of these algorithms as disentangling the unknown underlying factors of variation that explain the observed data.This tutorial reviews the basics of feature learning and deep learning, as well as recent work relating these subjects to probabilistic modeling and manifold learning. An objective is to raise questions and issues about the appropriate objectives for learning good representations, for computing representations (i.e., inference), and the geometrical connections between representation learning, density estimation and manifold learning.

Prediction, Belief, and Markets

12:30 - 15:00. Venue: AT LT1.

Presenters: Jennifer Wortman Vaughan (UCLA) and Jacob Abernethy (University of Pennsylvania)

Tuesday, 26th June 2012

Prediction markets are financial markets designed to aggregate opinions across large populations of traders. A typical prediction market offers a set of securities with payoffs determined by the future state of the world. For example, a market might offer a security worth $1 if Barack Obama is re-elected in 2012 and $0 otherwise. Roughly speaking, a trader who believes the probability of Obama's re-election is p should be willing to buy this security at any price less than $p and (short) sell this security at any price greater than $p. For this reason, the going price of this security could be interpreted as traders' collective belief about the likelihood of Obama's re-election. Prediction markets have been used to generate accurate forecasts in a variety of domains including politics, disease surveillance, business, and entertainment, and are cited in the media increasingly often.

This tutorial will cover some of the basic mathematical ideas used in the design of prediction markets, and illustrate several fundamental connections between these ideas and techniques used in machine learning. We will begin with an overview of proper scoring rules, which can be used to measure the accuracy of a single entity's prediction, and are closely related to proper loss functions. We will then discuss market scoring rules, automated market makers based on proper scoring rules which can be used to aggregate the predictions of many forecasters, and describe how market scoring rules can be implemented as inventory-based markets in which securities are bought and sold. We will describe recent research exploring a duality-based interpretation of market scoring rules which can be exploited to design new markets that can be run efficiently over very large state spaces. Finally, we will explore the fundamental mathematical connections between market scoring rules and two areas of machine learning: online 'no-regret' learning and variational inference with exponential families.

This tutorial will be self-contained. No background on markets or specific areas of machine learning is required.

PAC-Bayesian Analysis in Supervised, Unsupervised, and Reinforcement Learning

15:30 - 18:00. Venue: AT LT1.

Presenters: Yevgeny Seldin (Max Planck Institute for Intelligent Systems, Tübingen, Germany), François Laviolette (Université Laval, Québec, Canada) and John Shawe-Taylor (University College London, London, UK)

Tuesday, 26th June 2012

PAC-Bayesian analysis is a basic and very general tool for data-dependent analysis in machine learning. By now, it has been applied in such diverse areas as supervised learning, unsupervised learning, and reinforcement learning, leading to state-of-the-art algorithms and accompanying generalization bounds. PAC-Bayesian analysis, in a sense, takes the best out of Bayesian methods and PAC learning and puts it together: (1) it provides an easy way to exploit prior knowledge (like Bayesian methods); (2) it provides strict and explicit generalization guarantees (like VC theory); and (3) it is data-dependent and provides an easy and strict way of exploiting benign conditions (like Rademacher complexities). In addition, PAC-Bayesian bounds directly lead to efficient learning algorithms. We will start with a general introduction to PAC-Bayesian analysis, which should be accessible to an average student, who is familiar with machine learning at the basic level. Then, we will survey multiple forms of PAC-Bayesian bounds and their numerous applications in different fields (including supervised and unsupervised learning, finite and continuous domains, and the very recent extension to martingales and reinforcement learning). Some of these applications will be explained in more details, while others will be surveyed at a high level. We will also describe the relations and distinctions between PAC-Bayesian analysis, Bayesian learning, VC theory, and Rademacher complexities. We will discuss the role, value, and shortcomings of frequentist bounds that are inspired by Bayesian analysis.

Causal Inference - Conditional Independency and Beyond

15:30 - 18:00. Venue: AT LT4.

Presenters: Dominik Janzing and Bernhard Schölkopf (Max Planck Institute for Intelligent Systems, Tübingen, Germany)

Tuesday, 26th June 2012

Machine learning has traditionally been focused on prediction. Given observations that have been generated by an unknown stochastic dependency, the goal is to infer a law that will be able to correctly predict future observations generated by the same dependency. As opposed to this goal, causal inference tries to infer the causal structure underlying the observed dependencies. More precisely, one tries to infer the behavior of a system under interventions without performing them, which does not fit into any traditional prediction scenario. Apart from the fact that it is still debated whether this is possible at all, it is a priori not clear, given that it is, why machine learning tools should be helpful for this task.

Since the Eighties there has been a community of researchers, mostly from statistics, philosophy, and computer science who have developed methods aiming at inferring causal relationships from observational data. The pioneering work of Glymour, Scheines, Spirtes, and Pearl describes assumptions that link conditional statistical dependences to causality, which then renders many causal inference problems solvable. The typical task, which is solved by the corresponding algorithms, reads: given observations from the joint distribution on the variables $X_1,...,X_n$ with $n \geq 3$, infer the causal directed acyclic graph (or parts of it). Recently, this work has been complemented by several researchers from machine learning who described methods that do not rely on conditional independences alone, but employ other properties of joint probability distributions. These methods use established tools of machine learning like, for instance, regression and reproducing kernel Hilbert spaces. As opposed to the above approaches, the causal direction can sometimes be inferred when only two variables are observed. Remarkably, this can be helpful for more traditional machine learning tasks like prediction under changing background conditions, because the task has different solutions depending on whether the predicted variable is the cause or the effect.

Spectral Approaches to Learning Latent Variable Models

15:30 - 18:00. Venue: AT LT5.

Presenters: Geoffrey J. Gordon (CMU), Le Song (Georgia Institute of Technology) and Byron Boots (CMU)

Tuesday, 26th June 2012

Many problems in machine learning involve collecting high-dimensional multivariate observations or sequences of observations, and then hypothesizing a compact model which explains these observations. An appealing representation for such a model is a latent variable model that relates a set of observed variables to an additional set of unobserved or hidden variables. Examples of popular latent variable models include latent tree graphical models and dynamical system models, both of which occupy a fundamental place in engineering, control theory, economics as well as the physical, biological, and social sciences. Unfortunately, to discover the right latent state representation and model parameters, we must solve difficult structural and temporal credit assignment problems. Work on learning latent variable structure has predominantly relied on likelihood maximization and local search heuristics such as expectation maximization (EM); these heuristics often lead to a search space with a host of bad local optima, and may therefore require impractically many restarts to reach a prescribed training precision. This tutorial will focus on a recently-discovered class of spectral learning algorithms. These algorithms hold the promise of overcoming these problems and enabling learning of latent structure in tree and dynamical system models. Unlike the EM algorithm, spectral methods are computationally efficient, statistically consistent, and have no local optima; in addition, they can be simple to implement, and have state-of-the-art practical performance for many interesting learning problems. We will describe the main theoretical, algorithmic, and empirical results related to spectral learning algorithms, starting with an overview of linear system identification results obtained in the last two decades, and then focusing on the remarkable recent progress in learning nonlinear dynamical systems, latent tree graphical models, and kernel-based nonparametric models.

4 Main Conference, Day 1, June 27

8:30	Welcome				
8:40	Invited talk. Sethu Muthukrishnan. AT LT2 / LT 4 / LT 5				
9:40	ICML Best paper award. AT LT 2 / LT 4 / LT 5				
10:00	*Coffee: Appleton Tower and Informatics Forum*				
	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
10:30	1A: Optimization Algorithms 1	1B: Reinforcement Learning 1	1C: Neural Networks & Deep Learning 1	1D: Structured Ouput Prediction	COLT
12:10	*Lunch*				
14:00	Invited talk. Dimitris Achlioptas. AT LT 2 / LT 4 / LT 5				
15:00	COLT Best paper award. AT LT 2 / LT 4 / LT 5				
15:30	*Coffee: Appleton Tower and Informatics Forum*				
	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
16:00	2A: Kernel Methods 1	2B: Reinforcement Learning	2C:Gaussian Processes	2D: Statistical Methods	COLT
17:40	Poster Session. Appleton Tower and Informatics Forum				
18:00	Open Problem Session. AT LT 3				

Wednesday June 27, 2012

Schedule of COLT Sessions can be found on page 169.

Plenary Session

08:30–10:00

08:30–08:40 *Welcome*

08:40–09:40 *Invited talk: Modern Algorithmic Tools for Analyzing Data Streams*
Sethu Muthukrishnan (p. 75)

09:40–10:00 *ICML Best paper award: Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring*
Sungjin Ahn, Anoop Korattikara, Max Welling (p. 76)

1A: Optimization algorithms 1

10:30–12:10, AT LT 4. Chair: Elad Hazan

10:30–10:50 *On the Equivalence between Herding and Conditional Gradient Algorithms*
Francis Bach, Simon Lacoste-Julien, Guillaume Obozinski (p. 76)

10:50–11:10 *Similarity Learning for Provably Accurate Sparse Linear Classification*
Aurélien Bellet, Amaury Habrard, Marc Sebban (p. 77)

11:10–11:30 *Stochastic Smoothing for Nonsmooth Minimizations: Accelerating SGD by Exploiting Structure*
Hua Ouyang, Alexander Gray (p. 77)

11:30–11:50 *To Average or Not to Average? Making Stochastic Gradient Descent Optimal for Strongly Convex Problems*
Alexander Rakhlin, Ohad Shamir, Karthik Sridharan (p. 77)

11:50–11:55 *Scaling Up Coordinate Descent Algorithms for Large ℓ_1 Regularization Problems*
Chad Scherrer, Mahantesh Halappanavar, Ambuj Tewari, David Haglin (p. 78)

11:55–12:00 *Quasi-Newton Methods: A New Direction*
Philipp Hennig, Martin Kiefel (p. 78)

12:00–12:05 *A Hybrid Algorithm for Convex Semidefinite Optimization*
Soeren Laue (p. 78)

12:05–12:10 *Efficient and Practical Stochastic Subgradient Descent for Nuclear Norm Regularization*
Haim Avron, Satyen Kale, Shiva Kasiviswanathan, Vikas Sindhwani (p. 78)

1B: Reinforcement learning 1

10:30–12:10, AT LT 5. Chair: David Silver

10:30–10:50 *Policy Gradients with Variance Related Risk Criteria*
Dotan Di Castro, Aviv Tamar, Shie Mannor (p. 79)

10:50–11:10 *Approximate Dynamic Programming By Minimizing Distributionally Robust Bounds*
Marek Petrik (p. 79)

11:10–11:30 *Statistical linear estimation with penalized estimators: an application to reinforcement learning*
Bernardo Avila Pires, Csaba Szepesvari (p. 79)

11:30–11:50 *Approximate Modified Policy Iteration*
Bruno Scherrer, Victor Gabillon, Mohammad Ghavamzadeh, Matthieu Geist (p. 80)

11:50–11:55 *A Dantzig Selector Approach to Temporal Difference Learning*
Matthieu Geist, Bruno Scherrer, Alessandro Lazaric, Mohammad Ghavamzadeh (p. 80)

11:55–12:00 *Linear Off-Policy Actor-Critic*
Thomas Degris, Martha White, Richard Sutton (p. 81)

12:00–12:05 *Lightning Does Not Strike Twice: Robust MDPs with Coupled Uncertainty*
Shie Mannor, Ofir Mebel, Huan Xu (p. 81)

12:05–12:10 *Bounded Planning in Passive POMDPs*
Roy Fox, Naftali Tishby (p. 81)

1C: Neural networks and deep learning 1

10:30–12:10, AT LT 1. Chair: Marc'Aurelio Ranzato

10:30–10:50 *Scene parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers*
Clément Farabet, Camille Couprie, Laurent Najman, Yann LeCun (p. 82)

10:50–11:10 *A Generative Process for Contractive Auto-Encoders*
Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio (p. 82)

11:10–11:30 *Deep Lambertian Networks*
Yichuan Tang, Ruslan Salakhutdinov, Geoffrey Hinton (p. 82)

11:30–11:50 *Deep Mixtures of Factor Analysers*
Yichuan Tang, Ruslan Salakhutdinov, Geoffrey Hinton (p. 83)

11:50–11:55 *Utilizing Static Analysis and Code Generation to Accelerate Neural Networks*
Lawrence McAfee, Kunle Olukotun (p. 83)

11:55–12:00 *Estimating the Hessian by Back-propagating Curvature*
James Martens, Ilya Sutskever, Kevin Swersky (p. 84)

12:00–12:05 *Training Restricted Boltzmann Machines on Word Observations*
George Dahl, Ryan Adams, Hugo Larochelle (p. 84)

12:05–12:10 *A fast and simple algorithm for training neural probabilistic language models*
Andriy Mnih, Yee Whye Teh (p. 84)

1D: Structured output prediction

10:30–12:10, AT LT 2. Chair: David McAllester

10:30–10:50 *Learning to Identify Regular Expressions that Describe Email Campaigns*
Paul Prasse, Christoph Sawade, Niels Landwehr, Tobias Scheffer (p. 85)

10:50–11:10 *Efficient Structured Prediction with Latent Variables for General Graphical Models*
Alexander Schwing, Tamir Hazan, Marc Pollefeys, Raquel Urtasun (p. 85)

11:10–11:30 *Output Space Search for Structured Prediction*
Janardhan Rao Doppa, Alan Fern, Prasad Tadepalli (p. 86)

11:30–11:50 *Efficient Decomposed Learning for Structured Prediction*
Rajhans Samdani, Dan Roth (p. 86)

11:50–12:10 *Modeling Latent Variable Uncertainty for Loss-based Learning*
M. Pawan Kumar, Ben Packer, Daphne Koller (p. 86)

Plenary Session

14:00–15:30

14:00–15:00 *Invited talk: Algorithmic Phase Transitions in Constraint Satisfaction Problems*
Dimitris Achlioptas (p. 73)

15:00–15:30 *COLT Best paper award*

2A: Kernel methods 1

16:00–17:40, AT LT 4. Chair: Arthur Gretton

16:00–16:20 *On the Size of the Online Kernel Sparsification Dictionary*
Yi Sun, Faustino Gomez, Juergen Schmidhuber (p. 87)

16:20–16:40 *Improved Nystrom Low-rank Decomposition with Priors*
Kai Zhang, Liang Lan, Jun Liu, andreas Rauber (p. 87)

16:40–17:00 *Bayesian Efficient Multiple Kernel Learning*
Mehmet Gönen (p. 88)

17:00–17:20 *A Binary Classification Framework for Two-Stage Multiple Kernel Learning*
Abhishek Kumar, Alexandru Niculescu-Mizil, Koray Kavukcuoglu, Hal Daume III (p. 88)

17:20–17:25 *Multiple Kernel Learning from Noisy Labels by Stochastic Programming*
Tianbao Yang, Rong Jin, Yang Zhou, Mehrdad Mahdavi, Lijun Zhang (p. 88)

17:25–17:30 *Subgraph Matching Kernels for Attributed Graphs*
Nils Kriege, Petra Mutzel (p. 89)

17:30–17:35 *Fast Computation of Subpath Kernel for Trees*
Daisuke Kimura, Hisashi Kashima (p. 89)

17:35–17:40 *Hypothesis testing using pairwise distances and associated kernels*
Dino Sejdinovic, Arthur Gretton, Bharath Sriperumbudur, Kenji Fukumizu (p. 89)

2B: Reinforcement learning 2

16:00–17:40, AT LT 5. Chair: Geoff Gordon

16:00–16:20 *No-Regret Learning in Extensive-Form Games with Imperfect Recall*
Marc Lanctot, Richard Gibson, Neil Burch, Michael Bowling (p. 90)

16:20–16:40 *Near-Optimal BRL using Optimistic Local Transitions*
Mauricio Araya, Olivier Buffet, Vincent Thomas (p. 90)

16:40–17:00 *Continuous Inverse Optimal Control with Locally Optimal Examples*
Sergey Levine, Vladlen Koltun (p. 91)

17:00–17:20 *Monte Carlo Bayesian Reinforcement Learning*
Yi Wang, Kok Sung Won, David Hsu, Wee Sun Lee (p. 91)

17:20–17:40 *Apprenticeship Learning for Model Parameters of Partially Observable Environments*
Takaki Makino, Johane Takeuchi (p. 91)

2C: Gaussian processes

16:00–17:40, AT LT 1. Chair: Ryan Adams

16:00–16:20 *Gaussian Process Regression Networks*
Andrew Wilson, David Knowles, Zoubin Ghahramani (p. 92)

16:20–16:40 *Infinite Tucker Decomposition: Nonparametric Bayesian Models for Multiway Data Analysis*
Zenglin Xu, Feng Yan, Alan Qi (p. 92)

16:40–17:00 *State-Space Inference for Non-Linear Latent Force Models with Application to Satellite Orbit Prediction*
Jouni Hartikainen, Mari Seppänen, Simo Särkkä (p. 93)

17:00–17:20 *Gaussian Process Quantile Regression using Expectation Propagation*
Alexis Boukouvalas, Remi Barillec, Dan Cornford (p. 93)

17:20–17:25 *Residual Components Analysis*
Alfredo Kalaitzis, Neil Lawrence (p. 93)

17:25–17:30 *Manifold Relevance Determination*
Andreas Damianou, Carl Ek, Michalis Titsias, Neil Lawrence (p. 93)

2D: Statistical methods

16:00–17:40, AT LT 2. Chair: Lawrence Carin

16:00–17:30 *Lognormal and Gamma Mixed Negative Binomial Regression*
Mingyuan Zhou, Lingbo Li, David Dunson, Lawrence Carin (p. 94)

16:20–17:30 *Group Sparse Additive Models*
Junming Yin, Xi Chen, eric xing (p. 94)

16:40–17:00 *Variance Function Estimation in High-dimensions*
Mladen Kolar, James Sharpnack (p. 95)

17:00–17:20 *Sparse Additive Functional and Kernel CCA*
Sivaraman Balakrishnan, Kriti Puniyani, John Lafferty (p. 95)

17:20–17:25 *Consistent Covariance Selection From Data With Missing Values*
Mladen Kolar, eric xing (p. 96)

17:25–17:30 *Conditional Sparse Coding and Grouped Multivariate Regression*
Min Xu, John Lafferty (p. 96)

17:30–17:35 *Is margin preserved after random projection?*
Qinfeng Shi, Chunhua Shen, Rhys Hill, Anton van den Hengel (p. 96)

Poster Session

17:40, AT and IF

Posters 1–48 are in the Informatics Forum (IF). Posters 49–96 are in Appleton Tower (AT). See floor plans at end of book.

(1) *Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring*
Sungjin Ahn, Anoop Korattikara, Max Welling (p. 76)

(2) *On the Equivalence between Herding and Conditional Gradient Algorithms*
Francis Bach, Simon Lacoste-Julien, Guillaume Obozinski (p. 76)

(3) *Similarity Learning for Provably Accurate Sparse Linear Classification*
Aurélien Bellet, Amaury Habrard, Marc Sebban (p. 77)

(4) *Stochastic Smoothing for Nonsmooth Minimizations: Accelerating SGD by Exploiting Structure*
Hua Ouyang, Alexander Gray (p. 77)

(5) *To Average or Not to Average? Making Stochastic Gradient Descent Optimal for Strongly Convex Problems*
Alexander Rakhlin, Ohad Shamir, Karthik Sridharan (p. 77)

(6) *Scaling Up Coordinate Descent Algorithms for Large ℓ_1 Regularization Problems*
Chad Scherrer, Mahantesh Halappanavar, Ambuj Tewari, David Haglin (p. 78)

(7) *Quasi-Newton Methods: A New Direction*
Philipp Hennig, Martin Kiefel (p. 78)

(8) *A Hybrid Algorithm for Convex Semidefinite Optimization*
Soeren Laue (p. 78)

(9) *Efficient and Practical Stochastic Subgradient Descent for Nuclear Norm Regularization*
Haim Avron, Satyen Kale, Shiva Kasiviswanathan, Vikas Sindhwani (p. 78)

(10) *Policy Gradients with Variance Related Risk Criteria*
Dotan Di Castro, Aviv Tamar, Shie Mannor (p. 79)

(11) *Approximate Dynamic Programming By Minimizing Distributionally Robust Bounds*
Marek Petrik (p. 79)

(12) *Statistical linear estimation with penalized estimators: an application to reinforcement learning*
Bernardo Avila Pires, Csaba Szepesvari (p. 79)

(56) *Is margin preserved after random projection?*
Qinfeng Shi, Chunhua Shen, Rhys Hill, Anton van den Hengel
(p. 96)

COLT Posters

17:40, AT

COLT posters are in Appleton Tower (AT). See floor plans at end of book.

(57) *A Conjugate Property between Loss Functions and Uncertainty Sets in Classification Problems*
Takafumi Kanamori, Akiko Takeda, Taiji Suzuki

(58) *Rare Probability Estimation under Regularly Varying Heavy Tails*
Mesrob Ohannessian, Munther Dahleh

(59) *L_1 Covering Numbers for Uniformly Bounded Convex Functions*
Adityanand Guntuboyina, Bodhisattva Sen

(60) *Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences*
Benjamin Recht, Christopher Re

(61) *Attribute-Efficient Learning and Weight-Degree Tradeoffs for Polynomial Threshold Functions*
Rocco Servedio, Li-Yang Tan, Justin Thaler

(62) *Distance Preserving Embeddings for General n-Dimensional Manifolds*
Nakul Verma

(63) *A Method of Moments for Hidden Markov Models and Multi-view Mixture Models*
Animashree Anandkumar, Daniel Hsu, Sham Kakade

(64) *Autonomous Exploration For Navigating In MDPs*
Shiau Hong Lim, Peter Auer

(65) *The Optimality of Jeffreys Prior for Online Density Estimation and the Asymptotic Normality of Maximum Likelihood Estimators*
Fares Hedayati, Peter Bartlett

(66) *Toward understanding complex spaces: graph Laplacians on manifolds with singularities and boundaries*
Mikhail Belkin, Qichao Que, Yusu Wang, Xueyuan Zhou

(67) *Differentially Private Online Learning*
Prateek Jain, Pravesh Kothari, Abhradeep Thakurta

(68) *Active Learning Using Smooth Relative Regret Approximations with Applications*
Nir Ailon, Ron Begleiter, Esther Ezra

(69) *Distributed Learning, Communication Complexity and Privacy*
Maria-Florina Balcan, Avrim Blum, Shai Fine, Yishay Mansour

(70) *Divergences and Risks for Multiclass Experiments*
Dario Garcia Garcia, Robert C. Williamson

(71) *Tight Bounds on Proper Equivalence Query Learning of DNF*
Lisa Hellerstein, Devorah Kletenik, Linda Sellie, Rocco Servedio

(72) *Generalization Bounds for Online Learning Algorithms with Pairwise Loss Functions*
Yuyang Wang, Roni Khardon, Dmitry Pechyony, Rosie Jones

(73) *Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model*
Kamalika Chaudhuri, Fan Chung, Alexander Tsiatas

(74) *Exact Recovery of Sparsely-Used Dictionaries*
Daniel Spielman, Huan Wang, John Wright

(75) *Analysis of Thompson Sampling for the multi-armed bandit problem*
Shipra Agrawal, Navin Goyal

(76) *Learning Valuation Functions*
Maria-Florina Balcan, Florin Constantin, Satoru Iwata, Lei Wang

(77) *A Correlation Clustering Approach to Link Classification in Signed Networks*
Nicolò Cesa-Bianchi, Claudio Gentile, Fabio Vitale, Giovanni Zappella

(78) *Differentially Private Convex Optimization for Empirical Risk Minimization with Applications to High-dimensional Regression*
Daniel Kifer, Adam Smith, Abhradeep Thakurta

(79) *Competitive Classification and Closeness Testing*
Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, Ananda Theertha Suresh

(80) *Unified Algorithms for Online Learning and Competitive Analysis*
Niv Buchbinder, Shahar Chen, Joseph Naor, Ohad Shamir

(81) *Robust Interactive Learning*
Maria-Florina Balcan, Steve Hanneke

(82) *Online Optimization with Gradual Variations*
Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, Shenghuo Zhu

Student Scholarship Programme Posters

17:40, AT

(83) *Kernel Methods for Multiple-Instance Classification and Regression*
Gary Doran

(84) *Competing Against Strategies*
Wei Han

(85) *Minimizing On-Line Learning by Exploiting Stationary Properties in Nonstationary Environments*
Timothy Arthur Mann

COLT/ICML Open Problems

18:00, AT LT 3

Regret Bounds for Thompson Sampling
Lihong Li, Olivier Chapelle

Better Bounds for Online Logistic Regression
H. Brendan McMahan, Matthew Streeter

Does AdaBoost Always Cycle?
Cynthia Rudin, Robert E. Schapire, Ingrid Daubechies

Learning dynamic network models from a static snapshot
Jan Ramon, Constantin Comendant

Is Averaging Needed for Strongly Convex Stochastic Gradient Descent?
Ohad Shamir

5 Main Conference, Day 2, June 28

	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
08:40	3A: Optimization Algorithms 2	3B: Clustering 1	3C: Privacy, Anonymity and Security	3D: Ranking and Preference Learning	3E: Nonparametric Bayesian Inference
10:00	*Coffee: Appleton Tower and Informatics Forum*				
	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
10:30	4A: Feature Selection and Dimensionality Reduction 1	4B: Online Learning	4C: Supervised Learning 1	4D: Transfer and Multi-Task Learning	4E: Graphical Models
12:10	*Lunch*				
14:00	Invited talk. David MacKay. AT LT2 / LT 4 / LT 5				
15:00	Test-of-time award talk. Risi Kondor and John Lafferty. AT LT2 / LT 4 / LT 5				
15:30	*Coffee: Appleton Tower and Informatics Forum*				
	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
16:00	5A: Learning Theory	5B: Online Learning 2	5C: Neural Networks and Deep Learning 2	5D: Sparsity and Compressed Sensing	5E: Latent-Variable Models and Topic Models
17:40	Poster Session. Appleton Tower and Informatics Forum				

Thursday, June 28, 2012

3A: Optimization algorithms 2

08:40–10:00, AT LT 4. Chair: Tong Zhang

3B: Clustering 1

08:40–09:55, AT LT 5. Chair: Raquel Urtasun

3C: Privacy, Anonymity, and Security

08:40–10:00, AT LT 1. Chair: Tobias Scheffer

09:20–09:40 *Convergence Rates for Differentially Private Statistical Estimation*
Kamalika Chaudhuri, Daniel Hsu (p. 101)

09:40–10:00 *Finding Botnets Using Minimal Graph Clusterings*
Peter Haider, Tobias Scheffer (p. 101)

3D: Ranking and Preference Learning

08:40–09:55, AT LT 2. Chair: Balazs Kegl

08:40–09:00 *Incorporating Domain Knowledge in Matching Problems via Harmonic Analysis*
Deepti Pachauri, Maxwell Collins, Vikas SIngh (p. 102)

09:00–09:20 *Consistent Multilabel Ranking through Univariate Losses*
Krzysztof Dembczyński, Wojciech Kotłowski, Eyke Huellermeier (p. 102)

09:20–09:40 *Predicting Consumer Behavior in Commerce Search*
Or Sheffet, Nina Mishra, Samuel Ieong (p. 102)

09:40–09:45 *Adaptive Regularization for Similarity Measures*
Koby Crammer, Gal Chechik (p. 103)

09:45–09:50 *Online Structured Prediction via Coactive Learning*
Pannaga Shivaswamy, Thorsten Joachims (p. 103)

09:50–09:55 *TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings*
Chao Liu (p. 104)

3E: Nonparametric Bayesian inference

08:40–10:00, AT LT 3. Chair: Sharon Goldwater

08:40–09:00 *Factorized Asymptotic Bayesian Hidden Markov Models*
Ryohei Fujimaki, Kohei Hayashi (p. 104)

09:00–09:20 *An Infinite Latent Attribute Model for Network Data*
David Knowles, Konstantina Palla, Zoubin Ghahramani (p. 104)

09:20–09:40 *The Nonparametric Metadata Dependent Relational Model*
Dae Il Kim, Michael Hughes, Erik Sudderth (p. 105)

09:40–09:45 *Dependent Hierarchical Normalized Random Measures for Dynamic Topic Modeling*
Changyou Chen, Nan Ding, Wray Buntine (p. 105)

09:45–09:50 *A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling*
Drausin Wulsin, Shane Jensen, Brian Litt (p. 105)

09:50–09:55 *Modeling Images using Transformed Indian Buffet Processes*
KE ZHAI, Yuening Hu, Jordan Boyd-Graber, Sinead Williamson
(p. 106)

09:55–10:00 *A Topic Model for Melodic Sequences*
Athina Spiliopoulou, Amos Storkey (p. 106)

4A: Feature selection and dimensionality reduction 1

10:30–12:10, AT LT 4. Chair: Kilian Weinberger

10:30–10:50 *Discovering Support and Affiliated Features from Very High Dimensions*
Yiteng Zhai, Mingkui Tan, Ivor Tsang, Yew Soon Ong (p. 106)

10:50–11:10 *Inferring Latent Structure From Mixed Real and Categorical Relational Data*
Esther Salazar, Lawrence Carin (p. 107)

11:10–11:30 *Conditional Likelihood Maximization: A Unifying Framework for Information Theoretic Feature Selection*
Gavin Brown, Adam Pocock, Ming-Jie Zhao, Mikel Lujan (p. 107)

11:30–11:50 *Dimensionality Reduction by Local Discriminative Gaussians*
Nathan Parrish, Maya Gupta (p. 108)

11:50–12:10 *Fast Prediction of New Feature Utility*
Hoyt Koepke, Mikhail Bilenko (p. 108)

4B: Online learning 1

10:30–12:10, AT LT 5. Chair: Satyen Kale

10:30–10:50 *An Online Boosting Algorithm with Theoretical Justifications*
Shang-Tse Chen, Hsuan-Tien Lin, Chi-Jen Lu (p. 109)

10:50–11:10 *An adaptive algorithm for finite stochastic partial monitoring*
Gabor Bartok, Navid Zolghadr, Csaba Szepesvari (p. 109)

11:10–11:30 *Online Alternating Direction Method*
Huahua Wang, Arindam Banerjee (p. 109)

11:30–11:50 *Projection-free Online Learning*
Elad Hazan, Satyen Kale (p. 109)

11:50–11:55 *PAC Subset Selection in Stochastic Multi-armed Bandits*
Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, Peter Stone
(p. 110)

11:55–12:00 *On Local Regret*
Michael Bowling, Martin Zinkevich (p. 110)

12:00–12:05 *Exact Soft Confidence-Weighted Learning*
Steven C.H. Hoi, Jialei Wang, Peilin Zhao (p. 110)

12:05–12:10 *Compact Hyperplane Hashing with Bilinear Functions*
Wei Liu, Jun Wang, Yadong Mu, Sanjiv Kumar, Shih-Fu Chang
(p. 111)

4C: Supervised learning 1

10:30–12:10, AT LT 1. Chair: Cynthia Rudin

10:30–10:50 *Improved Information Gain Estimates for Decision Tree Induction*
Sebastian Nowozin (p. 111)

10:50–11:10 *Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation*
Kendrick Boyd, Jesse Davis, David Page, Vitor Santos Costa
(p. 112)

11:10–11:30 *Bootstrapping Big Data*
Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, Michael Jordan
(p. 112)

11:30–11:50 *Robust Classification with Adiabatic Quantum Optimization*
Vasil Denchev, Nan Ding, SVN Vishwanathan, Hartmut Neven
(p. 112)

11:50–11:55 *Nonparametric Link Prediction in Dynamic Networks*
Purnamrita Sarkar, Deepayan Chakrabarti, Michael Jordan (p. 113)

11:55–12:00 *A Unified Robust Classification Model*
Akiko Takeda, Hiroyuki Mitsugi , Takafumi Kanamori (p. 113)

12:00–12:05 *Maximum Margin Output Coding*
Yi Zhang, Jeff Schneider (p. 113)

12:05–12:10 *Structured Learning from Partial Annotations*
Xinghua Lou, Fred Hamprecht (p. 114)

4D: Transfer and Multi-Task Learning

10:30–12:10, AT LT 2. Chair: Jenn Wortman Vaughan

10:30–10:50 *Marginalized Denoising Autoencoders for Domain Adaptation*
Minmin Chen, Zhixiang Xu, Kilian Weinberger, Fei Sha (p. 114)

10:50–11:10 *Information-Theoretical Learning of Discriminative Clusters for Unsupervised Domain Adaptation*
Yuan Shi, Fei Sha (p. 114)

11:10–11:30 *Learning Task Grouping and Overlap in Multi-task Learning*
Abhishek Kumar, Hal Daume III (p. 115)

11:30–11:50 *A Convex Feature Learning Formulation for Latent Task Structure Discovery*
Pratik Jawanpuria, J. Saketha Nath (p. 115)

11:50–11:55 *Convex Multitask Learning with Flexible Task Clusters*
Wenliang Zhong, James Kwok (p. 116)

11:55–12:00 *A Complete Analysis of the $l_{1,p}$ Group-Lasso*
Julia Vogt, Volker Roth (p. 116)

12:00–12:05 *Learning with Augmented Features for Heterogeneous Domain Adaptation*
Lixin Duan, Dong Xu, Ivor Tsang (p. 116)

12:05–12:10 *Cross-Domain Multitask Learning with Latent Probit Models*
Shaobo Han, Xuejun Liao, Lawrence Carin (p. 117)

4E: Graphical models

10:30–12:10, AT LT 3. Chair: Matthias Seeger

10:30–10:50 *High Dimensional Semiparametric Gaussian Copula Graphical Models*
Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman (p. 117)

10:50–11:10 *Convergence Rates of Biased Stochastic Optimization for Learning Sparse Ising Models*
Jean Honorio (p. 118)

11:10–11:30 *On the Partition Function and Random Maximum A-Posteriori Perturbations*
Tamir Hazan, Tommi Jaakkola (p. 118)

11:30–11:50 *Anytime Marginal MAP Inference*
Denis Maua, Cassio De Campos (p. 118)

11:50–11:55 *Exact Maximum Margin Structure Learning of Bayesian Networks*
Robert Peharz, Franz Pernkopf (p. 119)

11:55–12:00 *LPQP for MAP: Putting LP Solvers to Better Use*
Patrick Pletscher, Sharon Wulff (p. 119)

12:00–12:05 *How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing*
Yoram Bachrach, Thore Graepel, Tom Minka, John Guiver (p. 119)

12:05–12:10 *Smoothness and Structure Learning by Proxy*
Benjamin Yackley, Terran Lane (p. 120)

Plenary Session

14:00–15:30

14:00–15:00 *Invited talk: Information Theory and Sustainable Energy*
David MacKay FRS (p. 75)

15:00–15:30 *Test-of-time award talk: Diffusion Kernels on Graphs and Other Discrete Input Spaces*
Risi Kondor, John Lafferty (p. 73)

5A: Learning theory

16:00–17:40, AT LT 4. Chair: Daniel Hsu

16:00–16:20 *Linear Regression with Limited Observation*
Elad Hazan, Tomer Koren (p. 120)

16:20–16:40 *Optimizing F-measure: A Tale of Two Approaches*
Ye Nan, Kian Ming Chai, Wee Sun Lee, Hai Leong Chieu (p. 120)

16:40–17:00 *Conditional mean embeddings as regressors*
Steffen Grunewalder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, Massi Pontil (p. 121)

17:00–17:20 *PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification*
Emilie Morvant, Sokol Koço, Liva Ralaivola (p. 121)

17:20–17:25 *Tighter Variational Representations of f-Divergences via Restriction to Probability Measures*
Avraham Ruderman, Mark Reid, Darío García-García, James Petterson (p. 122)

17:25–17:30 *Agglomerative Bregman Clustering*
Matus Telgarsky, Sanjoy Dasgupta (p. 122)

17:30–17:35 *The Convexity and Design of Composite Multiclass Losses*
Mark Reid, Robert Williamson, Peng Sun (p. 122)

17:35–17:40 *Minimizing The Misclassification Error Rate Using a Surrogate Convex Loss*
Shai Ben-David, David Loker, Nathan Srebro, Karthik Sridharan (p. 122)

5B: Online learning 2

16:00–17:40, AT LT 5. Chair: Csaba Szepesvari

16:00–16:20 *Hierarchical Exploration for Accelerating Contextual Bandits*
Yisong Yue, Sue Ann Hong, Carlos Guestrin (p. 122)

16:20–16:40 *Online Bandit Learning against an Adaptive Adversary: from Regret to Policy Regret*
Ofer Dekel, Ambuj Tewari, Raman Arora (p. 123)

16:40–17:00 *Decoupling Exploration and Exploitation in Multi-Armed Bandits*
Orly Avner, Shie Mannor, Ohad Shamir (p. 123)

17:00–17:20 *Learning the Experts for Online Sequence Prediction*
Elad Eban, Amir Globerson, Shai Shalev-Schwartz, Aharon Birnbaum (p. 124)

17:20–17:25 *Plug-in martingales for testing exchangeability on-line*
Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, Volodya Vovk (p. 124)

17:25–17:30 *Parallelizing Exploration-Exploitation Tradeoffs with Gaussian Process Bandit Optimization*
Thomas Desautels, Andreas Krause, Joel Burdick (p. 124)

17:30–17:35 *On-Line Portfolio Selection with Moving Average Reversion*
Bin Li, Steven C.H. Hoi (p. 125)

17:35–17:40 *Exponential Regret Bounds for Gaussian Process Bandits with Deterministic Observations*
Nando de Freitas, Alex Smola, Masrour Zoghi (p. 125)

5C: Neural networks and deep learning 2

16:00–17:40, AT LT 1. Chair: Yoshua Bengio

16:00–16:20 *Large-Scale Feature Learning With Spike-and-Slab Sparse Coding*
Ian Goodfellow, Aaron Courville, Yoshua Bengio (p. 126)

16:20–16:40 *Learning Local Transformation Invariance with Restricted Boltzmann Machines*
Kihyuk Sohn, Honglak Lee (p. 126)

16:40–17:00 *Building high-level features using large scale unsupervised learning*
Quoc Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Jeff Dean, Andrew Ng (p. 127)

17:00–17:20 *On multi-view feature learning*
Roland Memisevic (p. 127)

17:20–17:40 *Learning to Label Aerial Images from Noisy Data*
Volodymyr Mnih, Geoffrey Hinton (p. 127)

5D: Sparsity and compressed sensing

16:00–17:40, AT LT 2. Chair: Pradeep Ravikumar

16:00–16:20 *Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering*
Gael Varoquaux, Alexandre Gramfort, Bertrand Thirion (p. 128)

16:20–16:40 *Estimation of Simultaneously Sparse and Low Rank Matrices*
Pierre-André Savalle, Emile Richard, Nicolas Vayatis (p. 128)

16:40–17:00 *Multi-level Lasso for Sparse Multi-task Regression*
Aurelie Lozano (p. 128)

17:00–17:20 *Efficient Euclidean Projections onto the Intersection of Norm Balls*
Wei YU, Hao Su, Li Fei-Fei (p. 129)

17:20–17:40 *Learning Efficient Structured Sparse Models*
Alex Bronstein, Pablo Sprechmann, Guillermo Sapiro (p. 129)

5E: Latent-Variable Models and Topic Models

16:00–17:35, AT LT 3. Chair: Jordan Boyd-Graber

16:00–16:20 *Max-Margin Nonparametric Latent Feature Models for Link Prediction*
Jun Zhu (p. 130)

16:20–16:40 *Canonical Trends: Detecting Trend Setters in Web Data*
Felix Biessmann, Jens-Michalis Papaioannou, Mikio Braun, Andreas Harth (p. 130)

16:40–17:00 *Variational Inference in Non-negative Factorial Hidden Markov Models for Efficient Audio Source Separation*
Gautham Mysore, Maneesh Sahani (p. 130)

17:00–17:20 *Sparse stochastic inference for latent Dirichlet allocation*
David Mimno, Matt Hoffman, David Blei (p. 131)

17:20–17:25 *Dirichlet Process with Mixed Random Measures: A Nonparametric Topic Model for Labeled Data*
Dongwoo Kim, Suin Kim, Alice Oh (p. 131)

17:25–17:30 *Rethinking Collapsed Variational Bayes Inference for LDA*
Issei Sato, Hiroshi Nakagawa (p. 131)

17:30–17:35 *Capturing topical content with frequency and exclusivity*
Jonathan Bischof, Edoardo Airoldi (p. 131)

Poster Session

17:40, AT and IF

Posters 1–48 are in the Informatics Forum (IF). Posters 49–96 are in Appleton Tower (AT). See floor plans at end of book.

(1) *Integer Optimization Methods for Supervised Ranking*
Allison Chang, Cynthia Rudin, Dimitris Bertsimas (p. 97)

(2) *A Proximal-Gradient Homotopy Method for the L1-Regularized Least-Squares Problem*
Lin Xiao, Tong Zhang (p. 97)

(3) *Complexity Analysis of the Lasso Regularization Path*
Julien Mairal, Bin Yu (p. 97)

(4) *Randomized Smoothing for (Parallel) Stochastic Optimization*
John Duchi, Martin Wainwright, Peter Bartlett (p. 98)

(5) *Demand-Driven Clustering in Relational Domains for Predicting Adverse Drug Events*
Jesse Davis, Vitor Santos Costa, Elizabeth Berg, David Page, Peggy Peissig, Michael Caldwell (p. 98)

(6) *Clustering to Maximize the Ratio of Split to Diameter*
Jiabing Wang, Jiaye Chen (p. 98)

(7) *An Iterative Locally Linear Embedding Algorithm*
Deguang Kong, Chris H.Q. Ding (p. 99)

(8) *Robust Multiple Manifold Structure Learning*
Dian Gong, Xuemei Zhao, Gerard Medioni (p. 99)

(9) *A Split-Merge Framework for Comparing Clusterings*
Qiaoliang Xiang, Qi Mao, Kian Ming Chai, Hai Leong Chieu, Ivor Tsang, Zhenddong Zhao (p. 99)

(10) *On the Difficulty of Nearest Neighbor Search*
Junfeng He, Sanjiv Kumar, Shih-Fu Chang (p. 100)

(11) *Bayesian Watermark Attacks*
Ivo Shterev, David Dunson (p. 100)

(12) *Poisoning Attacks against Support Vector Machines*
Battista Biggio, Blaine Nelson, Pavel Laskov (p. 100)

(13) *Convergence Rates for Differentially Private Statistical Estimation*
Kamalika Chaudhuri, Daniel Hsu (p. 101)

(14) *Finding Botnets Using Minimal Graph Clusterings*
Peter Haider, Tobias Scheffer (p. 101)

(75) *Decoupling Exploration and Exploitation in Multi-Armed Bandits*
Orly Avner, Shie Mannor, Ohad Shamir (p. 123)

(76) *Learning the Experts for Online Sequence Prediction*
Elad Eban, Amir Globerson, Shai Shalev-Schwartz, Aharon Birnbaum (p. 124)

(77) *Plug-in martingales for testing exchangeability on-line*
Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, Volodya Vovk (p. 124)

(78) *Parallelizing Exploration-Exploitation Tradeoffs with Gaussian Process Bandit Optimization*
Thomas Desautels, Andreas Krause, Joel Burdick (p. 124)

(79) *On-Line Portfolio Selection with Moving Average Reversion*
Bin Li, Steven C.H. Hoi (p. 125)

(80) *Exponential Regret Bounds for Gaussian Process Bandits with Deterministic Observations*
Nando de Freitas, Alex Smola, Masrour Zoghi (p. 125)

(81) *Large-Scale Feature Learning With Spike-and-Slab Sparse Coding*
Ian Goodfellow, Aaron Courville, Yoshua Bengio (p. 126)

(82) *Learning Local Transformation Invariance with Restricted Boltzmann Machines*
Kihyuk Sohn, Honglak Lee (p. 126)

(83) *Building high-level features using large scale unsupervised learning*
Quoc Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Jeff Dean, Andrew Ng (p. 127)

(84) *On multi-view feature learning*
Roland Memisevic (p. 127)

(85) *Learning to Label Aerial Images from Noisy Data*
Volodymyr Mnih, Geoffrey Hinton (p. 127)

(86) *Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering*
Gael Varoquaux, Alexandre Gramfort, Bertrand Thirion (p. 128)

(87) *Estimation of Simultaneously Sparse and Low Rank Matrices*
Pierre-André Savalle, Emile Richard, Nicolas Vayatis (p. 128)

(88) *Multi-level Lasso for Sparse Multi-task Regression*
Aurelie Lozano (p. 128)

(89) *Efficient Euclidean Projections onto the Intersection of Norm Balls*
Wei YU, Hao Su, Li Fei-Fei (p. 129)

(90) *Learning Efficient Structured Sparse Models*
Alex Bronstein, Pablo Sprechmann, Guillermo Sapiro (p. 129)

(91) *Max-Margin Nonparametric Latent Feature Models for Link Prediction*
Jun Zhu (p. 130)

(92) *Canonical Trends: Detecting Trend Setters in Web Data*
Felix Biessmann, Jens-Michalis Papaioannou, Mikio Braun, Andreas Harth (p. 130)

(93) *Variational Inference in Non-negative Factorial Hidden Markov Models for Efficient Audio Source Separation*
Gautham Mysore, Maneesh Sahani (p. 130)

(94) *Sparse stochastic inference for latent Dirichlet allocation*
David Mimno, Matt Hoffman, David Blei (p. 131)

(95) *Dirichlet Process with Mixed Random Measures: A Nonparametric Topic Model for Labeled Data*
Dongwoo Kim, Suin Kim, Alice Oh (p. 131)

(96) *Rethinking Collapsed Variational Bayes Inference for LDA*
Issei Sato, Hiroshi Nakagawa (p. 131)

(97) *Capturing topical content with frequency and exclusivity*
Jonathan Bischof, Edoardo Airoldi (p. 131)

6 Main Conference, Day 3, June 29

	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
08:40	6A: Semi-supervised Learning	6B: Reinforcement Learning 3	6C: Applications	6D: Time-Series Analysis	6E: Graph-based Learning
10:00	*Coffee: Appleton Tower and Informatics Forum*				
	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
10:30	7A: Invited Applications	7B: Reinforcement Learning 4	7C: Clustering 2	7D: Supervised Learning 2	7E: Probabilistic Models
12:10	*Lunch*				
	AT LT 4	AT LT 5	AT LT 1	AT LT 2	AT LT 3
14:00	8A: Kernel Methods 2	8B: Active Methods and Cost-Sensitive Learning	8C: Feature Selection and Dimensionality Reduction 2	8D: Recommendation and Matrix Factorisation	8E: Graphical Models
15:40	*Coffee: Appleton Tower and Informatics Forum*				
16:10	Invited talk. Yann LeCun. AT LT2 / LT 4 / LT 5				
17:10	Applications talk. Kiri Wagstaff. AT LT2 / LT 4 / LT 5				
17:40	Business meeting. AT LT2				
17:40	Poster Session. Appleton Tower and Informatics Forum				

Friday, June 29, 2012

6A: Semi-supervised learning

08:40–10:00, AT LT 4. Chair: Maria Florina Balcan

08:40–09:00 *A convex relaxation for weakly supervised classifiers*
Armand Joulin, Francis Bach (p. 132)

09:00–09:20 *Semi-Supervised Learning of Class Balance under Class-Prior Change by Distribution Matching*
Marthinus Du Plessis, Masashi Sugiyama (p. 132)

09:20–09:40 *A Simple Algorithm for Semi-supervised Learning with Improved Generalization Error Bound*
Ming Ji, Tianbao Yang, Binbin Lin, Rong Jin, Jiawei Han (p. 132)

09:40–09:45 *Information-theoretic Semi-supervised Metric Learning via Entropy Regularization*
Gang Niu, Bo Dai, Makoto Yamada, Masashi Sugiyama (p. 133)

09:45–09:50 *Cross Language Text Classification via Subspace Co-regularized Multi-view Learning*
Yuhong Guo, Min Xiao (p. 133)

09:50–09:55 *Using CCA to improve CCA: A new spectral method for estimating vector models of words*
Paramveer Dhillon, Jordan Rodu, Dean Foster, Lyle Ungar (p. 133)

09:55–10:00 *Semi-Supervised Collective Classification via Hybrid Label Regularization*
Luke McDowell, David Aha (p. 134)

6B: Reinforcement learning 3

08:40–10:00, AT LT 5. Chair: Ron Parr

08:40–09:00 *Compositional Planning Using Optimal Option Models*
David Silver, Kamil Ciosek (p. 134)

09:00–09:20 *Learning Parameterized Skills*
Bruno Da Silva, George Konidaris, Andrew Barto (p. 134)

09:20–09:40 *Safe Exploration in Markov Decision Processes*
Teodor Mihai Moldovan, Pieter Abbeel (p. 135)

09:40–10:00 *Modelling transition dynamics in MDPs with RKHS embeddings*
Steffen Grunewalder, Guy Lever, Luca Baldassarre, Massi Pontil, Arthur Gretton (p. 135)

6C: Applications

08:40–10:00, AT LT 1. Chair: Tom Dietterich

6D: Time-Series Analysis

08:40–10:00, AT LT 2. Chair: Naoki Abe

6E: Graph-based learning

08:40–10:00, AT LT 3. Chair: Charles Elkan

09:45–09:50 *The Most Persistent Soft-Clique in a Set of Sampled Graphs*
Novi Quadrianto, Chao Chen, Christoph Lampert (p. 140)

09:50–09:55 *Two Manifold Problems with Applications to Nonlinear System Identification*
Byron Boots, Geoff Gordon (p. 140)

09:55–10:00 *Incorporating Causal Prior Knowledge as Path-Constraints in Bayesian Networks and Maximal Ancestral Graphs*
Giorgos Borboudakis, Ioannis Tsamardinos (p. 141)

7A: Invited Applications

10:30–12:10, AT LT 4. Chair: Samy Bengio

10:30–10:50 *Conversational Speech Transcription Using Context-Dependent Deep Neural Networks*
Dong Yu, Frank Seide, Gang Li (p. 141)

10:50–11:10 *Data-driven Web Design*
Ranjitha Kumar, Jerry Talton, Salman Ahmad, Scott Klemmer (p. 142)

11:10–11:30 *Learning the Central Events and Participants in Unlabeled Text*
Nathanael Chambers, Dan Jurafsky (p. 142)

11:30–11:50 *Exemplar-SVMs for Visual Object Detection, Label Transfer and Image Retrieval*
Tomasz Malisiewicz, Abhinav Shrivastava, Abhinav Gupta, Alexei Efros (p. 142)

11:50–12:10 *Learning Force Control Policies for Compliant Robotic Manipulation*
Mrinal Kalakrishnan, Ludovic Righetti, Peter Pastor, Stefan Schaal (p. 142)

7B: Reinforcement learning 4

10:30–12:10, AT LT 5. Chair: Michael Bowling

10:30–10:50 *Agnostic System Identification for Model-Based Reinforcement Learning*
Stephane Ross, Drew Bagnell (p. 143)

10:50–11:10 *Greedy Algorithms for Sparse Reinforcement Learning*
Christopher Painter-Wakefield, Ronald Parr (p. 143)

11:10–11:30 *On the Sample Complexity of Reinforcement Learning with a Generative Model*
Mohammad Gheshlaghi Azar, Remi Munos, Bert Kappen (p. 144)

11:30–11:50 *Artist Agent: A Reinforcement Learning Approach to Automatic Stroke Generation in Oriental Ink Painting*
Ning Xie, Hirotaka Hachiya, Masashi Sugiyama (p. 144)

11:50–12:10 *Path Integral Policy Improvement with Covariance Matrix Adaptation*
Freek Stulp, Olivier Sigaud (p. 144)

7C: Clustering 2

10:30–12:10, AT LT 1. Chair: Shai Ben-David

10:30–10:50 *On causal and anticausal learning*
Bernhard Schoelkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, Joris Mooij (p. 145)

10:50–11:10 *Revisiting k-means: New Algorithms via Bayesian Nonparametrics*
Brian Kulis, Michael Jordan (p. 145)

11:10–11:30 *Approximate Principal Direction Trees*
Mark McCartin-Lim, Andrew McGregor, Rui Wang (p. 146)

11:30–11:50 *Clustering using Max-norm Constrained Optimization*
Ali Jalali, Nathan Srebro (p. 146)

11:50–11:55 *Efficient Active Algorithms for Hierarchical Clustering*
Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, Aarti Singh (p. 146)

11:55–12:00 *Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients*
Iftekhar Naim, Daniel Gildea (p. 146)

12:00–12:05 *Groupwise Constrained Reconstruction for Subspace Clustering*
Ruijiang Li, Bin Li, Cheng Jin, Xiangyang Xue (p. 147)

12:05–12:10 *Clustering by Low-Rank Doubly Stochastic Matrix Decomposition*
Zhirong Yang, Erkki Oja (p. 147)

7D: Supervised learning 2

10:30–12:10, AT LT 2. Chair: Leon Bottou

10:30–10:50 *Total Variation and Euler's Elastica for Supervised Learning*
Tong Lin, Hanlin Xue, Ling Wang, Hongbin Zha (p. 148)

10:50–11:10 *A General Framework for Inferring Latent Task Structures*
Alexandre Passos, Piyush Rai, Jacques Wainer, Hal Daume III (p. 148)

11:10–11:30 *Fast classification using sparse decision DAGs*
Robert Busa-Fekete, Djalel Benbouzid, Balazs Kegl (p. 148)

11:30–11:50 *An Efficient Approach to Sparse Linear Discriminant Analysis*
Luis Francisco Sánchez Merchante, Yves Grandvalet, Gérrad Govaert (p. 149)

7E: Probabilistic Models

10:30–12:10, AT LT 3. Chair: Erik Sudderth

8A: Kernel methods 2

14:00–15:40, AT LT 4. Chair: Mario Marchand

14:20–14:40 *The Kernelized Stochastic Batch Perceptron*
Andrew Cotter, Shai Shalev-Schwartz, Nathan Srebro (p. 154)

14:40–15:00 *Fast Bounded Online Gradient Descent Algorithms for Scalable Kernel-Based Online Learning*
Steven C.H. Hoi, Jialei Wang, Peilin Zhao, Rong Jin, Pengcheng Wu (p. 154)

15:00–15:20 *Distributed Tree Kernels*
Fabio Massimo Zanzotto, Lorenzo Dell'Arciprete (p. 155)

15:20–15:40 *Analysis of Kernel Mean Matching under Covariate Shift*
Yaoliang Yu, Csaba Szepesvari (p. 155)

8B: Active and cost-sensitive learning

14:00–15:35, AT LT 5. Chair: Andreas Krause

14:00–14:20 *The Greedy Miser: Learning under Test-time Budgets*
Zhixiang Xu, Kilian Weinberger, Olivier Chapelle (p. 155)

14:20–14:40 *Joint Optimization and Variable Selection of High-dimensional Gaussian Processes*
Bo Chen, Rui Castro, Andreas Krause (p. 156)

14:40–15:00 *Comparison-Based Learning with Rank Nets*
Amin Karbasi, Stratis Ioannidis, laurent Massoulie (p. 156)

15:00–15:20 *Bayesian Optimal Active Search and Surveying*
Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, Richard Mann (p. 156)

15:20–15:25 *Hybrid Batch Bayesian Optimization*
Javad Azimi, Ali Jalali, Xiaoli Zhang-Fern (p. 157)

15:25–15:30 *Batch Active Learning via Coordinated Matching*
Javad Azimi, Alan Fern, Xiaoli Zhang-Fern, Glencora Borradaile , Brent Heeringa (p. 157)

15:30–15:35 *Bayesian Nonexhaustive Learning for Online Discovery and Modeling of Emerging Classes*
Murat Dundar, Ferit Akova, Alan Qi, Bartek Rajwa (p. 158)

8C: Feature selection and dimensionality reduction

14:00–15:40, AT LT 1. Chair: Andrea Danyluk

14:00–14:20 *Robust PCA in High-dimension: A Deterministic Approach*
Jiashi Feng, Huan Xu, Shuicheng Yan (p. 158)

14:20–14:40 *Communications Inspired Linear Discriminant Analysis*
Minhua Chen, William Carson, Miguel Rodrigues, Lawrence Carin, Robert Calderbank (p. 158)

8D: Recommendation and Matrix Factorization

14:00–15:35, AT LT 2. Chair: Thorsten Joachims

15:35–15:40 *A Graphical Model Formulation of Collaborative Filtering Neighbourhood Methods with Fast Maximum Entropy Training*
Aaron Defazio, Tiberio Caetano (p. 163)

8E: Graphical models

14:00–15:40, AT LT 3. Chair: Ricardo Silva

14:00–14:20 *Variational Bayesian Inference with Stochastic Search*
John Paisley, David Blei, Michael Jordan (p. 163)

14:20–14:40 *Large Scale Variational Bayesian Inference for Structured Scale Mixture Models*
Young Jun Ko, Matthias Seeger (p. 164)

14:40–15:00 *A Generalized Loop Correction Method for Approximate Inference in Graphical Models*
Siamak Ravanbakhsh, Chun-Nam Yu, Russell Greiner (p. 164)

15:00–15:20 *Distributed Parameter Estimation via Pseudo-likelihood*
Qiang Liu, alexander ihler (p. 164)

15:20–15:40 *High-Dimensional Covariance Decomposition into Sparse Markov and Independence Domains*
Majid Janzamin, Animashree Anandkumar (p. 165)

Plenary Session

16:10–17:40

16:10–17:10 *Invited talk: Learning Hierarchies of Invariant Features*
Yann LeCun (p. 74)

17:10–17:40 *Applications talk: Machine Learning that Matters*
Kiri Wagstaff (p. 165)

Poster Session

17:40, AT and IF

Posters 1–48 are in the Informatics Forum (IF). Posters 49–96 are in Appleton Tower (AT). See floor plans at end of book.

(1) *A convex relaxation for weakly supervised classifiers*
Armand Joulin, Francis Bach (p. 132)

(2) *Semi-Supervised Learning of Class Balance under Class-Prior Change by Distribution Matching*
Marthinus Du Plessis, Masashi Sugiyama (p. 132)

(3) *A Simple Algorithm for Semi-supervised Learning with Improved Generalization Error Bound*
Ming Ji, Tianbao Yang, Binbin Lin, Rong Jin, Jiawei Han (p. 132)

(4) *Information-theoretic Semi-supervised Metric Learning via Entropy Regularization*
Gang Niu, Bo Dai, Makoto Yamada, Masashi Sugiyama (p. 133)

(5) *Cross Language Text Classification via Subspace Co-regularized Multi-view Learning*
Yuhong Guo, Min Xiao (p. 133)

(6) *Using CCA to improve CCA: A new spectral method for estimating vector models of words*
Paramveer Dhillon, Jordan Rodu, Dean Foster, Lyle Ungar (p. 133)

(7) *Semi-Supervised Collective Classification via Hybrid Label Regularization*
Luke McDowell, David Aha (p. 134)

(8) *Compositional Planning Using Optimal Option Models*
David Silver, Kamil Ciosek (p. 134)

(9) *Learning Parameterized Skills*
Bruno Da Silva, George Konidaris, Andrew Barto (p. 134)

(10) *Safe Exploration in Markov Decision Processes*
Teodor Mihai Moldovan, Pieter Abbeel (p. 135)

(11) *Modelling transition dynamics in MDPs with RKHS embeddings*
Steffen Grunewalder, Guy Lever, Luca Baldassarre, Massi Pontil, Arthur Gretton (p. 135)

(12) *A Joint Model of Language and Perception for Grounded Attribute Learning*
Cynthia Matuszek, Nichoals FitzGerald, Luke Zettlemoyer, Liefeng Bo, Dieter Fox (p. 136)

(13) *Predicting Manhole Events in New York City*
Cynthia Rudin, Rebecca Passonneau, Axinia Radeva, Steve Ierome, Delfina Isaac (p. 136)

(28) *Data-driven Web Design*
Ranjitha Kumar, Jerry Talton, Salman Ahmad, Scott Klemmer (p. 142)

(29) *Learning the Central Events and Participants in Unlabeled Text*
Nathanael Chambers, Dan Jurafsky (p. 142)

(30) *Exemplar-SVMs for Visual Object Detection, Label Transfer and Image Retrieval*
Tomasz Malisiewicz, Abhinav Shrivastava, Abhinav Gupta, Alexei Efros (p. 142)

(31) *Learning Force Control Policies for Compliant Robotic Manipulation*
Mrinal Kalakrishnan, Ludovic Righetti, Peter Pastor, Stefan Schaal (p. 142)

(32) *Agnostic System Identification for Model-Based Reinforcement Learning*
Stephane Ross, Drew Bagnell (p. 143)

(33) *Greedy Algorithms for Sparse Reinforcement Learning*
Christopher Painter-Wakefield, Ronald Parr (p. 143)

(34) *On the Sample Complexity of Reinforcement Learning with a Generative Model*
Mohammad Gheshlaghi Azar, Remi Munos, Bert Kappen (p. 144)

(35) *Artist Agent: A Reinforcement Learning Approach to Automatic Stroke Generation in Oriental Ink Painting*
Ning Xie, Hirotaka Hachiya, Masashi Sugiyama (p. 144)

(36) *Path Integral Policy Improvement with Covariance Matrix Adaptation*
Freek Stulp, Olivier Sigaud (p. 144)

(37) *On causal and anticausal learning*
Bernhard Schoelkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, Joris Mooij (p. 145)

(38) *Revisiting k-means: New Algorithms via Bayesian Nonparametrics*
Brian Kulis, Michael Jordan (p. 145)

(39) *Approximate Principal Direction Trees*
Mark McCartin-Lim, Andrew McGregor, Rui Wang (p. 146)

(40) *Clustering using Max-norm Constrained Optimization*
Ali Jalali, Nathan Srebro (p. 146)

(85) *A Bayesian Approach to Approximate Joint Diagonalization of Square Matrices*
Mingjun Zhong (p. 162)

(86) *Collaborative Topic Regression with Social Matrix Factorization for Recommendation Systems*
Sanjay Purushotham, Yan Liu (p. 162)

(87) *Active Learning for Matching Problems*
Laurent Charlin, Rich Zemel, Craig Boutilier (p. 163)

(88) *A Graphical Model Formulation of Collaborative Filtering Neighbourhood Methods with Fast Maximum Entropy Training*
Aaron Defazio, Tiberio Caetano (p. 163)

(89) *Variational Bayesian Inference with Stochastic Search*
John Paisley, David Blei, Michael Jordan (p. 163)

(90) *Large Scale Variational Bayesian Inference for Structured Scale Mixture Models*
Young Jun Ko, Matthias Seeger (p. 164)

(91) *A Generalized Loop Correction Method for Approximate Inference in Graphical Models*
Siamak Ravanbakhsh, Chun-Nam Yu, Russell Greiner (p. 164)

(92) *Distributed Parameter Estimation via Pseudo-likelihood*
Qiang Liu, alexander ihler (p. 164)

(93) *High-Dimensional Covariance Decomposition into Sparse Markov and Independence Domains*
Majid Janzamin, Animashree Anandkumar (p. 165)

(94) *Machine Learning that Matters*
Kiri Wagstaff (p. 165)

7 Invited Talks

Test of Time Award: Risi Kondor and John Lafferty: Diffusion Kernels on Graphs and Other Discrete Input Spaces

Risi Kondor. University of Chicago

The application of kernel-based learning algorithms has, so far, largely been confined to real-valued data and a few special data types, such as strings. In this paper we propose a general method of constructing natural families of kernels over discrete structures, based on the matrix exponentiation idea. In particular, we focus on generating kernels on graphs, for which we propose a special class of exponential kernels called diffusion kernels, which are based on the heat equation and can be regarded as the discretization of the familiar Gaussian kernel of Euclidean space.

page 46

Algorithmic Phase Transitions in Constraint Satisfaction Problems

Dimitris Achlioptas. CTI & UC Santa Cruz

Constraint Satisfaction Problems (CSPs) are the common abstraction of real-life problems ranging from air-traffic control to protein folding. Their ubiquity presses forward a fundamental question: why are certain CSP instances exceptionally hard while other, seemingly similar, instances are quite easy? To study this phenomenon we consider probability distributions over CSP instances formed by adding constraints one-by-one, uniformly and independently. We will see that for many random CSPs there exists a broad regime in which exponentially many solutions provably exist, yet all known efficient algorithms fail to find even one solution. To understand the origin of this algorithmic barrier we examine how the solution-space geometry evolves as constraints are added. We prove in a precise mathematical sense that the

barrier faced by algorithms corresponds to a phase transition in solution-space geometry: at some point, the set of solutions shatters and goes from resembling a single giant ball to exponentially many clusters of well-separated solutions. More speculatively, we will also discuss how this shattering phenomenon can perhaps be used to reframe the clustering problem in machine learning using ideas from statistical physics.

page 28

Learning Hierarchies of Invariant Features

Yann LeCun. Courant Institute of Mathematical Sciences and Center for Neural Science, NYU

Intelligent perceptual tasks such as vision and audition require the construction of good internal representations. Machine Learning has been very successful for producing classifiers, but the next big challenge for ML, computer vision, and computational neuroscience is to devise learning algorithms that can learn features and internal representations automatically.

Theoretical and empirical evidence suggest that the perceptual world is best represented by a multi-stage hierarchy in which features in successive stages are increasingly global, invariant, and abstract. An important question is to devise "deep learning" methods for multi-stage architecture than can automatically learn invariant feature hierarchies from labeled and unlabeled data.

A number of unsupervised methods for learning invariant features will be described that are based on sparse coding and sparse auto-encoders: convolutional sparse auto-encoders, invariance through group sparsity, invariance through lateral inhibition, and invariance through temporal constancy. The methods are used to pre-train convolutional networks (ConvNets). ConvNets are biologically-inspired architectures consisting of multiple stages of filter banks, interspersed with non-linear operations, spatial pooling, and contrast normalization operations.

Several applications will be shown through videos and live demos, including a a pedestrian detector, a category-level object recognition system that can be trained on the fly, and a system that can label every pixel in an image with the category of the object it belongs to (scene parsing). Specialized hardware architecture that run these systems in real time will also be described.

page 65

Modern Algorithmic Tools for Analyzing Data Streams

Sethu Muthukrishnan. Professor of Computer Science, Rutgers University

We now have a second generation of algorithmic tools for analyzing data streams, that go beyond the initial tools for summarizing a single stream in small space. The new tools deal with distributed data, stochastic models, dynamic graph and matrix objects and others; they optimize communication, number of parallel rounds and privacy among other things. I will provide an overview of these tools and explore their application to Machine Learning problems.

page 26

Information Theory and Sustainable Energy

David MacKay FRS. Chief Scientific Advisor, DECC
Professor of Natural Philosophy, University of Cambridge

How easy is it to get off our fossil fuel habit? Can European countries live on their own renewables? What do the fundamental limits of physics say? How does our current energy consumption compare with our sustainable energy options? This talk will offer a straight-talking assessment of the numbers; will discuss how to make energy plans that add up; and will hunt for connections between machine learning, climate change, and sustainable energy.

page 45

8 Main Conference Abstracts

Main Conference Abstracts: Wednesday, June 27

Plenary Session, 09:40–10:00

Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring

Sungjin Ahn, Anoop Korattikara, Max Welling

In this paper we address the following question: "Can we approximately sample from a Bayesian posterior distribution if we are only allowed to touch a small mini-batch of data-items for every sample we generate?". An algorithm based on the Langevin equation with stochastic gradients (SGLD) was previously proposed to solve this, but its mixing rate was slow. By leveraging the Bayesian Central Limit Theorem, we extend the SGLD algorithm so that at high mixing rates it will sample from a normal approximation of the posterior, while for slow mixing rates it will mimic the behavior of SGLD with a pre-conditioner matrix. As a bonus, the proposed algorithm is reminiscent of Fisher scoring (with stochastic gradients) and as such an efficient optimizer during burn-in.

1A: Optimization algorithms 1, 10:30–12:10, AT LT 4

On the Equivalence between Herding and Conditional Gradient Algorithms

Francis Bach, Simon Lacoste-Julien, Guillaume Obozinski

We show the equivalence of the herding procedure of Welling (2009) with a standard convex optimization algorithm–namely a conditional gradient algorithm minimizing a quadratic moment discrepancy. This link enables us to harness convergence results from convex optimization as well as suggest faster alternatives for the task of approximating integrals in a reproducing kernel Hilbert space. We study the behavior of the different variants with numerical simulations. The experiments indicate that while we can improve on the task of approximating integrals, the original herding algorithm tends to approach more often the maximum entropy distribution, shedding more light on the learning bias behind herding.

Similarity Learning for Provably Accurate Sparse Linear Classification

Aurélien Bellet, Amaury Habrard, Marc Sebban

In recent years, the crucial importance of metrics in machine learning algorithms has led to an increasing interest for optimizing distance and similarity functions. Most of the state of the art focus on learning Mahalanobis distances (requiring to fulfill a constraint of positive semi-definiteness) for use in a local k-NN algorithm. However, no theoretical link is established between the learned metrics and their performance in classification. In this paper, we make use of the formal framework of good similarities introduced by Balcan et al. to propose an algorithm for learning a non PSD linear similarity optimized in a nonlinear feature space, tailored to a use in a global linear classifier. We show that our approach has the property of stability, allowing us to derive a generalization bound on the classification error. Experiments performed on various datasets confirm the effectiveness of our approach compared to state-of-the-art methods and provide evidence that (i) it is fast, (ii) robust to overfitting and (iii) produces very sparse models.

Stochastic Smoothing for Nonsmooth Minimizations: Accelerating SGD by Exploiting Structure

Hua Ouyang, Alexander Gray

In this work we consider the stochastic minimization of nonsmooth convex loss functions, a central problem in machine learning. We propose a novel algorithm called Accelerated Nonsmooth Stochastic Gradient Descent (ANSGD), which exploits the structure of common nonsmooth loss functions to achieve optimal convergence rates for a class of problems including SVMs. It is the first stochastic algorithm that can achieve the optimal O(1/t) rate for minimizing nonsmooth loss functions. The fast rates are confirmed by empirical comparisons, in which ANSGD significantly outperforms previous subgradient descent algorithms including SGD.

To Average or Not to Average? Making Stochastic Gradient Descent Optimal for Strongly Convex Problems

Alexander Rakhlin, Ohad Shamir, Karthik Sridharan

Stochastic gradient descent (SGD) is a simple and popular method to solve stochastic optimization problems which arise in machine learning. For strongly convex problems, its convergence rate was known to be O(log(T)/T), by running SGD for T iterations and returning the average point. However, recent results showed that using a different algorithm, one can get an optimal O1/T) rate. This might lead one to believe that standard SGD is suboptimal, and maybe should even be replaced as a method of choice. In this paper, we investigate the optimality of SGD in a stochastic setting. We show that for smooth problems, the algorithm attains the optimal O(1/T) rate. However, for non-smooth problems, the convergence rate with averaging might really

be $\Omega(\log(T)/T)$, and this is not just an artifact of the analysis. On the flip side, we show that a simple modification of the averaging step suffices to recover the $O(1/T)$ rate, and no other change of the algorithm is necessary. We also present experimental results which support our findings, and point out open problems.

Scaling Up Coordinate Descent Algorithms for Large ℓ_1 Regularization Problems

Chad Scherrer, Mahantesh Halappanavar, Ambuj Tewari, David Haglin

We present a generic framework for parallel coordinate descent (CD) algorithms that has as special cases the original sequential algorithms of Cyclic CD and Stochastic CD, as well as the recent parallel Shotgun algorithm of Bradley et al. We introduce two novel parallel algorithms that are also special cases—Thread-Greedy CD and Coloring-Based CD—and give performance measurements for an OpenMP implementation of these.

Quasi-Newton Methods: A New Direction

Philipp Hennig, Martin Kiefel

Four decades after their invention, quasi-Newton methods are still state of the art in unconstrained numerical optimization. Although not usually interpreted thus, these are learning algorithms that fit a local quadratic approximation to the objective function. We show that many, including the most popular, quasi-Newton methods can be interpreted as approximations of Bayesian linear regression under varying prior assumptions. This new notion elucidates some shortcomings of classical algorithms, and lights the way to a novel nonparametric quasi-Newton method, which is able to make more efficient use of available information at computational cost similar to its predecessors.

A Hybrid Algorithm for Convex Semidefinite Optimization

Soeren Laue

We present a hybrid algorithm for optimizing a convex, smooth function over the cone of positive semidefinite matrices. Our algorithm converges to the global optimal solution and can be used to solve general large-scale semidefinite programs and hence can be readily applied to a variety of machine learning problems. We show experimental results on three machine learning problems (matrix completion, metric learning, and sparse PCA) . Our approach outperforms state-of-the-art algorithms.

Efficient and Practical Stochastic Subgradient Descent for Nuclear Norm Regularization

Haim Avron, Satyen Kale, Shiva Kasiviswanathan, Vikas Sindhwani

We describe novel subgradient methods for a broad class of matrix optimization problems involving nuclear norm regularization. Unlike existing approaches, our method executes very cheap iterations by combining low-rank stochastic subgradients with

efficient incremental SVD updates, made possible by highly optimized and parallelizable dense linear algebra operations on small matrices. Our practical algorithms always maintain a low-rank factorization of iterates that can be conveniently held in memory and efficiently multiplied to generate predictions in matrix completion settings. Empirical comparisons confirm that our approach is highly competitive with several recently proposed state-of-the-art solvers for such problems.

1B: Reinforcement learning 1, 10:30–12:10, AT LT 5

Policy Gradients with Variance Related Risk Criteria

Dotan Di Castro, Aviv Tamar, Shie Mannor

Managing risk in dynamic decision problems is of cardinal importance in many fields such as finance and process control. The most common approach to defining risk is through various variance related criteria such as the Sharpe Ratio or the standard deviation adjusted reward. It is known that optimizing many of the variance related risk criteria is NP-hard. In this paper we devise a framework for local policy gradient style algorithms for reinforcement learning for variance related criteria. Our starting point is a new formula for the variance of the cost-to-go in episodic tasks. Using this formula we develop policy gradient algorithms for criteria that involve both the expected cost and the variance of the cost. We prove the convergence of these algorithms to local minima and demonstrate their applicability in a portfolio planning problem.

Approximate Dynamic Programming By Minimizing Distributionally Robust Bounds

Marek Petrik

Approximate dynamic programming is a popular method for solving large Markov decision processes. This paper describes a new class of approximate dynamic programming (ADP) methods—distributionally robust ADP—that address the curse of dimensionality by minimizing a pessimistic bound on the policy loss. This approach turns ADP into an optimization problem, for which we derive new mathematical program formulations and analyze itsp properties. DRADP improves on the theoretical guarantees of existing ADP methods—it guarantees convergence and L_1 norm-based error bounds. The empirical evaluation of DRADP shows that the theoretical guarantees translate well into good performance on benchmark problems.

Statistical linear estimation with penalized estimators: an application to reinforcement learning

Bernardo Avila Pires, Csaba Szepesvari

Motivated by value function estimation in reinforcement learning, we study statistical linear inverse problems, i.e., problems where the coefficients of a linear system

to be solved are observed in noise. We consider penalized estimators, where performance is evaluated using a matrix-weighted two-norm of the defect of the estimator measured with respect to the true, unknown coefficients. Two objective functions are considered depending whether the error of the defect measured with respect to the noisy coefficients is squared or unsquared. We propose simple, yet novel and theoretically well-founded data-dependent choices for the regularization parameters for both cases that avoid data-splitting. A distinguishing feature of our analysis is that we derive deterministic error bounds in terms of the error of the coefficients, thus allowing the complete separation of the analysis of the stochastic properties of these errors. We show that our results lead to new insights and bounds for linear value function estimation in reinforcement learning.

Approximate Modified Policy Iteration

Bruno Scherrer, Victor Gabillon, Mohammad Ghavamzadeh, Matthieu Geist

Modified policy iteration (MPI) is a dynamic programming (DP) algorithm that contains the two celebrated policy and value iteration methods. Despite its generality, MPI has not been thoroughly studied, especially its approximation form which is used when the state and/or action spaces are large or infinite. In this paper, we propose three approximate MPI (AMPI) algorithms that are extensions of the well-known approximate DP algorithms: fitted-value iteration, fitted-Q iteration, and classification-based policy iteration. We provide an error propagation analysis for AMPI that unifies those for approximate policy and value iteration. We also provide a finite-sample analysis for the classification-based implementation of AMPI (CBMPI), which is more general (and somehow contains) than the analysis of the other presented AMPI algorithms. An interesting observation is that the MPI's parameter allows us to control the balance of errors (in value function approximation and in estimating the greedy policy) in the final performance of the CBMPI algorithm.

A Dantzig Selector Approach to Temporal Difference Learning

Matthieu Geist, Bruno Scherrer, Alessandro Lazaric, Mohammad Ghavamzadeh

LSTD is one of the most popular reinforcement learning algorithms for value function approximation. Whenever the number of samples is larger than the number of features, LSTD must be paired with some form of regularization. In particular, L1-regularization methods tends to perform feature selection by promoting sparsity and thus they are particularly suited in high–dimensional problems. Nonetheless, since LSTD is not a simple regression algorithm but it solves a fixed–point problem, the integration with L1-regularization is not straightforward and it might come with some drawbacks (see e.g., the P-matrix assumption for LASSO-TD). In this paper we introduce a novel algorithm obtained by integrating LSTD with the Dantzig Selector. In

particular, we investigate the performance of the algorithm and its relationship with existing regularized approaches, showing how it overcomes some of the drawbacks of existing solutions.

Linear Off-Policy Actor-Critic

Thomas Degris, Martha White, Richard Sutton

This paper presents the first off-policy actor-critic reinforcement learning algorithm with a per-time-step complexity that scales linearly with the number of learned parameters. Previous work on actor-critic algorithms is limited to the on-policy setting and does not take advantage of the recent advances in off-policy gradient temporal-difference learning. Off-policy techniques, such as Greedy-GQ, enable a target policy to be learned while following and obtaining data from another (behavior) policy. For many problems, however, actor-critic methods are more practical than action value methods (like Greedy-GQ) because they explicitly represent the policy; consequently, the policy can be stochastic and utilize a large action space. In this paper, we illustrate how to practically combine the generality and learning potential of off-policy learning with the flexibility in action selection given by actor-critic methods. We derive an incremental, linear time and space complexity algorithm that includes eligibility traces, prove convergence under assumptions similar to previous off-policy algorithms, and empirically show better or comparable performance to existing algorithms on standard reinforcement-learning benchmark problems.

Lightning Does Not Strike Twice: Robust MDPs with Coupled Uncertainty

Shie Mannor, Ofir Mebel, Huan Xu

We consider Markov decision processes under parameter uncertainty. Previous studies all restrict to the case that uncertainties among different states are uncoupled, which leads to conservative solutions. In contrast, we introduce an intuitive concept, termed 'Lightning Does not Strike Twice,' to model coupled uncertain parameters. Specifically, we require that the system can deviate from its nominal parameters only a bounded number of times. We give probabilistic guarantees indicating that this model represents real life situations and devise tractable algorithms for computing optimal control policies using this concept.

Bounded Planning in Passive POMDPs

Roy Fox, Naftali Tishby

In Passive POMDPs actions do not affect the world state, but still incur costs. When the agent is bounded by information-processing constraints, it can only keep an approximation of the belief. We present a variational principle for the problem of maintaining the information which is most useful for minimizing the cost, and introduce an efficient and simple algorithm for finding an optimum.

1C: Neural networks and deep learning 1, 10:30–12:10, AT LT 1

Scene parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers

Clément Farabet, Camille Couprie, Laurent Najman, Yann LeCun

Scene parsing consists in labeling each pixel in an image with the category of the object it belongs to. We propose a method that uses a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centered on each pixel. The method alleviates the need for engineered features. In parallel to feature extraction, a tree of segments is computed from a graph of pixel dissimilarities. The feature vectors associated with the segments covered by each node in the tree are aggregated and fed to a classifier which produces an estimate of the distribution of object categories contained in the segment. A subset of tree nodes that cover the image are then selected so as to maximize the average 'purity' of the class distributions, hence maximizing the overall likelihood that each segment will contain a single object. The system yields record accuracies on the the Sift Flow Dataset (33 classes) and the Barcelona Dataset (170 classes) and near-record accuracy on Stanford Background Dataset (8 classes), while being an order of magnitude faster than competing approaches, producing a 320x240 image labeling in less than 1 second, including feature extraction.

A Generative Process for Contractive Auto-Encoders

Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio

The contractive auto-encoder learns a representation of the input data that captures the local manifold structure around each data point, through the leading singular vectors of the Jacobian of the transformation from input to representation. The corresponding singular values specify how much local variation is plausible in directions associated with the corresponding singular vectors, while remaining in a high-density region of the input space. This paper proposes a procedure for generating samples that are consistent with the local structure captured by a contractive auto-encoder. The associated stochastic process defines a distribution from which one can sample, and which experimentally appears to converge quickly and mix well between modes, compared to Restricted Boltzmann Machines and Deep Belief Networks. The intuitions behind this procedure can also be used to train the second layer of contraction that pools lower-level features and learns to be invariant to the local directions of variation discovered in the first layer. We show that this can help learn and represent invariances present in the data and improve classification error.

Deep Lambertian Networks

Yichuan Tang, Ruslan Salakhutdinov, Geoffrey Hinton

Visual perception is a challenging problem in part due to illumination variations. A

possible solution is to first estimate an illumination invariant representation before using it for recognition. The object albedo and surface normals are examples of such representation. In this paper, we introduce a multilayer generative model where the latent variables include the albedo, surface normals, and the light source. Combining Deep Belief Nets with the Lambertian reflectance assumption, our model can learn good priors over the albedo from 2D images. Illumination variations can be explained by changing only the lighting latent variable in our model. By transferring learned knowledge from similar objects, albedo and surface normals estimation from a *single* image is possible in our model. Experiments demonstrate that our model is able to generalize as well as improve over standard baselines in *one-shot* face recognition.

Deep Mixtures of Factor Analysers

Yichuan Tang, Ruslan Salakhutdinov, Geoffrey Hinton

An efficient way to learn deep density models that have many layers of latent variables is to learn one layer at a time using a model that has only one layer of latent variables. After learning each layer, samples from the posterior distributions for that layer are used as training data for learning the next layer. This approach is commonly used with Restricted Boltzmann Machines, which are *undirected* graphical models with a single hidden layer, but it can also be used with Mixtures of Factor Analysers (MFAs) which are *directed* graphical models. In this paper, we present a greedy layer-wise learning algorithm for Deep Mixtures of Factor Analysers (DMFAs). Even though a DMFA can be converted to an equivalent shallow MFA by multiplying together the factor loading matrices at different levels, learning and inference are much more efficient in a DMFA and the sharing of each lower-level factor loading matrix by many different higher level MFAs prevents overfitting. We demonstrate empirically that DMFAs learn better density models than both MFAs and two types of Restricted Boltzmann Machines on a wide variety of datasets.

Utilizing Static Analysis and Code Generation to Accelerate Neural Networks

Lawrence McAfee, Kunle Olukotun

As datasets continue to grow, neural network (NN) applications are becoming increasingly limited by both the amount of available computational power and the ease of developing high-performance applications. Researchers often must have expert systems knowledge to make their algorithms run efficiently. Although available computing power approximately doubles every two years, algorithm efficiency is not able to keep pace due to the use of general purpose compilers, which are not able to fully optimize specialized application domains. Within the domain of NNs, we have the added knowledge that network architecture remains constant during training, meaning the architecture's data structure can be statically optimized by a compiler.

In this paper, we present SONNC, a compiler for NNs that utilizes static analysis to generate optimized parallel code. We show that SONNC's use of static optimizations make it able to outperform optimized MATLAB code by up to 242X. Additionally, we show that use of SONNC significantly reduces code complexity when using structurally sparse networks.

Estimating the Hessian by Back-propagating Curvature

James Martens, Ilya Sutskever, Kevin Swersky

In this work we develop Curvature Propagation (CP), a general technique for efficiently computing unbiased approximations of the Hessian of any function that is computed using a computational graph. At the cost of roughly two gradient evaluations, CP can give a rank-1 approximation of the whole Hessian, and can be repeatedly applied to give increasingly precise unbiased estimates of any or all of the entries of the Hessian. Of particular interest is the diagonal of the Hessian, for which no general approach is known to exist that is both efficient and accurate. We show in experiments that CP turns out to work well in practice, giving very accurate estimates of the Hessian of neural networks, for example, with a relatively small amount of work. We also apply CP to Score Matching, where a diagonal of a Hessian plays an integral role in the Score Matching objective, and where it is usually computed exactly using inefficient algorithms which do not scale to larger and more complex models.

Training Restricted Boltzmann Machines on Word Observations

George Dahl, Ryan Adams, Hugo Larochelle

The restricted Boltzmann machine (RBM) is a flexible tool for modeling complex data, however there have been significant computational difficulties in using RBMs to model high-dimensional multinomial observations. In natural language processing applications, words are naturally modeled by K-ary discrete distributions, where K is determined by the vocabulary size and can easily be in the hundreds of thousands. The conventional approach to training RBMs on word observations is limited because it requires sampling the states of K-way softmax visible units during block Gibbs updates, an operation that takes time linear in K. In this work, we address this issue by employing a more general class of Markov chain Monte Carlo operators on the visible units, yielding updates with computational complexity independent of K. We demonstrate the success of our approach by training RBMs on hundreds of millions of word n-grams using larger vocabularies than previously feasible and using the learned features to improve performance on chunking and sentiment classification tasks, achieving state-of-the-art results on the latter.

A fast and simple algorithm for training neural probabilistic language models

Andriy Mnih, Yee Whye Teh

Neural probabilistic language models (NPLMs) have recently superseded smoothed n-gram models as the best-performing model class for language modelling. Unfortunately, the adoption of NPLMs is held back by their notoriously long training times, which can be measured in weeks even for moderately-sized datasets. These are a consequence of the models being explicitly normalized, which leads to having to consider all words in the vocabulary when computing the log-likelihood gradients. We propose a fast and simple algorithm for training NPLMs based on noise-contrastive estimation, a newly introduced procedure for estimating unnormalized continuous distributions. We investigate the behaviour of the algorithm on the Penn Treebank corpus and show that it reduces the training times by more than an order of magnitude without affecting the quality of the resulting models. The algorithm is also more efficient and much more stable than importance sampling because it requires far fewer noise samples to perform well. We demonstrate the scalability of the proposed approach by training several neural language models on a 47M-word corpus with a 80K-word vocabulary, obtaining state-of-the-art results in the Microsoft Research Sentence Completion Challenge.

1D: Structured output prediction, 10:30–12:10, AT LT 2

Learning to Identify Regular Expressions that Describe Email Campaigns

Paul Prasse, Christoph Sawade, Niels Landwehr, Tobias Scheffer

This paper addresses the problem of inferring a regular expression from a given set of strings that resembles, as closely as possible, the regular expression that a human expert would have written to identify the language. This is motivated by our goal of automating the task of postmasters of an email service who use regular expressions to describe and blacklist email spam campaigns. Training data contains batches of messages and corresponding regular expressions that an expert postmaster feels confident to blacklist. We model this task as a learning problem with structured output spaces and an appropriate loss function, derive a decoder and the resulting optimization problem, and a report on a case study conducted with an email service.

Efficient Structured Prediction with Latent Variables for General Graphical Models

Alexander Schwing, Tamir Hazan, Marc Pollefeys, Raquel Urtasun

In this paper we propose a unified framework for structured prediction with hidden variables which includes hidden conditional random fields and latent structural support vector machines as special cases. We describe an approximation for this general structured prediction formulation using duality, which is based on a local entropy approximation. We then derive an efficient message passing algorithm that

is guaranteed to converge, and demonstrate its effectiveness in the tasks of image segmentation as well as 3D indoor scene understanding from single images, showing that our approach is superior to latent structural support vector machines and hidden conditional random fields.

Output Space Search for Structured Prediction

Janardhan Rao Doppa, Alan Fern, Prasad Tadepalli

We consider a framework for structured prediction based on search in the space of complete structured outputs. Given a structured input, an output is produced by running a time-bounded search procedure, guided by a learned cost function, and then returning the least cost output uncovered during the search. This framework can be instantiated for a wide range of search spaces and search procedures, and easily incorporates arbitrary structured-prediction loss functions. In this paper, we make two main technical contributions within this framework. First, we describe a novel approach to automatically defining an effective search space over structured outputs, which is able to leverage the availability of powerful classification learning algorithms. In particular, we define the limited-discrepancy search space and relate the quality of that space to the quality of learned classifiers. Second, we give a generic cost function learning approach that can be instantiated for a wide range of search procedures. The key idea is to learn a cost function that attempts to mimic the behavior of conducting searches guided by the true loss function. Our experiments on six benchmark domains demonstrate that using our framework with only a small amount of search is sufficient for significantly improving on state-of-the-art structured-prediction performance.

Efficient Decomposed Learning for Structured Prediction

Rajhans Samdani, Dan Roth

Structured prediction is the cornerstone of several machine learning applications. Unfortunately, in structured prediction settings with expressive inter-variable interactions, inference and hence exact learning is often intractable. We present a new way, Decomposed Learning (DecL), for performing efficient learning over structured output spaces. In DecL, we restrict the inference step to a limited part of the output space. We use characterizations based on the structure, target parameters, and gold labels to guarantee that DecL with limited inference is equivalent to exact learning. We also show that in real world settings, where our theoretical assumptions may not hold exactly, DecL-based algorithms are significantly more efficient and provide accuracies close to exact learning.

Modeling Latent Variable Uncertainty for Loss-based Learning

M. Pawan Kumar, Ben Packer, Daphne Koller

We consider the problem of parameter estimation using weakly supervised datasets,

where a training sample consists of the input and a partially specified annotation (called the output). In addition, the missing information in the annotation is modeled using latent variables. Traditional methods, such as expectation-maximization, overburden a single distribution with two separate tasks: (i) modeling the uncertainty in the latent variables during training; and (ii) making accurate predictions for the output and the latent variables during testing. We propose a novel framework that separates the demands of the two tasks using two distributions: (i) a conditional distribution to model the uncertainty of the latent variables for a given input-output pair; and (ii) a delta distribution to predict the output and the latent variables for a given input. During learning, we encourage agreement between the two distributions by minimizing a loss-based dissimilarity coefficient. Our approach generalizes latent SVM in two important ways: (i) it models the uncertainty over latent variables instead of relying on a pointwise estimate; and (ii) it allows the use of loss functions that depend on latent variables, which greatly increases its applicability. We demonstrate the efficacy of our approach on two challenging problems—object detection and action detection—using publicly available datasets.

2A: Kernel methods 1, 16:00–17:40, AT LT 4

On the Size of the Online Kernel Sparsification Dictionary

Yi Sun, Faustino Gomez, Juergen Schmidhuber

We analyze the size of the dictionary constructed from online kernel sparsification, using a novel formula that expresses the expected determinant of the kernel Gram matrix in terms of the eigenvalues of the covariance operator. Using this formula, we are able to connect the cardinality of the dictionary with the eigen-decay of the covariance operator. In particular, we show that for bounded kernels, the size of the dictionary always grows sub-linearly in the number of data points, and, as a consequence, the kernel linear regressor constructed from the resulting dictionary is consistent.

Improved Nystrom Low-rank Decomposition with Priors

Kai Zhang, Liang Lan, Jun Liu, andreas Rauber

Low-rank matrix decomposition has gained great popularity recently in scaling up kernel methods to large amount of data. However, some limitations could prevent them from working effectively in certain domains. For example, many existing approaches are intrinsically unsupervised, which does not incorporate side information (e.g., class labels) to produce task specific decomposition; also, they typically work "transductively" and the factorization does not generalize to new samples, causing lot of inconvenience for updated environment. To solve these problems, in this paper we propose an "inductive"-flavored method for low-rank kernel decomposition with priors or side information. We achieve this by generalizing the Nyström method in a

novel way. On the one hand, our approach employs a highly flexible, nonparametric structure that allows us to generalize the low-rank factors to arbitrarily new samples; on the other hand, it has linear time and space complexities, which can be orders of magnitudes faster than existing approaches and renders great efficiency in learning a low-rank kernel decomposition. Empirical results demonstrate the efficacy and efficiency of the proposed method.

Bayesian Efficient Multiple Kernel Learning

Mehmet Gönen

Multiple kernel learning algorithms are proposed to combine kernels in order to obtain a better similarity measure or to integrate feature representations coming from different data sources. Most of the previous research on such methods is focused on the computational efficiency issue. However, it is still not feasible to combine many kernels using existing Bayesian approaches due to their high time complexity. We propose a fully conjugate Bayesian formulation and derive a deterministic variational approximation, which allows us to combine hundreds or thousands of kernels very efficiently. We briefly explain how the proposed method can be extended for multiclass learning and semi-supervised learning. Experiments with large numbers of kernels on benchmark data sets show that our inference method is quite fast, requiring less than a minute. On one bioinformatics and three image recognition data sets, our method outperforms previously reported results with better generalization performance.

A Binary Classification Framework for Two-Stage Multiple Kernel Learning

Abhishek Kumar, Alexandru Niculescu-Mizil, Koray Kavukcuoglu, Hal Daume III

With the advent of kernel methods, automating the task of specifying a suitable kernel has become increasingly important. In this context, the Multiple Kernel Learning (MKL) problem of finding a combination of pre-specified base kernels that is suitable for the task at hand has received significant attention from researchers. In this paper we show that Multiple Kernel Learning can be framed as a standard binary classification problem with additional constraints that ensure the positive definiteness of the learned kernel. Framing MKL in this way has the distinct advantage that it makes it easy to leverage the extensive research in binary classification to develop better performing and more scalable MKL algorithms that are conceptually simpler, and, arguably, more accessible to practitioners. Experiments on nine data sets from different domains show that, despite their simplicity, the proposed techniques can significantly outperform current leading MKL approaches.

Multiple Kernel Learning from Noisy Labels by Stochastic Programming

Tianbao Yang, Rong Jin, Yang Zhou, Mehrdad Mahdavi, Lijun Zhang

We study the problem of multiple kernel learning from noisy labels. This is in con-

trast to most of the previous studies on multiple kernel learning that mainly focus on developing efficient algorithms and assume perfectly labeled training examples. Directly applying the existing multiple kernel learning algorithms to noisily labeled examples often leads to suboptimal performance due to the incorrect class assignments. We address this challenge by casting multiple kernel learning from noisy labels into a stochastic programming problem, and presenting a minimax formulation. We develop an efficient algorithm for solving the related convex-concave optimization problem with a fast convergence rate of $O(1/T)$ where T is the number of iterations. Empirical studies on UCI data sets verify both the effectiveness of the proposed framework and the efficiency of the proposed optimization algorithm.

Subgraph Matching Kernels for Attributed Graphs

Nils Kriege, Petra Mutzel

We propose graph kernels based on subgraph matchings, i.e. structure-preserving bijections between subgraphs. While recently proposed kernels based on common subgraphs (Wale et al., 2008; Shervashidze et al., 2009) in general can not be applied to attributed graphs, our approach allows to rate mappings of subgraphs by a flexible scoring scheme comparing vertex and edge attributes by kernels. We show that subgraph matching kernels generalize several known kernels. To compute the kernel we propose a graph-theoretical algorithm inspired by a classical relation between common subgraphs of two graphs and cliques in their product graph observed by Levi (1973). Encouraging experimental results on a classification task of real-world graphs are presented.

Fast Computation of Subpath Kernel for Trees

Daisuke Kimura, Hisashi Kashima

The kernel method is a potential approach to analyzing structured data such as sequences, trees, and graphs; however, unordered trees have not been investigated extensively. Kimura et al. (2011) proposed a kernel function for unordered trees on the basis of their subpaths, which are vertical substructures of trees responsible for hierarchical information in them. Their kernel exhibits practically good performance in terms of accuracy and speed; however, linear-time computation is not guaranteed theoretically, unlike the case of the other unordered tree kernel proposed by Vishwanathan and Smola (2003). In this paper, we propose a theoretically guaranteed linear-time kernel computation algorithm that is practically fast, and we present an efficient prediction algorithm whose running time depends only on the size of the input tree. Experimental results show that the proposed algorithms are quite efficient in practice.

Hypothesis testing using pairwise distances and associated kernels

Dino Sejdinovic, Arthur Gretton, Bharath Sriperumbudur, Kenji Fukumizu

We provide a unifying framework linking two classes of statistics used in two-sample and independence testing: on one hand, the energy distances and distance covariances from the statistics literature; on the other, distances between embeddings of distributions to reproducing kernel Hilbert spaces (RKHS), as established in machine learning. The equivalence holds when energy distances are computed with semimetrics of negative type, in which case a kernel may be defined such that the RKHS distance between distributions corresponds exactly to the energy distance. We determine the class of probability distributions for which kernels induced by semimetrics are characteristic (that is, for which embeddings of the distributions to an RKHS are injective). Finally, we investigate the performance of this family of kernels in two-sample and independence tests: we show in particular that the energy distance most commonly employed in statistics is just one member of a parametric family of kernels, and that other choices from this family can yield more powerful tests.

2B: Reinforcement learning 2, 16:00–17:40, AT LT 5

No-Regret Learning in Extensive-Form Games with Imperfect Recall

Marc Lanctot, Richard Gibson, Neil Burch, Michael Bowling

Counterfactual Regret Minimization (CFR) is an efficient no-regret learning algorithm for decision problems modeled as extensive games. CFR's regret bounds depend on the requirement of perfect recall: players always remember information that was revealed to them and the order in which it was revealed. In games without perfect recall, however, CFR's guarantees do not apply. In this paper, we present the first regret bound for CFR when applied to a general class of games with imperfect recall. In addition, we show that CFR applied to any abstraction belonging to our general class results in a regret bound not just for the abstract game, but for the full game as well. We verify our theory and show how imperfect recall can be used to trade a small increase in regret for a significant reduction in memory in three domains: die-roll poker, phantom tic-tac-toe, and Bluff.

Near-Optimal BRL using Optimistic Local Transitions

Mauricio Araya, Olivier Buffet, Vincent Thomas

Model-based Bayesian Reinforcement Learning (BRL) allows a found formalization of the problem of acting optimally while facing an unknown environment, i.e., avoiding the exploration-exploitation dilemma. However, algorithms explicitly addressing BRL suffer from such a combinatorial explosion that a large body of work relies on heuristic algorithms. This paper introduces BOLT, a simple and (almost) deterministic heuristic algorithm for BRL which is optimistic about the transition function. We analyze BOLT's sample complexity, and show that under certain parameters, the algorithm is near-optimal in the Bayesian sense with high probability. Then, experimental results highlight the key differences of this method compared to previous

work.

Continuous Inverse Optimal Control with Locally Optimal Examples

Sergey Levine, Vladlen Koltun

Inverse optimal control, also known as inverse reinforcement learning, is the problem of recovering an unknown reward function in a Markov decision process from expert demonstrations of the optimal policy. Prior methods for inverse optimal control assume that the demonstrations are globally optimal (or near-optimal), and are often restricted to tasks with small, discrete state spaces. We introduce an inverse optimal control algorithm designed for high dimensional, continuous state spaces. Our algorithm does not suffer from the curse of dimensionality, and is suitable even for domains where computing a full policy is impractical. Our method also only assumes that the demonstrations are locally optimal, rather than requiring global optimality. This allows it to learn from examples that are unsuitable for prior methods.

Monte Carlo Bayesian Reinforcement Learning

Yi Wang, Kok Sung Won, David Hsu, Wee Sun Lee

Bayesian reinforcement learning (BRL) encodes prior knowledge of the world in a model and represents uncertainty in model parameters by maintaining a posterior distribution over them. This paper proposes a simple and general approach to BRL. The idea is to sample a priori a finite set of hypotheses for the model parameter values and form a discrete partially observable Markov decision process (POMDP) whose state space is a cross product of the state space for the reinforcement learning task and the sampled model parameter space. The POMDP does not require conjugate distributions for belief representation, as earlier works do, and can be solved relatively easily with point-based approximation algorithms. Our approach naturally handles both fully and partially observable worlds. Theoretical and experimental results show that the discrete POMDP approximates the underlying BRL problem well with guaranteed performance.

Apprenticeship Learning for Model Parameters of Partially Observable Environments

Takaki Makino, Johane Takeuchi

We consider apprentice learning — i.e., having an agent learn a task by observing an expert demonstrating the task — in a partially observable environment when the model of the environment is uncertain. This setting is useful in applications where the explicit modeling of the environment is difficult, such as a dialogue system. We show that we can extract information about the environment model by inferring action selection process behind the demonstration, under the assumption that the expert is choosing optimal actions based on knowledge of the true model of the target environment. We show that our proposed algorithms can estimate the parameters of

the environment model with much shorter demonstration compared to learning the model only from the reaction from the environment.

2C: Gaussian processes, 16:00–17:40, AT LT 1

Gaussian Process Regression Networks

Andrew Wilson, David Knowles, Zoubin Ghahramani

We introduce a new regression framework, Gaussian process regression networks (GPRN), which combines the structural properties of Bayesian neural networks with the nonparametric flexibility of Gaussian processes. GPRN accommodates input (predictor) dependent signal and noise correlations between multiple output (response) variables, input dependent length-scales and amplitudes, and heavy-tailed predictive distributions. We derive both elliptical slice sampling and variational Bayes inference procedures for GPRN. We apply GPRN as a multiple output regression and multivariate volatility model, demonstrating substantially improved performance over eight popular multiple output (multi-task) Gaussian process models and three multivariate volatility models on real datasets, including a 1000 dimensional gene expression dataset.

Infinite Tucker Decomposition: Nonparametric Bayesian Models for Multiway Data Analysis

Zenglin Xu, Feng Yan, Alan Qi

Tensor decomposition is a powerful computational tool for multiway data analysis. Many popular tensor decomposition approaches—such as the Tucker decomposition and CANDECOMP/PARAFAC (CP)—amount to multi-linear factorization. They are insufficient to model (i) complex interactions between data entities, (ii) various data types (eg missing data and binary data), and (iii) noisy observations and outliers. To address these issues, we propose tensor-variate latent nonparametric Bayesian models, coupled with efficient inference methods, for multiway data analysis. We name these models InfTucker. Using these InfTucker, we conduct Tucker decomposition in an infinite feature space. Unlike classical tensor decomposition models, our new approaches handle both continuous and binary data in a probabilistic framework. Unlike previous Bayesian models on matrices and tensors, our models are based on latent Gaussian or t processes with nonlinear covariance functions. To efficiently learn the InfTucker from data, we develop a variational inference technique on tensors. Compared with classical implementation, the new technique reduces both time and space complexities by several orders of magnitude. Our experimental results on chemometrics and social network datasets demonstrate that our new models achieved significantly higher prediction accuracy than the most state-of-art tensor decomposition approaches.

State-Space Inference for Non-Linear Latent Force Models with Application to Satellite Orbit Prediction

Jouni Hartikainen, Mari Seppänen, Simo Särkkä

Latent force models (LFMs) are flexible models that combine mechanistic modelling principles (i.e., physical models) with non-parametric data-driven components. Several key applications of LFMs need non-linearities, which results in analytically intractable inference. In this work we show how non-linear LFMs can be represented as non-linear white noise driven state-space models and present an efficient non-linear Kalman filtering and smoothing based method for approximate state and parameter inference. We illustrate the performance of the proposed methodology via two simulated examples, and apply it to a real-world problem of long-term prediction of GPS satellite orbits.

Gaussian Process Quantile Regression using Expectation Propagation

Alexis Boukouvalas, Remi Barillec, Dan Cornford

Direct quantile regression involves estimating a given quantile of a response variable as a function of input variables. We present a new framework for direct quantile regression where a Gaussian process model is learned, minimising the expected tilted loss function. The integration required in learning is not analytically tractable so to speed up the learning we employ the Expectation Propagation algorithm. We describe how this work relates to other quantile regression methods and apply the method on both synthetic and real data sets.

Residual Components Analysis

Alfredo Kalaitzis, Neil Lawrence

Probabilistic principal component analysis (PPCA) seeks a low dimensional representation of a data set in the presence of independent spherical Gaussian noise, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. The maximum likelihood solution for the model is an eigenvalue problem on the sample covariance matrix. In this paper we consider the situation where the data variance is already partially explained by other factors, e.g. conditional dependencies between the covariates, or temporal correlations leaving some residual variance. We decompose the residual variance into its components through a generalised eigenvalue problem, which we call residual component analysis (RCA). We explore a range of new algorithms that arise from the framework, including one that factorises the covariance of a Gaussian density into a low-rank and a sparse-inverse component. We illustrate the ideas on the recovery of a protein-signaling network, a gene expression time-series data set and the recovery of the human skeleton from motion capture 3-D cloud data.

Manifold Relevance Determination

Andreas Damianou, Carl Ek, Michalis Titsias, Neil Lawrence

In this paper we present a fully Bayesian latent variable model which exploits conditional non- linear (in)-dependence structures to learn an efficient latent representation. The model is capable of learning from extremely high-dimensional data such as directly modelling high resolution images. The latent representation is factorized to represent shared and private information from multiple views of the data. Bayesian techniques allow us to automatically estimate the dimensionality of the latent spaces. We demonstrate the model by prediction of human pose in an ambiguous setting. Our Bayesian representation allows us to perform disambiguation in a principled manner by including priors which incorporate the dynamics structure of the data. We demonstrate the ability of the model to capture structure underlying extremely high dimensional spaces by learning a low-dimensional representation of a set of facial images under different illumination conditions. The model correctly automatically creates a factorized representation where the lighting variance is represented in a separate latent space from the variance associated with different faces. We show that the model is capable of generating morphed faces and images from novel light directions.

2D: Statistical methods, 16:00–17:40, AT LT 2

Lognormal and Gamma Mixed Negative Binomial Regression

Mingyuan Zhou, Lingbo Li, David Dunson, Lawrence Carin

In regression analysis of counts, a lack of simple and efficient algorithms for posterior computation has made Bayesian approaches appear unattractive and thus underdeveloped. We propose a lognormal and gamma mixed negative binomial (NB) regression model for counts, and present efficient closed-form Bayesian inference; unlike conventional Poisson models, the proposed approach has two free parameters to include two different kinds of dispersion, and allows the incorporation of prior information, such as sparsity in the regression coefficients. By placing a gamma distribution prior on the NB dispersion parameter r, and connecting a lognormal distribution prior with the logit of the NB probability parameter p, efficient Gibbs sampling and variational Bayes inference are both developed. The closed-form updates are obtained by exploiting conditional conjugacy via both a compound Poisson representation and a Polya-Gamma distribution based data augmentation approach. The proposed Bayesian inference can be implemented routinely, while being easily generalizable to more complex settings involving multivariate dependence structures. The algorithms are illustrated using real examples.

Group Sparse Additive Models

Junming Yin, Xi Chen, eric xing

We consider the problem of sparse variable selection in nonparametric additive models, with the prior knowledge of the structure among the covariates to encourage

those variables within a group to be selected jointly. Previous works either study the group sparsity in the parametric setting (e.g., group lasso), or address the variable selection problem in the nonparametric setting without exploiting the structural information (e.g., sparse additive models (SpAM)). In this paper, we present a new method, called group sparse additive models (GroupSpAM), which can handle group sparsity in nonparametric additive models. We generalize the l1/l2 norm to Hilbert spaces as the sparsity-inducing penalty in GroupSpAM. Moreover, we derive a novel thresholding condition for identifying the functional sparsity at the group level, and propose an efficient block coordinate descent algorithm for constructing the estimate. We demonstrate by simulation that GroupSpAM substantially outperforms competing methods in terms of support recovery and prediction accuracy in additive models, and also conduct a comparative experiment on a real breast cancer dataset.

Variance Function Estimation in High-dimensions

Mladen Kolar, James Sharpnack

We consider the high-dimensional heteroscedastic regression model, where the mean and the log variance are modeled as a linear combination of input variables. Existing literature on high-dimensional linear regression models has largely ignored non-constant error variances, even though they commonly occur in a variety of applications ranging from biostatistics to finance. In this paper we study a class of non-convex penalized pseudolikelihood estimators for both the mean and variance parameters. We show that the Heteroscedastic Iterative Penalized Pseudolikelihood Optimizer (HIPPO) achieves the oracle property, that is, we prove that the rates of convergence are the same as if the true model was known. We demonstrate numerical properties of the procedure on a simulation study and real world data.

Sparse Additive Functional and Kernel CCA

Sivaraman Balakrishnan, Kriti Puniyani, John Lafferty

Canonical Correlation Analysis (CCA) is a classical tool for finding correlations among the components of two random vectors. In recent years, CCA has been widely applied to the analysis of genomic data, where it is common for researchers to perform multiple assays on a single set of patient samples. Recent work has proposed sparse variants of CCA to address the high dimensionality of such data. However, classical and sparse CCA are based on linear models, and are thus limited in their ability to find general correlations. In this paper, we present two approaches to high-dimensional nonparametric CCA, building on recent developments in high-dimensional nonparametric regression. We present estimation procedures for both approaches, and analyze their theoretical properties in the high-dimensional setting. We demonstrate the effectiveness of these procedures in discovering nonlinear correlations via extensive simulations, as well as through experiments with genomic data.

Consistent Covariance Selection From Data With Missing Values

Mladen Kolar, eric xing

Data sets with missing values arise in many practical problems and domains. However, correct statistical analysis of these data sets is difficult. A popular likelihood approach to statistical inference from partially observed data is the expectation maximization (EM) algorithm, which leads to non-convex optimization and estimates that are difficult to analyze theoretically. We study a simple two step procedure for covariance selection, which is tractable in high-dimensions and does not require imputation of the missing values. We provide rates of convergence for this estimator in the spectral norm, Frobenius norm and element-wise ℓ_∞ norm. Simulation studies show that this estimator compares favorably with the EM algorithm. Our results have important practical consequences as they show that standard tools for covariance selection can be used when data contains missing values, without resorting to the iterative EM algorithm that can be slow to converge in practice for large problems.

Conditional Sparse Coding and Grouped Multivariate Regression

Min Xu, John Lafferty

We study the problem of multivariate regression where the data are naturally grouped, and a regression matrix is to be estimated for each group. We propose an approach in which a dictionary of low rank parameter matrices is estimated across groups, and a sparse linear combination of the dictionary elements is estimated to form a model within each group. We refer to the method as conditional sparse coding since it is a coding procedure for the response vectors Y conditioned on the covariate vectors X. This approach captures the shared information across the groups while adapting to the structure within each group. It exploits the same intuition behind sparse coding that has been successfully developed in computer vision and computational neuroscience. We propose an algorithm for conditional sparse coding, analyze its theoretical properties in terms of predictive accuracy, and present the results of simulation experiments that compare the new technique to reduced rank regression.

Is margin preserved after random projection?

Qinfeng Shi, Chunhua Shen, Rhys Hill, Anton van den Hengel

Random projections have been applied in many machine learning algorithms. However, whether margin is preserved after random projection is non-trivial and not well studied. In this paper we analyse margin distortion after random projection, and give the conditions of margin preservation for binary classification problems. We also extend our analysis to margin for multiclass problems, and provide theoretical bounds on multiclass margin on the projected data.

Main Conference Abstracts: Thursday, June 28

3A: Optimization algorithms 2, 08:40–10:00, AT LT 4

Integer Optimization Methods for Supervised Ranking

Allison Chang, Cynthia Rudin, Dimitris Bertsimas

We consider supervised ranking problems, where the goal is to optimize a "rank statistic,' which is a quality measure of a ranked list. We present several mixed integer optimization (MIO) formulations for supervised ranking problems. Our formulations encompass a large set of rank statistics, including the Area Under the ROC Curve (AUC), the local AUC, and the Discounted Cumulative Gain (DCG). Other methods for supervised ranking approximate the ranking quality measure by a convex function in order to accommodate extremely large problems, at the expense of exact solutions. As our MIO approach provides exact modeling for ranking problems, our solutions are benchmarks for the other non-exact methods. We report computational results that demonstrate significant advantages for MIO methods over current state-of-the-art.

A Proximal-Gradient Homotopy Method for the L1-Regularized Least-Squares Problem

Lin Xiao, Tong Zhang

We consider the ℓ_1-regularized least-squares problem for sparse recovery and compressed sensing. Since the objective function is not strongly convex, standard proximal gradient methods only achieve sublinear convergence. We propose a homotopy continuation strategy, which employs a proximal gradient method to solve the problem with a sequence of decreasing regularization parameters. It is shown that under common assumptions in compressed sensing, the proposed method ensures that all iterates along the homotopy solution path are sparse, and the objective function is effectively strongly convex along the solution path. This observation allows us to obtain a global geometric convergence rate for the procedure. Empirical results are presented to support our theoretical analysis.

Complexity Analysis of the Lasso Regularization Path

Julien Mairal, Bin Yu

The regularization path of the Lasso can be shown to be piecewise linear, making it possible to "follow" and explicitly compute the entire path. We analyze in this paper this popular strategy, and prove that its worst case complexity is exponential in the number of variables. We then oppose this pessimistic result to an (optimistic) approximate analysis: We show that an approximate path with at most $O(1/\sqrt{\varepsilon})$ linear segments can always be obtained, where every point on the path is guaranteed to be optimal up to a relative ε-duality gap. We complete our theoretical analysis with a practical algorithm to compute these approximate paths.

Randomized Smoothing for (Parallel) Stochastic Optimization

John Duchi, Martin Wainwright, Peter Bartlett

By combining randomized smoothing techniques with accelerated gradient methods, we obtain convergence rates for stochastic optimization procedures, both in expectation and with high probability, that have optimal dependence on the variance of the gradient estimates. To the best of our knowledge, these are the first variance-based convergence guarantees for non-smooth optimization. A combination of our techniques with recent work on decentralized optimization yields order-optimal parallel stochastic optimization algorithms. We give applications of our results to several statistical machine learning problems, providing experimental results demonstrating the effectiveness of our algorithms.

3B: Clustering 1, 08:40–09:55, AT LT 5

Demand-Driven Clustering in Relational Domains for Predicting Adverse Drug Events

Jesse Davis, Vitor Santos Costa, Elizabeth Berg, David Page, Peggy Peissig, Michael Caldwell

Learning from electronic medical records (EMR) is challenging due to their relational nature and the need to capture the uncertain dependence between a patient's past and future health status. Statistical relational learning (SRL), which combines relational and probabilistic representations, is a natural fit for analyzing EMRs. SRL is less adept at handling the inherent latent structure, such as connections between related medications or diseases, in EMRs. One solution is to capture the latent structure via a relational clustering of objects. We propose a novel approach that, instead of pre-clustering the objects, performs a demand-driven clustering during the learning process. We evaluate our algorithm on three real-world tasks where the goal is to use EMRs to predict whether a patient will have an adverse reaction to a medication. We find that the proposed approach is more accurate than performing no clustering, pre-clustering and using expert-constructed medical heterarchies.

Clustering to Maximize the Ratio of Split to Diameter

Jiabing Wang, Jiaye Chen

Given a weighted and complete graph G = (V, E), V denotes the set of n objects to be clustered, and the weight d(u, v) associated with an edge (u, v) belonging to E denotes the dissimilarity between objects u and v. The diameter of a cluster is the maximum dissimilarity between pairs of objects in the cluster, and the split of a cluster is the minimum dissimilarity between objects within the cluster and objects outside the cluster. In this paper, we propose a new criterion for measuring the goodness of clusters:?the ratio of the minimum split to the maximum diameter, and the objective is to maximize the ratio. For k = 2, we present an exact algorithm. For k

>= 3, we prove that the problem is NP-hard and present a factor of 2 approximation algorithm on the precondition that the weights associated with edges of G satisfy the triangle inequality. The worst-case runtime of both algorithms is O(n3). We compare the proposed algorithms with the Normalized Cut by applying them to image segmentation. The experimental results on both natural and synthetic images demonstrate the effectiveness of the proposed algorithms.

An Iterative Locally Linear Embedding Algorithm

Deguang Kong, Chris H.Q. Ding

Local Linear embedding(LLE) is a popular dimension reduction method. In this paper, we first show LLE with nonnegative constraint is equivalent to the widely used Laplacian embedding. We further propose to iterate the two steps in LLE repeatedly to improve the results. Thirdly, we relax the kNN constraint of LLE and present a sparse similarity learning algorithm. The final Iterative LLE combines these three improvements. Extensive experiment results show that iterative LLE algorithm significantly improve both classification and clustering results.

Robust Multiple Manifold Structure Learning

Dian Gong, Xuemei Zhao, Gerard Medioni

We present a robust multiple manifold structure learning (RMMSL) scheme to robustly estimate data structures under the multiple low intrinsic dimensional manifolds assumption. In the local learning stage, RMMSL efficiently estimates local tangent space by weighted low-rank matrix factorization. In the global learning stage, we propose a robust manifold clustering method based on local structure learning results. The proposed clustering method is designed to get the flattest manifolds clusters by introducing a novel curved-level similarity function. Our approach is evaluated and compared to state-of-the-art methods on synthetic data, handwritten digit images, human motion capture data and motorbike videos. We demonstrate the effectiveness of the proposed approach, which yields higher clustering accuracy, and produces promising results for challenging tasks of human motion segmentation and motion flow learning from videos.

A Split-Merge Framework for Comparing Clusterings

Qiaoliang Xiang, Qi Mao, Kian Ming Chai, Hai Leong Chieu, Ivor Tsang, Zhenddong Zhao

Clustering evaluation measures are frequently used to evaluate the performance of algorithms. However, most measures are not suitable for this task mainly because they are not normalized properly and ignore some information in the inherent structure of clusterings such as dependency and consistency. We model two clusterings as a bipartite graph and propose a general component-based decomposition formula based on the components of the graph. Under this formula, most existing measures can be reinterpreted. In order to capture dependency and satisfy consistency in the

component, we further propose a split-merge framework for comparing clusterings of different data sets. The proposed framework is conditionally normalized, flexible to utilize previous measures, and extensible to take into account data point information, such as feature vectors and pairwise distances. Moreover, experimental results on a real world data set demonstrate that one instantiated measure of the framework outperforms other measures.

On the Difficulty of Nearest Neighbor Search

Junfeng He, Sanjiv Kumar, Shih-Fu Chang

Fast approximate nearest neighbor search in large databases is becoming popular. Several powerful learning-based formulations have been proposed recently. However, not much attention has been paid to a more fundamental question: how difficult is (approximate) nearest neighbor search in a given data set? More broadly, which data properties affect the nearest neighbor search and how? This paper introduces the first concrete measure called Relative Contrast that can be used to evaluate the influence of several crucial data characteristics such as dimensionality, sparsity, and database size simultaneously in arbitrary normed metric spaces. To further justify why relative contrast is an important and effective measure, we present a theoretical analysis to prove how relative contrast determines/affects the performance/complexity of Locality Sensitive Hashing, a popular hashing based approximate nearest neighbor search method. Finally, relative contrast also provides an explanation for a family of heuristic hashing algorithms based on PCA with good practical performance.

3C: Privacy, Anonymity, and Security, 8:40–10:00, AT LT 1

Bayesian Watermark Attacks

Ivo Shterev, David Dunson

This paper presents an application of statistical machine learning to the field of watermarking. We develop a new attack model on spread-spectrum watermarking systems, using Bayesian statistics. Our model jointly infers the watermark signal and embedded message bitstream, directly from the watermarked signal. No access to the watermark decoder is required. We develop an efficient Markov chain Monte Carlo sampler for updating the model parameters from their conjugate full conditional posteriors. We also provide a variational Bayesian solution, which further increases the convergence speed of the algorithm. Experiments with synthetic and real image signals demonstrate that the attack model is able to correctly infer a large part of the message bitstream, while at the same time obtaining a very accurate estimate of the watermark signal.

Poisoning Attacks against Support Vector Machines

Battista Biggio, Blaine Nelson, Pavel Laskov

We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks amount to injecting specially crafted training data so as to increase the test error of the trained SVM. Central to the motivation for these attacks is the fact that most learning algorithms assume that their training data comes from a natural or well-behaved distribution. However, this iassumption does not generally hold in security-sensitive settings. As we demonstrate in this contribution, an intelligent adversary can to some extent predict the change of the SVM decision function in response to malicious input and use this ability to construct malicious data points. The proposed attack uses a gradient ascent strategy in which the gradient is computed based on properties of the SVM's optimal solution. The gradient ascent method can be easily kernelized and enables the attack to be constructed in the input space even for non-linear kernels. We experimentally demonstrate that our gradient ascent procedure reliably identifies good local maxima of the non-convex validation error surface and inflicts a significant damage on the test error of the trained classifier.

Convergence Rates for Differentially Private Statistical Estimation

Kamalika Chaudhuri, Daniel Hsu

Differential privacy is a cryptographically-motivated privacy definition which has gained significant attention over the past few years. Differentially private solutions enforce privacy by adding random noise to the data or a function computed on the data, and the challenge in designing such algorithms is to optimize the privacy-accuracy-sample size tradeoff. This work studies differentially-private statistical estimation, and shows upper and lower bounds on the convergence rates of differentially private approximations to statistical estimators. Our results reveal a connection between differential privacy and the notion of B-robustness in robust statistics, by showing that unless an estimator is B-robust, we cannot approximate it well with differential privacy over a large class of distributions. We then provide an upper bound on the convergence rate of a differentially private approximation to a B-robust estimator with a bounded range. We show that the bounded range condition is necessary if we wish to ensure a strict form of differential privacy.

Finding Botnets Using Minimal Graph Clusterings

Peter Haider, Tobias Scheffer

We study the problem of identifying botnets and the IP addresses which they comprise, based on the observation of a fraction of the global email spam traffic. Observed mailing campaigns constitute evidence for joint botnet membership, they are represented by cliques in the graph of all messages. No evidence against an association of nodes is ever available. We reduce the problem of identifying botnets to a problem of finding a minimal clustering of the graph of messages. We directly model the distribution of clusterings given the input graph; this avoids potential errors caused

by distributional assumptions of a generative model. We report on a case study in which we evaluate the model by its ability to predict the spam campaign that a given IP address is going to participate in.

3D: Ranking and Preference Learning, 08:40-09:55, AT LT 2

Incorporating Domain Knowledge in Matching Problems via Harmonic Analysis

Deepti Pachauri, Maxwell Collins, Vikas SIngh

Matching one set of objects to another is a ubiquitous task in machine learning and computer vision that often reduces to some form of the quadratic assignment problem (QAP). The QAP is known to be notoriously hard, both in theory and in practice. Here, we investigate if this difficulty can be mitigated when some additional piece of information is available: (a) that all QAP instances of interest come from the same application, and (b) the correct solution for a set of such QAP instances is given. We propose a new approach to accelerating the solution of QAPs based on *learning* parameters for a modified objective function from prior QAP instances. A key feature of our approach is that it takes advantage of the algebraic structure of permutations, in conjunction with special methods for optimizing functions over the symmetric group S_n in Fourier space. Experiments show that in practical domains the new method can significantly outperform existing approaches.

Consistent Multilabel Ranking through Univariate Losses

Krzysztof Dembczyński, Wojciech Kotłowski, Eyke Huellermeier

We consider the problem of rank loss minimization in the setting of multilabel classification, which is commonly tackled by means of convex surrogate losses defined on pairs of labels. Very recently, this approach was put into question by the negative result showing that any such loss function is necessarily inconsistent. In this paper, we show a positive result which is arguably surprising in light of the previous one: despite being simpler, common convex surrogates for binary classification, defined on single labels instead of label pairs, are consistent for rank loss minimization. Instead of directly proving convergence, we give a much stronger result by deriving regret bounds and convergence rates. The proposed losses suggest efficient and scalable algorithms, which are tested experimentally.

Predicting Consumer Behavior in Commerce Search

Or Sheffet, Nina Mishra, Samuel Ieong

Traditional approaches to ranking in web search follow the paradigm of rank-by-score: a learned function gives each query-URL combination an absolute score and URLs are ranked according to this score. This paradigm ensures that if the score of one URL is better than another then one will always be ranked higher than the

other. Scoring contradicts prior work in behavioral economics that showed that users' preferences between two items depend not only on the items but also on the presented alternatives. Thus, for the same query, users' preference between items A and B depends on the presence/absence of item C. We propose a new model of ranking, the Random Shopper Model, that allows and explains such behavior. In this model, each feature is viewed as a Markov chain over the items to be ranked, and the goal is to find a weighting of the features that best reflects their importance. We show that our model can be learned under the empirical risk minimization framework, and give an efficient learning algorithm. Experiments on commerce search logs demonstrate that our algorithm outperforms scoring-based approaches including regression and listwise ranking.

Adaptive Regularization for Similarity Measures

Koby Crammer, Gal Chechik

Algorithms for learning distributions over weight-vectors, such as AROW were recently shown empirically to achieve state-of-the-art performance at various problems, with strong theoretical guaranties. Extending these algorithms to matrix models poses a challenge since the number of free parameters in the covariance of the distribution scales as n^4 with the dimension n of the matrix. We describe, analyze and experiment with two new algorithms for learning distribution of matrix models. Our first algorithm maintains a diagonal covariance over the parameters and is able to handle large covariance matrices. The second algorithm factores the covariance capturing some inter-features correlation while keeping the number of parameters linear in the size of the original matrix. We analyze the diagonal algorithm in the mistake bound model and show the superior precision of our approach over other algorithms in two tasks: retrieving similar images, and ranking similar documents. The second algorithms is shown to attain faster convergence rate.

Online Structured Prediction via Coactive Learning

Pannaga Shivaswamy, Thorsten Joachims

We propose Coactive Learning as a model of interaction between a learning system and a human user, where both have the common goal of providing results of maximum utility to the user. At each step, the system (e.g. search engine) receives a context (e.g. query) and predicts an object (e.g. ranking). The user responds by correcting the system if necessary, providing a slightly improved – but not necessarily optimal – object as feedback. We argue that such feedback can be inferred from observable user behavior, specifically clicks in web search. Evaluating predictions by their cardinal utility to the user, we propose efficient learning algorithms that have O(1/sqrtT) average regret, even though the learning algorithm never observes cardinal utility values. We demonstrate the applicability of our model and learning algorithms on a

movie recommendation task, as well as ranking for web search.

TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings

Chao Liu

This paper revisits the problem of analyzing multiple ratings given by different judges. Different from previous work that focuses on distilling the true labels from noisy crowdsourcing ratings, we emphasize gaining diagnostic insights into our in-house well-trained judges. We generalize the well-known DawidSkene model (Dawid & Skene, 1979) to a spectrum of probabilistic models under the same "TrueLabel + Confusion" paradigm, and show that our proposed hierarchical Bayesian model, called HybridConfusion, consistently outperforms DawidSkene on both synthetic and real-world data sets.

3E: Nonparametric Bayesian inference, 08:40–10:00, AT LT 3

Factorized Asymptotic Bayesian Hidden Markov Models

Ryohei Fujimaki, Kohei Hayashi

This paper addresses the issue of model selection for hidden Markov models (HMMs). We generalize factorized asymptotic Bayesian inference (FAB), which has been recently developed for model selection on independent hidden variables (i.e., mixture models), for time-dependent hidden variables. As with FAB in mixture models, FAB for HMMs is derived as an iterative lower bound maximization algorithm of a factorized information criterion (FIC). It inherits, from FAB for mixture models, several desirable properties for learning HMMs, such as asymptotic consistency of FIC with marginal log-likelihood, a shrinkage effect for hidden state selection, monotonic increase of the lower FIC bound through the iterative optimization. Further, it does not have a tunable hyper-parameter, and thus its model selection process can be fully automated. Experimental results shows that FAB outperforms states-of-the-art variational Bayesian HMM and non-parametric Bayesian HMM in terms of model selection accuracy and computational efficiency.

An Infinite Latent Attribute Model for Network Data

David Knowles, Konstantina Palla, Zoubin Ghahramani

Latent variable models for network data extract a summary of the relational structure underlying an observed network. The simplest possible models subdivide nodes of the network into clusters; the probability of a link between any two nodes then depends only on their cluster assignment. Currently available models can be classified by whether clusters are disjoint or are allowed to overlap. These models can explain a "flat" clustering structure. Hierarchical Bayesian models provide a natural approach to capture more complex dependencies. We propose a model in which objects are

characterised by a latent feature vector. Each feature is itself partitioned into disjoint groups (subclusters), corresponding to a second layer of hierarchy. In experimental comparisons, the model achieves significantly improved predictive performance on social and biological link prediction tasks. The results indicate that models with a single layer hierarchy over-simplify real networks.

The Nonparametric Metadata Dependent Relational Model

Dae Il Kim, Michael Hughes, Erik Sudderth

We introduce the nonparametric metadata dependent relational (NMDR) model, a Bayesian nonparametric extension of previous stochastic block models. The NMDR allows the entities associated with each node to have mixed membership in an unbounded collection of latent communities. Learned regression models allow these memberships to depend on, and be predicted from, arbitrary node metadata. We develop efficient MCMC algorithms for learning NMDR models from (partially observed) node relationships. Retrospective MCMC methods allow our sampler to work directly with the infinite stick-breaking representation of the NMDR, avoiding the need for finite truncations. Our results demonstrate recovery of useful latent communities from real-world social and ecological networks, and the usefulness of metadata in link prediction tasks.

Dependent Hierarchical Normalized Random Measures for Dynamic Topic Modeling

Changyou Chen, Nan Ding, Wray Buntine

We develop dependent hierarchical normalized random measures and apply them to dynamic topic modeling. The dependency arises via superposition, subsampling and point transition on the underlying Poisson processes of these measures. The measures used include normalised generalised Gamma processes that demonstrate power law properties, unlike Dirichlet processes used previously in dynamic topic modeling. Inference for the model includes adapting a recently developed slice sampler to directly manipulate the underlying Poisson process. Experiments performed on news, blogs, academic and Twitter collections demonstrate the technique gives superior perplexity over a number of previous models.

A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling

Drausin Wulsin, Shane Jensen, Brian Litt

Driven by the multi-level structure of human intracranial electroencephalogram (iEEG) recordings of epileptic seizures, we introduce a new variant of a hierarchical Dirichlet Process—the multi-level clustering hierarchical Dirichlet Process (MLC-HDP)—that simultaneously clusters datasets on multiple levels. Our seizure dataset contains brain activity recorded in typically more than a hundred individual channels

for each seizure of each patient. The MLC-HDP model clusters over channels-types, seizure-types, and patient-types simultaneously. We describe this model and its implementation in detail. We also present the results of a simulation study comparing the MLC-HDP to a similar model, the Nested Dirichlet Process and finally demonstrate the MLC-HDP's use in modeling seizures across multiple patients. We find the MLC-HDP's clustering to be comparable to independent human physician clusterings. To our knowledge, the MLC-HDP model is the first in the epilepsy literature capable of clustering seizures within and between patients.

Modeling Images using Transformed Indian Buffet Processes

KE ZHAI, Yuening Hu, Jordan Boyd-Graber, Sinead Williamson

Latent feature models are attractive for image modeling; images generally contain multiple objects. However, many latent feature models ignore that objects can appear at different locations, or require pre-segmentation of images. While the transformed Indian buffet process (tIBP) provides a method for modeling transformation-invariant features in simple, unsegmented binary images, in its current form it is inappropriate for real images because of computational constraints and modeling assumptions. We combine the tIBP with likelihoods appropriate for real images. We also develop an efficient inference scheme using the cross-correlation between images and features that is both theoretically and empirically faster than existing inference techniques. We demonstrate that, using our method, we are able to discover reasonable components and achieve effective image reconstruction in natural images.

A Topic Model for Melodic Sequences

Athina Spiliopoulou, Amos Storkey

We examine the problem of learning a probabilistic model for melody directly from musical sequences belonging to the same genre. This is a challenging task as one needs to tackle not only the complex statistical dependencies that characterize a music genre, but also the element of novelty, as each music piece is a unique realization of a musical form. To address this problem we introduce the Variable-gram Topic Model, which couples the latent topic formalism with a systematic model for contextual information. We evaluate the model using a next-step prediction task. Additionally, we present a novel way of model evaluation, where we directly compare model samples with data sequences using the Maximum Mean Discrepancy (MMD) of string kernels, to assess how close is the model distribution to the data distribution. We show that the model has the highest performance under both evaluation measures when compared to LDA, the Topic Bigram and related non-topic models.

4A: Feature selection and dimensionality reduction 1, 10:30–12:00, AT LT 4

Discovering Support and Affiliated Features from Very High Dimensions

Yiteng Zhai, Mingkui Tan, Ivor Tsang, Yew Soon Ong

In this paper, a novel learning paradigm is p- resented to automatically identify groups of in- formative and correlated features from very high dimensions. Specifically, we explicitly incorpo- rate correlation measures as constraints and then develop an efficient embedded feature selection method using recently developed cutting plane s- trategy. The advantages of the proposed algorith- m are two-fold. First, it can identify the opti- mal discriminative and uncorrelated feature sub- set to the output labels, denoted as Support fea- tures, whichgivesgreatimprovementsinprediction performance over other state-of-the-art fea- ture selection methods considered in the paper. Second, during the learning process, the underly- ing group structures of correlated features asso- ciated with each support feature, denoted as Affiliated features, can also be discovered without any additional cost. The affiliated features im- prove interpretation of the learning tasks. Exten- sive empirical studies on both synthetic and real- world very high dimensional datasets verify the validity and efficiency of the proposed method.

Inferring Latent Structure From Mixed Real and Categorical Relational Data

Esther Salazar, Lawrence Carin

We consider analysis of relational data (a matrix), in which the rows correspond to subjects (e.g., people) and the columns correspond to attributes. The elements of the matrix may be a mix of real and categorical. Each subject and attribute is characterized by a latent binary feature vector, and an inferred matrix maps each row-column pair of binary feature vectors to an observed matrix element. The latent binary features of the rows are modeled via a multivariate Gaussian distribution with low-rank covariance matrix, and the Gaussian random variables are mapped to latent binary features via a probit link. The same type construction is applied jointly to the columns. The model infers latent, low-dimensional binary features associated with each row and each column, as well correlation structure between all rows and between all columns. The Bayesian construction is successfully applied to real-world data, demonstrating an ability to infer meaningful low-dimensional structure from high-dimensional relational data.

Conditional Likelihood Maximization: A Unifying Framework for Information Theoretic Feature Selection

Gavin Brown, Adam Pocock, Ming-Jie Zhao, Mikel Lujan

We present a unifying framework for information theoretic feature selection, bringing almost two decades of research on heuristic filter criteria under a single theoretical interpretation. This is in response to the question: “what are the implicit statistical assumptions of feature selection criteria based on mutual information?”. To answer

this, we adopt a different strategy than is usual in the feature selection literature-instead of trying to define a criterion, we derive one, directly from a clearly specified objective function: the conditional likelihood of the training labels. While many hand-designed heuristic criteria try to optimize a definition of feature 'relevancy' and 'redundancy', our approach leads to a probabilistic framework which naturally incorporates these concepts. As a result we unify the numerous criteria published over the last two decades, and show them to be low-order approximations to the exact (but intractable) optimization problem. The primary contribution is to show that common heuristics for information based feature selection (including Markov Blanket algorithms as a special case) are approximate iterative maximizers of the conditional likelihood. A large empirical study provides strong evidence to favor certain classes of criteria, in particular those that balance the relative size of the relevancy/redundancy terms. Overall we conclude that the JMI criterion (Yang and Moody, 1999; Meyer et al., 2008) provides the best tradeoff in terms of accuracy, stability, and flexibility with small data samples.

Dimensionality Reduction by Local Discriminative Gaussians

Nathan Parrish, Maya Gupta

We present local discriminative Gaussian (LDG) dimensionality reduction, a supervised linear dimensionality reduction technique that acts locally to each training point in order to find a mapping where similar data can be discriminated from dissimilar data. We focus on the classification setting; however, our algorithm can be applied whenever training data is accompanied with similarity or dissimilarity constraints. Our experiments show that LDG is superior to other state-of-the-art linear dimensionality reduction techniques when the number of features in the original data is large. We also adapt LDG to the transfer learning setting, and show that it achieves good performance when the test data distribution differs from that of the training data.

Fast Prediction of New Feature Utility

Hoyt Koepke, Mikhail Bilenko

We study the new feature utility prediction problem: statistically testing whether adding a new feature to the data representation can improve predictive accuracy on a supervised learning task. In many applications, identifying new informative features is the primary pathway for improving performance. However, evaluating every potential feature by re-training the predictor with it can be costly computationally, logistically and financially. The paper describes an efficient, learner-independent technique for estimating new feature utility without re-training based on the current predictor's outputs. The method is obtained by deriving a connection between loss reduction potential and the new feature's correlation with the loss gradient of the

current predictor. This leads to a simple yet powerful hypothesis testing procedure, for which we prove consistency. Our theoretical analysis is accompanied by empirical evaluation on standard benchmarks and a large-scale industrial dataset.

4B: Online learning 1, 10:30–12:00, AT LT 5

An Online Boosting Algorithm with Theoretical Justifications

Shang-Tse Chen, Hsuan-Tien Lin, Chi-Jen Lu

We study the task of online boosting, which aims to combine online weak learners into an online strong learner. While boosting in the batch setting has a sound theoretical foundation, not much theory is known for the online setting. In this paper, we propose an online boosting algorithm and we provide a theoretical guarantee. In addition, we also perform experiments on real-world datasets and the results show that our algorithm compares favorably with existing algorithms.

An adaptive algorithm for finite stochastic partial monitoring

Gabor Bartok, Navid Zolghadr, Csaba Szepesvari

We present a new anytime algorithm that achieves near-optimal regret for any instance of finite stochastic partial monitoring. In particular, the new algorithm achieves the minimax regret, within logarithmic factors, for both "easy" and "hard" problems. For easy problems, it additionally achieves logarithmic individual regret. Most importantly, the algorithm is adaptive in the sense that if the opponent strategy is in an "easy region" of the strategy space then the regret grows as if the problem was easy. As an implication, we show that under some reasonable additional assumptions, the algorithm enjoys an $O(\sqrt{T})$ regret in Dynamic Pricing, proven to be hard by Bartok et al. (2011).

Online Alternating Direction Method

Huahua Wang, Arindam Banerjee

Online optimization has emerged as powerful tool in large scale optimization. In this paper, we introduce efficient online algorithms based on the alternating directions method (ADM). We introduce a new proof technique for ADM in the batch setting, which yields a linear rate of convergence of ADM and forms the basis of regret analysis in the online setting. We consider two scenarios in the online setting, based on whether the solution needs to lie in the feasible set or not. In both settings, we establish regret bounds for both the objective function as well as constraint violation for general and strongly convex functions. Preliminary results are presented to illustrate the performance of the proposed algorithms.

Projection-free Online Learning

Elad Hazan, Satyen Kale

We present efficient online learning algorithms that eschew projections in favor of

linear optimizations using the Frank-Wolfe technique. We obtain regret bounds that vary from the optimal $\tilde{O}(\sqrt{T})$ for stochastic online smooth convex optimization to $\tilde{O}(T^{3/4})$ for general online convex optimization. Besides the computational advantage, other desirable features of our algorithms are that they are parameter-free in the stochastic case and produce sparse decisions.

PAC Subset Selection in Stochastic Multi-armed Bandits

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, Peter Stone

We consider the problem of selecting, from among the arms of a stochastic n-armed bandit, a subset of size m of those arms with the highest expected rewards, based on efficiently sampling the arms. This "subset selection" problem finds application in a variety of areas. Kalyanakrishnan & Stone (2010) frame this problem under a PAC setting (denoting it "Explore-m") and analyze corresponding sampling algorithms both formally and experimentally. Whereas their formal analysis is restricted to the worst case sample complexity of algorithms, in this paper, we design and analyze an algorithm ("LUCB") with improved expected sample complexity. Interestingly LUCB bears a close resemblance to the well-known UCB algorithm for regret minimization. We obtain a sample complexity bound for LUCB that matches the best existing bound for single-arm selection (that is, when m = 1). We also provide a lower bound on the worst case sample complexity of PAC algorithms for Explore-m.

On Local Regret

Michael Bowling, Martin Zinkevich

Online learning typically aims to perform nearly as well as the best hypothesis in hindsight. For some hypothesis classes, though, even finding the best hypothesis offline is challenging. In such offline cases, local search techniques are often employed and only local optimality guaranteed. For online decision-making with such hypothesis classes, we introduce local regret, a generalization of regret that aims to perform nearly as well as only nearby hypotheses. We then present a general algorithm that can minimize local regret for arbitrary locality graphs. We also show that certain forms of structure in the graph can be exploited to drastically simplify learning. These algorithms are then demonstrated on a diverse set of online problems (some previously unexplored): online disjunct learning, online Max-SAT, and online decision tree learning.

Exact Soft Confidence-Weighted Learning

Steven C.H. Hoi, Jialei Wang, Peilin Zhao

In this paper, we propose a new Soft Confidence-Weighted (SCW) online learning scheme, which enables the conventional confidence-weighted learning method to handle non-separable cases. Unlike the previous confidence-weighted learning algorithms, the proposed soft confidence-weighted learning method enjoys all the four salient

properties: (i) large margin training, (ii) confidence weighting, (iii) capability to handle non-separable data, and (iv) adaptive margin. Our experimental results show that SCW performs significantly better than the original CW algorithm. When comparing with the state-of-the-art AROW algorithm, we found that SCW in general achieves better or at least comparable predictive performance, but enjoys considerably better efficiency performance (i.e., producing less number of updates and spending less time cost).

Compact Hyperplane Hashing with Bilinear Functions

Wei Liu, Jun Wang, Yadong Mu, Sanjiv Kumar, Shih-Fu Chang

Hyperplane hashing aims at rapidly searching nearest points to a hyperplane, and has shown practical impact in scaling up active learning with SVMs. Unfortunately, the existing randomized methods need long hash codes and a number of hash tables to achieve reasonable search accuracy. Thus, they suffer from reduced search speed and large memory overhead. To this end, this paper proposes a novel hyperplane hashing technique which yields compact hash codes. The key idea is the bilinear form of the proposed hash functions, which leads to higher collision probability than the existing hyperplane hash functions when using random projections. To further increase the performance, we propose a learning based framework in which bilinear functions are directly learned from the data. This yields compact yet discriminative codes, and also increases the search performance over the random projection based solutions. Large-scale active learning experiments carried out on two datasets with up to one million samples demonstrate the overall superiority of the proposed approach.

4C: Supervised learning 1, 10:30–12:00, AT LT 1

Improved Information Gain Estimates for Decision Tree Induction

Sebastian Nowozin

Ensembles of classification and regression trees remain popular machine learning methods because they define flexible non-parametric models that predict well and are computationally efficient both during training and testing. During induction of decision trees one aims to find predicates that are maximally informative about the prediction target. To select good predicates most approaches estimate an information-theoretic scoring function, the information gain, both for classification and regression problems. We point out that the common estimation procedures are biased and show that by replacing them with improved estimators of the discrete and the differential entropy we can obtain better decision trees. In effect our modifications yield improved predictive performance and are simple to implement in any decision tree code.

Unachievable Region in Precision-Recall Space and Its Effect on Empirical

Evaluation

Kendrick Boyd, Jesse Davis, David Page, Vitor Santos Costa

Precision-recall (PR) curves and the areas under them are widely used to summarize machine learning results, especially for data sets exhibiting class skew. They are often used analogously to ROC curves and the area under ROC curves. It is already known that PR curves vary as class skew varies. What was not recognized before this paper is that there is a region of PR space that is completely unachievable, and the size of this region varies only with the skew. This paper precisely characterizes the size of that region and discusses its implications for empirical evaluation methodology in machine learning.

Bootstrapping Big Data

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, Michael Jordan

The bootstrap provides a simple and powerful means of assessing the quality of estimators. However, in settings involving large datasets, the computation of bootstrap-based quantities can be prohibitively demanding. As an alternative, we present BLB, a new procedure which incorporates features of both the bootstrap and subsampling to obtain a robust, computationally efficient means of assessing the quality of estimators. BLB is well suited to modern parallel and distributed computing architectures and retains the generic applicability, statistical efficiency, and favorable theoretical properties of the bootstrap. We provide a theoretical analysis elucidating the properties of BLB, as well as an extensive empirical investigation, including a study of its statistical correctness, its large-scale implementation and performance, selection of hyperparameters, and performance on real data.

Robust Classification with Adiabatic Quantum Optimization

Vasil Denchev, Nan Ding, SVN Vishwanathan, Hartmut Neven

We propose a non-convex training objective for robust binary classification of data sets in which label noise is present. The design is guided by the intention of solving the resulting problem by adiabatic quantum optimization. Two choices are prompted by the engineering constraints of existing quantum hardware: training problems are formulated as quadratic unconstrained binary optimization; and model parameters are represented as binary expansions of low bit-depth. In the present work we validate this approach by using a heuristic classical solver as a stand-in for quantum hardware. Testing on several popular data sets and comparing with a number of existing convex and non-convex losses solved by convex optimization, we find substantial advantages in robustness as measured by test error under increasing label noise. Robustness is enabled by the non-convexity of our hardware-compatible loss function, which we name *q-loss*. We emphasize that we do not claim any theoretical superiority of q-loss over other non-convex losses. In fact if solving them to optimality was possible, they

could be just as good or better, but q-loss has one important property that all others are lacking—namely that it is compatible with quantum hardware.

Nonparametric Link Prediction in Dynamic Networks

Purnamrita Sarkar, Deepayan Chakrabarti, Michael Jordan

We propose a non-parametric link prediction algorithm for a sequence of graph snapshots over time. The model predicts links based on the features of its endpoints, as well as those of the *local neighborhood* around the endpoints. This allows for different *types* of neighborhoods in a graph, each with its own dynamics (e.g, growing or shrinking communities). We prove the consistency of our estimator, and give a fast implementation based on locality-sensitive hashing. Experiments with simulated as well as five real-world dynamic graphs show that we outperform the state of the art, especially when sharp fluctuations or non-linearities are present.

A Unified Robust Classification Model

Akiko Takeda, Hiroyuki Mitsugi , Takafumi Kanamori

A wide variety of machine learning algorithms such as support vector machine (SVM), minimax probability machine (MPM), and Fisher discriminant analysis (FDA), exist for binary classification. The purpose of this paper is to provide a unified classification model that includes the above models through a robust optimization approach. This unified model has several benefits. One is that the extensions and improvements intended for SVM become applicable to MPM and FDA, and vice versa. Another benefit is to provide theoretical results to above learning methods at once by dealing with the unified model. We give a statistical interpretation of the unified classification model and propose a non-convex optimization algorithm that can be applied to non-convex variants of existing learning methods.

Maximum Margin Output Coding

Yi Zhang, Jeff Schneider

In this paper we study output coding for multi-label prediction. For a multi-label output coding to be discriminative, it is important that codewords for different label vectors are significantly different from each other. In the meantime, unlike in traditional coding theory, codewords in output coding are to be predicted from the input, so it is also critical to have a predictable label encoding. To find output codes that are both discriminative and predictable, we first propose a max-margin formulation that naturally captures these two properties. We then convert it to a metric learning formulation, but with an exponentially large number of constraints as commonly encountered in structured prediction problems. Without a label structure for tractable inference, we use overgenerating (i.e., relaxation) techniques combined with the cutting plane method for optimization. In our empirical study, the proposed output coding scheme outperforms a variety of existing multi-label prediction methods for

image, text and music classification.

Structured Learning from Partial Annotations

Xinghua Lou, Fred Hamprecht

Structured learning is appropriate when predicting structured outputs such as trees, graphs, or sequences. Most prior work requires the training set to consist of complete trees, graphs or sequences. Specifying such detailed ground truth can be tedious or infeasible for large outputs. Our main contribution is a large margin formulation that makes structured learning from only partially annotated data possible. The resulting optimization problem is non-convex, yet can be efficiently solve by concave-convex procedure (CCCP) with novel speedup strategies. We apply our method to a challenging tracking-by-assignment problem of a variable number of divisible objects. On this benchmark, using only 25% of a full annotation we achieve a performance comparable to a model learned with a full annotation. Finally, we offer a unifying perspective of previous work using the hinge, ramp, or max loss for structured learning, followed by an empirical comparison on their practical performance.

4D: Transfer and Multi-Task Learning, 10:30–12:00, AT LT 2

Marginalized Denoising Autoencoders for Domain Adaptation

Minmin Chen, Zhixiang Xu, Kilian Weinberger, Fei Sha

Glorot et al. (2011) have successfully used Stacked Denoising Autoencoders (SDAs) (Vincent et al., 2008) to learn new representations for domain adaptation, resulting in record accuracy levels on well-known benchmark datasets for sentiment analysis. The representations are learned by reconstructing input data from partial corruption. In this paper, we introduce a variation, marginalized SDA (mSDA). In contrast to the original SDA, mSDA requires no training through back-propagation as we explicitly marginalize out the feature corruption and solve for the parameters in closed form. mSDA learns representations that lead to comparable accuracy levels, but can be implemented in only 20 lines of MATLAB and reduces the computation time on large data sets from several days to mere minutes.

Information-Theoretical Learning of Discriminative Clusters for Unsupervised Domain Adaptation

Yuan Shi, Fei Sha

We study the problem of unsupervised domain adaptation, which aims to adapt classifiers trained on a labeled source domain to an unlabeled target domain. Many existing approaches first learn domain-invariant features and then construct classifiers with them. We propose a novel approach that jointly learn the both. Specifically, the proposed method identifies a feature space where data in the source and the target domains are similarly distributed. More importantly, our method learns a

discriminative feature space, optimizing an information-theoretic metric as an proxy to the expected misclassification error on the target domain. We show how this optimization can be effectively carried out with simple gradient-based methods and how hyperparameters can be cross validated without demanding any labeled data from the target domain. Empirical studies on the benchmark tasks of visual object recognition and sentiment analysis validate our modeling assumptions and demonstrate significant improvement of our method over competing ones in classification accuracies.

Learning Task Grouping and Overlap in Multi-task Learning

Abhishek Kumar, Hal Daume III

In the paradigm of multi-task learning, mul- tiple related prediction tasks are learned jointly, sharing information across the tasks. We propose a framework for multi-task learn- ing that enables one to selectively share the information across the tasks. We assume that each task parameter vector is a linear combi- nation of a finite number of underlying basis tasks. The coefficients of the linear combina- tion are sparse in nature and the overlap in the sparsity patterns of two tasks controls the amount of sharing across these. Our model is based on on the assumption that task pa- rameters within a group lie in a low dimen- sional subspace but allows the tasks in differ- ent groups to overlap with each other in one or more bases. Experimental results on four datasets show that our approach outperforms competing methods.

A Convex Feature Learning Formulation for Latent Task Structure Discovery

Pratik Jawanpuria, J. Saketha Nath

This paper considers the multi-task learning problem and in the setting where some relevant features could be shared across few related tasks. Most of the existing methods assume the extent to which the given tasks are related or share a common feature space to be known apriori. In real-world applications however, it is desirable to automatically discover the groups of related tasks that share a feature space. In this paper we aim at searching the exponentially large space of all possible groups of tasks that may share a feature space. The main contribution is a convex formulation that employs a graph-based regularizer and simultaneously discovers few groups of related tasks, having close-by task parameters, as well as the feature space shared within each group. The regularizer encodes an important structure among the groups of tasks leading to an efficient algorithm for solving it: if there is no feature space under which a group of tasks has close-by task parameters, then there does not exist such a feature space for any of its supersets. An efficient active set algorithm that exploits this simplification and performs a clever search in the exponentially large space is presented. The algorithm is guaranteed to solve the proposed formulation

(within some precision) in a time polynomial in the number of groups of related tasks discovered. Empirical results on benchmark datasets show that the proposed formulation achieves good generalization and outperforms state-of-the-art multi-task learning algorithms in some cases.

Convex Multitask Learning with Flexible Task Clusters

Wenliang Zhong, James Kwok

Traditionally, multitask learning (MTL) assumes that all the tasks are related. This can lead to negative transfer when tasks are indeed incoherent. Recently, a number of approaches have been proposed that alleviate this problem by discovering the underlying task clusters or relationships. However, they are limited to modeling these relationships at the task level,which may be restrictive in some applications. In this paper, we propose a novel MTL formulation that captures task relationships at the feature-level. Depending on the interactions among tasks and features, the proposed method construct different task clusters for different features, without even the need of pre-specifying the number of clusters. Computationally, the proposed formulation is strongly convex, and can be efficiently solved by accelerated proximal methods. Experiments are performed on a number of synthetic and real-world data sets. Under various degrees of task relationships, the accuracy of the proposed method is consistently among the best. Moreover, the feature-specific task clusters obtained agree with the known/plausible task structures of the data.

A Complete Analysis of the $l_{1,p}$ Group-Lasso

Julia Vogt, Volker Roth

The Group-Lasso is a well-known tool for joint regularization in machine learning methods. While the $l_{1,2}$ and the $l_{1,\infty}$ version have been studied in detail and efficient algorithms exist, there are still open questions regarding other $l_{1,p}$ variants. We characterize conditions for solutions of the $l_{1,p}$ Group-Lasso for all p-norms with $1 <= p <= \infty$, and we present a unified active set algorithm. For all p-norms, a highly efficient projected gradient algorithm is presented. This new algorithm enables us to compare the prediction performance of many variants of the Group-Lasso in a multi-task learning setting, where the aim is to solve many learning problems in parallel which are coupled via the Group-Lasso constraint. We conduct large-scale experiments on synthetic data and on two real-world data sets. In accordance with theoretical characterizations of the different norms we observe that the weak-coupling norms with p between 1.5 and 2 consistently outperform the strong-coupling norms with $p >> 2$.

Learning with Augmented Features for Heterogeneous Domain Adaptation

Lixin Duan, Dong Xu, Ivor Tsang

We propose a new learning method for heterogeneous domain adaptation (HDA),

in which the data from the source domain and the target domain are represented by heterogeneous features with different dimensions. Using two different projection matrices, we first transform the data from two domains into a common subspace in order to measure the similarity between the data from two domains. We then propose two new feature mapping functions to augment the transformed data with their original features and zeros. The existing learning methods (e.g., SVM and SVR) can be readily incorporated with our newly proposed augmented feature representations to effectively utilize the data from both domains for HDA. Using the hinge loss function in SVM as an example, we introduce the detailed objective function in our method called Heterogeneous Feature Augmentation (HFA) for a linear case and also describe its kernelization in order to efficiently cope with the data with very high dimensions. Moreover, we also develop an alternating optimization algorithm to effectively solve the nontrivial optimization problem in our HFA method. Comprehensive experiments on two benchmark datasets clearly demonstrate that our HFA outperforms the existing HDA methods.

Cross-Domain Multitask Learning with Latent Probit Models

Shaobo Han, Xuejun Liao, Lawrence Carin

Learning multiple tasks across heterogeneous domains is a challenging problem since the feature space may not be the same for different tasks. We assume the data in multiple tasks are generated from a latent common domain via sparse domain transforms and propose a latent probit model (LPM) to jointly learn the domain transforms, and the shared probit classifier in the common domain. To learn meaningful task relatedness and avoid over-fitting in classification, we introduce sparsity in the domain transforms matrices, as well as in the common classifier. We derive theoretical bounds for the estimation error of the classifier in terms of the sparsity of domain transforms. An expectation-maximization algorithm is derived for learning the LPM. The effectiveness of the approach is demonstrated on several real datasets.

4E: Graphical models, 10:30–12:00, AT LT 3

High Dimensional Semiparametric Gaussian Copula Graphical Models

Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman

In this paper, we propose a semiparametric approach named nonparanormal SKEPTIC for efficiently and robustly estimating high dimensional undirected graphical models. To achieve modeling flexibility, we consider Gaussian Copula graphical models (or the nonparanormal) as proposed by Liu et al. (2009). To achieve estimation robustness, we exploit nonparametric rank-based correlation coefficient estimators, including Spearman's rho and Kendall's tau. In high dimensional settings, we prove that the nonparanormal SKEPTIC achieves the optimal parametric rate of convergence in both graph and parameter estimation. This celebrating result suggests that

the Gaussian copula graphical models can be used as a safe replacement of the popular Gaussian graphical models, even when the data are truly Gaussian. Besides theoretical analysis, we also conduct thorough numerical simulations to compare different estimators for their graph recovery performance under both ideal and noisy settings. The proposed methods are then applied on a large-scale genomic dataset to illustrate their empirical usefulness.

Convergence Rates of Biased Stochastic Optimization for Learning Sparse Ising Models

Jean Honorio

We study the convergence rate of stochastic optimization of exact (NP-hard) objectives, for which only biased estimates of the gradient are available. We motivate this problem in the context of learning the structure and parameters of Ising models. We first provide a convergence-rate analysis of deterministic errors for forward-backward splitting (FBS). We then extend our analysis to biased stochastic errors, by first characterizing a family of samplers and providing a high probability bound that allows understanding not only FBS, but also proximal gradient (PG) methods. We derive some interesting conclusions: FBS requires only a logarithmically increasing number of random samples in order to converge (although at a very low rate); the required number of random samples is the same for the deterministic and the biased stochastic setting for FBS and basic PG; accelerated PG is not guaranteed to converge in the biased stochastic setting.

On the Partition Function and Random Maximum A-Posteriori Perturbations

Tamir Hazan, Tommi Jaakkola

n this paper we relate the partition function to the max-statistics of random variables. In particular, we provide a novel framework for approximating and bounding the partition function using MAP inference on randomly perturbed models. As a result, we can directly use efficient MAP solvers such as graph-cuts to evaluate the corresponding partition function. We show that our method excels in the typical "high signal - high coupling" regime that results in ragged energy landscapes difficult for alternative approaches.

Anytime Marginal MAP Inference

Denis Maua, Cassio De Campos

This paper presents a new anytime algorithm for the marginal MAP problem in graphical models. The algorithm is described in detail, its complexity and convergence rate are studied, and relations to previous theoretical results for the problem are discussed. It is shown that the algorithm runs in polynomial-time if the underlying graph of the model has bounded tree-width, and that it provides guarantees to the

lower and upper bounds obtained within a fixed amount of computational resources. Experiments with both real and synthetic generated models highlight its main characteristics and show that it compares favorably against Park and Darwiche's systematic search, particularly in the case of problems with many MAP variables and moderate tree-width.

Exact Maximum Margin Structure Learning of Bayesian Networks

Robert Peharz, Franz Pernkopf

Recently, there has been much interest in finding globally optimal Bayesian network structures, which is NP-hard in general. These techniques were developed for generative scores and can not be directly extended to discriminative scores, as desired for classification. In this paper, we propose an exact method for finding network structures maximizing the probabilistic soft margin, a successfully applied discriminative score. Our method is based on branch-and-bound techniques within a linear programming framework and maintains an any-time solution, together with worst-case suboptimality bounds. We introduce a set of order constraints for enforcing the network structure to be acyclic, which allows a compact problem representation and the use of general-purpose optimization techniques. In classification experiments, our methods clearly outperform generatively trained network structures and compete with support vector machines.

LPQP for MAP: Putting LP Solvers to Better Use

Patrick Pletscher, Sharon Wulff

MAP inference for general energy functions remains a challenging problem. While most efforts are channeled towards improving the linear programming (LP) based relaxation, this work is motivated by the quadratic programming (QP) relaxation. We propose a novel MAP relaxation that penalizes the Kullback-Leibler divergence between the LP pairwise auxiliary variables, and QP equivalent terms given by the product of the unaries. We develop two efficient algorithms based on variants of this relaxation. The algorithms minimize the non-convex objective using belief propagation and dual decomposition as building blocks. Experiments on synthetic and real-world data show that the solutions returned by our algorithms substantially improve over the LP relaxation.

How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing

Yoram Bachrach, Thore Graepel, Tom Minka, John Guiver

We propose a new probabilistic graphical model that jointly models the difficulties of questions, the abilities of participants and the correct answers to questions in aptitude testing and crowdsourcing settings. We devise an active learning/adaptive testing scheme based on a greedy minimization of expected model entropy, which

allows a more efficient resource allocation by dynamically choosing the next question to be asked based on the previous responses. We present experimental results that confirm the ability of our model to infer the required parameters and demonstrate that the adaptive testing scheme requires fewer questions to obtain the same accuracy as a static test scenario.

Smoothness and Structure Learning by Proxy

Benjamin Yackley, Terran Lane

As data sets grow in size, the ability of learning methods to find structure in them is increasingly hampered by the time needed to search the large spaces of possibilities and generate a score for each that takes all of the observed data into account. For instance, Bayesian networks, the model chosen in this paper, have a super-exponentially large search space for a fixed number of variables. One possible method to alleviate this problem is to use a proxy, such as a Gaussian Process regressor, in place of the true scoring function, training it on a selection of sampled networks. We prove here that the use of such a proxy is well-founded, as we can bound the smoothness of a commonly-used scoring function for Bayesian network structure learning. We show here that, compared to an identical search strategy using the network's exact scores, our proxy-based search is able to get equivalent or better scores on a number of data sets in a fraction of the time.

5A: Learning theory, 16:00–17:40, AT LT 4

Linear Regression with Limited Observation

Elad Hazan, Tomer Koren

We consider the most common variants of linear regression, including Ridge, Lasso and Support-vector regression, in a setting where the learner is allowed to observe only a fixed number of attributes of each example at training time. We present simple and efficient algorithms for these problems: for Lasso and Ridge regression they need the same number of attributes (up to constants) as do full-information algorithms, for reaching a certain accuracy. For Support-vector regression, we require exponentially less attributes compared to the state of the art. By that, we resolve an open problem recently posed by (Cesa-Bianchi et al., 2010). Experiments show the theoretical bounds to be justified by superior performance compared to the state of the art.

Optimizing F-measure: A Tale of Two Approaches

Ye Nan, Kian Ming Chai, Wee Sun Lee, Hai Leong Chieu

F-measures are popular performance metrics, particularly for tasks with imbalanced data sets. Algorithms for learning to maximize F-measures follow two approaches: the empirical utility maximization (EUM) approach learns a classifier having optimal performance on training data, while the decision-theoretic approach learns a probabilistic model and then predict labels with maximum expected F-measure. In this

paper, we investigate the theoretical justifications and connections for these two approaches, and we study the conditions under which one approach is preferable to the other using synthetic and real datasets. Given accurate models, our results suggest that the two approaches are asymptotically equivalent given large training and test sets. Nevertheless, the EUM approach appears to be more robust against model misspecification, and given a good model, the decision-theoretic approach appears to be better for handling rare classes and for a common domain adaptation scenario.

Conditional mean embeddings as regressors

Steffen Grunewalder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, Massi Pontil

We demonstrate an equivalence between reproducing kernel Hilbert space (RKHS) embeddings of conditional distributions and vector-valued regressors. This connection introduces a natural regularized loss function which the RKHS embeddings minimise, providing an intuitive understanding of the embeddings and a solid justification for their use. Furthermore, the equivalence allows the application of vector-valued regression methods and results to the problem of learning conditional distributions. Using this link we derive a sparse version of the embedding by considering alternative formulations. Further, by applying convergence results for vector-valued regression to the embedding problem we derive minimax convergence rates which are approximately $O(\log(n)/n)$ – compared to current state of the art rates of $O(n^{-1/4})$ – and are valid under milder and more intuitive assumptions. These minimax lower rates coincide with upper rates up to a logarithmic factor and show that the embedding method achieves nearly optimal rates. We study our sparse embedding algorithm in a reinforcement learning task where the algorithm shows significant improvement in sparsity over a Cholesky decomposition.

PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification

Emilie Morvant, Sokol Koço, Liva Ralaivola

In this paper, we propose a PAC-Bayes bound for the generalization risk of the Gibbs classifier in the multi-class classification framework. The novelty of our work is the critical use of the confusion matrix of a classifier as an error measure; this puts our contribution in the line of work aiming at dealing with performance measure that are richer than mere scalar criterion such as the misclassification rate. Thanks to very recent and beautiful results on matrix concentration inequalities, we derive two bounds showing that the true confusion risk of the Gibbs classifier is upper-bounded by its empirical risk plus a term depending on the number of training examples in each class. To the best of our knowledge, this is the first PAC-Bayes bounds based on confusion matrices.

Tighter Variational Representations of f-Divergences via Restriction to Probability Measures

Avraham Ruderman, Mark Reid, Darío García-García, James Petterson

We show that the variational representations for f-divergences currently used in the literature can be tightened. This has implications to a number of methods recently proposed based on this representation. As an example application we use our tighter representation to derive a general f-divergence estimator based on two i.i.d. samples and derive the dual program for this estimator that performs well empirically. We also point out a connection between our estimator and MMD.

Agglomerative Bregman Clustering

Matus Telgarsky, Sanjoy Dasgupta

This manuscript develops the theory of agglomerative clustering with Bregman divergences. Geometric smoothing techniques are developed to deal with degenerate clusters. To allow for cluster models based on exponential families with overcomplete representations, Bregman divergences are developed for nondifferentiable convex functions.

The Convexity and Design of Composite Multiclass Losses

Mark Reid, Robert Williamson, Peng Sun

We consider composite loss functions for multiclass prediction comprising a proper (i.e., Fisher-consistent) loss over probability distributions and an inverse link function. We establish conditions for their (strong) convexity and explore their implications. We also show how the separation of concerns afforded by using this composite representation allows for the design of families of losses with the same Bayes risk.

Minimizing The Misclassification Error Rate Using a Surrogate Convex Loss

Shai Ben-David, David Loker, Nathan Srebro, Karthik Sridharan

We carefully study how well minimizing convex surrogate loss functions, corresponds to minimizing the misclassification error rate for the problem of binary classification with linear predictors. In particular, we show that amongst all convex surrogate losses, the hinge loss gives essentially the best possible bound, of all convex loss functions, for the misclassification error rate of the resulting linear predictor in terms of the best possible margin error rate. We also provide lower bounds for specific convex surrogates that show how different commonly used losses qualitatively differ from each other.

5B: Online learning 2, 16:00–17:40, AT LT 5

Hierarchical Exploration for Accelerating Contextual Bandits

Yisong Yue, Sue Ann Hong, Carlos Guestrin

Contextual bandit learning is a popular approach to optimizing recommender systems, but can be slow to converge in practice due to the need for explorating a large feature space. In this paper, we propose a coarse-to-fine hierarchical approach for encoding prior knowledge which drastically reduces the amount of exploration required. Intuitively, user preferences can be reasonably embedded in a coarse low-dimensional feature space that can be explored efficiently, requiring exploration in the high-dimensional space only as necessary. We introduce a bandit algorithm for exploring within this coarse-to-fine spectrum, and prove performance guarantees that depend on how well the coarser feature space captures the user's preferences. We demonstrate substantial improvement over conventional bandit algorithms through extensive simulation as well as a live user study in the setting of personalized news recommendation.

Online Bandit Learning against an Adaptive Adversary: from Regret to Policy Regret

Ofer Dekel, Ambuj Tewari, Raman Arora

Online learning algorithms are designed to learn even when their input is generated by an adversary. The widely-accepted formal definition of an online algorithm's ability to learn is the game-theoretic notion of regret. We argue that the standard definition of regret becomes inadequate if the adversary is allowed to adapt to the online algorithm's actions. We define the alternative notion of pol- icy regret, which attempts to provide a more meaningful way to measure an online algorithm's performance against adaptive adversaries. Focusing on the online bandit setting, we show that no bandit algorithm can guarantee a sublinear policy regret against an adaptive adversary with unbounded memory. On the other hand, if the adversary's memory is unbounded, we present a general technique that converts any bandit algorithm with a sublinear regret bound into an algorithm with a sublinear policy regret bound. We extend this result to other variants of regret, such as switching regret, internal regret, and swap regret.

Decoupling Exploration and Exploitation in Multi-Armed Bandits

Orly Avner, Shie Mannor, Ohad Shamir

We consider a multi-armed bandit problem where the decision maker can explore and exploit different arms at every round. The exploited arm adds to the decision maker's cumulative reward (without necessarily observing the reward) while the explored arm reveals its value. We devise algorithms for this setup and show that the dependence on the number of arms can be much better than the standard square root of the number of arms dependence, according to the behavior of the arms' reward sequences. For the important case of piecewise stationary stochastic bandits, we show a significant improvement over existing algorithms. Our algorithms are based

on a non-uniform sampling policy, which we show is essential to the success of any algorithm in the adversarial setup. We finally show some simulation results on an ultra-wide band channel selection inspired setting indicating the applicability of our algorithms.

Learning the Experts for Online Sequence Prediction

Elad Eban, Amir Globerson, Shai Shalev-Schwartz, Aharon Birnbaum

Online sequence prediction is the problem of predicting the next element of a sequence given previous elements. This problem has been extensively studied in the context of individual sequence prediction, where no prior assumptions are made on the origin of the sequence. Individual sequence prediction algorithms work quite well for long sequences, where the algorithm has enough time to learn the temporal structure of the sequence. However, they might give poor predictions for short sequences. A possible remedy is to rely on the general model of prediction with expert advice, where the learner has access to a set of r experts, each of which makes its own predictions on the sequence. It is well known that it is possible to predict almost as well as the best expert if the sequence length is order of $\log(r)$. But, without firm prior knowledge on the problem, it is not clear how to choose a small set of *good* experts. In this paper we describe and analyze a new algorithm that learns a good set of experts using a training set of previously observed sequences. We demonstrate the merits of our approach by experimenting with the task of click prediction on the web.

Plug-in martingales for testing exchangeability on-line

Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, Volodya Vovk

A standard assumption in machine learning is the exchangeability of data; in other words, the examples are assumed to be generated from the same probability distribution independently. This paper is devoted to testing the assumption of exchangeability on-line: the examples arrive one by one, and after receiving each example we would like to have a valid measure of the degree to which the assumption of exchangeability has been falsified. Such measure are provided by exchangeability martingales. We extend known techniques for constructing exchangeability martingales and show that our new method is competitive to martingales introduced before. Finally we investigate the performance of our testing method on two benchmark datasets, USPS and Statlog Satellite data; for the former, the known techniques give satisfactory results, but for the latter our new more flexible method becomes necessary.

Parallelizing Exploration-Exploitation Tradeoffs with Gaussian Process Bandit Optimization

Thomas Desautels, Andreas Krause, Joel Burdick

Can one parallelize complex exploration-exploitation tradeoffs? As an example, consider the problem of optimal high-throughput experimental design, where we wish

to sequentially design batches of experiments in order to simultaneously learn a surrogate function mapping stimulus to response, as well as identifying the maximum of the function. We formalize the task as a multi-armed bandit problem, where the unknown payoff function is sampled from a Gaussian process (GP), and instead of a single arm, in each round we pull a batch of several arms in parallel. We develop GP-BUCB, a principled algorithm for choosing batches, based on the GP-UCB algorithm for sequential GP optimization. We prove that, perhaps surprisingly, the cumulative regret of the parallel algorithm, compared to the sequential approach, only increases by a constant factor independent of the batch size B, for any fixed B. Our results provide rigorous theoretical support for exploiting parallelism in Bayesian global optimization. We demonstrate the effectiveness of our approach on two real-world applications.

On-Line Portfolio Selection with Moving Average Reversion

Bin Li, Steven C.H. Hoi

On-line portfolio selection has attracted increasing interests in machine learning and AI communities recently. Empirical evidences show that stock's high and low prices are temporary and stock price relatives are likely to follow the mean reversion phenomenon. While the existing mean reversion strategies are shown to achieve good empirical performance on many real datasets, they often make the *single-period mean reversion* assumption, which is not always satisfied in some real datasets, leading to poor performance when the assumption does not hold. To overcome the limitation, this article proposes a *multiple-period mean reversion*, or so-called "Moving Average Reversion' (MAR), and a new on-line portfolio selection strategy named "On-Line Moving Average Reversion" (OLMAR), which exploits MAR by applying powerful online learning techniques. From our empirical results, we found that OLMAR can overcome the drawback of existing mean reversion algorithms and achieve significantly better results, especially on the datasets where the existing mean reversion algorithms failed. In addition to superior trading performance, OLMAR also runs extremely fast, further supporting its practical applicability to a wide range of applications.

Exponential Regret Bounds for Gaussian Process Bandits with Deterministic Observations

Nando de Freitas, Alex Smola, Masrour Zoghi

This paper analyzes the problem of Gaussian process (GP) bandits with deterministic observations. The analysis uses a branch and bound algorithm that is related to the UCB algorithm of (Srinivas et al, 2010). For GPs with Gaussian observation noise, with variance strictly greater than zero, Srinivas et al proved that the regret vanishes at the approximate rate of $O(1/\sqrt{t})$, where t is the number of observations.

To complement their result, we attack the deterministic case and attain a much faster exponential convergence rate. Under some regularity assumptions, we show that the regret decreases asymptotically according to $O(e^{-\frac{\tau t}{(\ln t)^{d/4}}})$ with high probability. Here, d is the dimension of the search space and tau is a constant that depends on the behaviour of the objective function near its global maximum.

5C: Neural networks and deep learning 2, 16:00–17:40, AT LT 1

Large-Scale Feature Learning With Spike-and-Slab Sparse Coding

Ian Goodfellow, Aaron Courville, Yoshua Bengio

We consider the problem of object recognition with a large number of classes. In order to scale existing feature learning algorithms to this setting, we introduce a new feature learning and extraction procedure based on a factor model we call spike-and-slab sparse coding (S3C). Prior work on this model has not prioritized the ability to exploit parallel architectures and scale to the enormous problem sizes needed for object recognition. We present an inference proce- dure appropriate for use with GPUs which allows us to dramatically increase both the training set size and the amount of latent factors. We demonstrate that this approach improves upon the supervised learning capabilities of both sparse coding and the ss-RBM on the CIFAR-10 dataset. We use the CIFAR-100 dataset to demonstrate that our method scales to large numbers of classes better than previous methods. Finally, we use our method to win the NIPS 2011 Workshop on Challenges In Learning Hierarchical Models' Transfer Learning Challenge.

Learning Local Transformation Invariance with Restricted Boltzmann Machines

Kihyuk Sohn, Honglak Lee

The difficulty of developing feature learning algorithms that are robust to the novel transformations (e.g., scale, rotation, or translation) has been a challenge in many applications (e.g., object recognition problems). In this paper, we address this important problem of transformation invariant feature learning by introducing the transformation matrices into the energy function of the restricted Boltzmann machines. Specifically, the proposed transformation-invariant restricted Boltzmann machines not only learn the diverse patterns by explicitly transforming the weight matrix, but it also achieves the invariance of the feature representation via probabilistic max pooling of hidden units over the set of transformations. Furthermore, we show that our transformation-invariant feature learning framework is not limited to this specific algorithm, but can be also extended to many unsupervised learning methods, such as an autoencoder or sparse coding. To validate, we evaluate our algorithm on several benchmark image databases such as MNIST variation, CIFAR-10, and STL-10

as well as the customized digit datasets with significant transformations, and show very competitive classification performance to the state-of-the-art. Besides the image data, we apply the method for phone classification tasks on TIMIT database to show the wide applicability of our proposed algorithms to other domains, achieving state-of-the-art performance.

Building high-level features using large scale unsupervised learning

Quoc Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Jeff Dean, Andrew Ng

We consider the challenge of building feature detectors for high-level concepts from only unlabeled data. For example, we would like to understand if it is possible to learn a face detector using only unlabeled images downloaded from the Internet. To answer this question, we trained a 9-layered locally connected sparse autoencoder with pooling and local contrast normalization on a large dataset of images (which has 10 million images, each image has 200x200 pixels). On contrary to what appears to be a widely-held negative belief, our experimental results reveal that it is possible to achieve a face detector via only unlabeled data. Control experiments show that the feature detector is robust not only to translation but also to scaling and 3D rotation. Also via recognition and visualization, we find that the same network is sensitive to other high-level concepts such as cat faces and human bodies.

On multi-view feature learning

Roland Memisevic

Sparse coding is a common approach to learning local features for object recognition. Recently, there has been an increasing interest in learning features from spatio-temporal, binocular, or other multi-observation data, where the goal is to encode the relationship between images rather than the content of a single image. We discuss the role of multiplicative interactions and of squaring non-linearities in learning such relations. In particular, we show that training a sparse coding model whose filter responses are multiplied or squared amounts to jointly diagonalizing a set of matrices that encode image transformations. Inference amounts to detecting rotations in the shared eigenspaces. Our analysis helps explain recent experimental results showing that Fourier features and circular Fourier features emerge when training complex cell models on translating or rotating images. It also shows how learning about transformations makes it possible to learn invariant features.

Learning to Label Aerial Images from Noisy Data

Volodymyr Mnih, Geoffrey Hinton

When training a system to label images, the amount of labeled training data tends to be a limiting factor. We consider the task of learning to label aerial images from existing maps. These provide abundant labels, but the labels are often incomplete

and sometimes poorly registered. We propose two robust loss functions for dealing with these kinds of label noise and use the loss functions to train a deep neural network on two challenging aerial image datasets. The robust loss functions lead to big improvements in performance and our best system substantially outperforms the best published results on the task we consider.

5D: Sparsity and compressed sensing, 16:00–17:40, AT LT 2

Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering

Gael Varoquaux, Alexandre Gramfort, Bertrand Thirion

Functional neuroimaging can measure the brain's response to an external stimulus. It is used to perform brain mapping: identifying from these observations the brain regions involved. This problem can be cast into a linear supervised learning task where the neuroimaging data are used as predictors for the stimulus. Brain mapping is then seen as a support recovery problem. On functional MRI (fMRI) data, this problem is particularly challenging as i) the number of samples is small due to limited acquisition time and ii) the variables are strongly correlated. We propose to overcome these difficulties using sparse regression models over new variables obtained by clustering of the original variables. The use of randomization techniques, e.g. bootstrap samples, and hierarchical clustering of the variables improves the recovery properties of sparse methods. We demonstrate the benefit of our approach on an extensive simulation study as well as two publicly available fMRI datasets.

Estimation of Simultaneously Sparse and Low Rank Matrices

Pierre-André Savalle, Emile Richard, Nicolas Vayatis

The paper introduces a penalized matrix estimation procedure aiming at solutions which are sparse and low-rank at the same time. Such structures arise in the context of social networks or protein interactions where underlying graphs have adjacency matrices which are block-diagonal in the appropriate basis. We introduce a convex mixed penalty which involves ℓ_1-norm and trace norm simultaneously. We obtain an oracle inequality which indicates how the two effects interact according to the nature of the target matrix. We bound generalization error in the link prediction problem. We also develop proximal descent strategies to solve the the optimization problem efficiently and evaluate performance on synthetic and real data sets.

Multi-level Lasso for Sparse Multi-task Regression

Aurelie Lozano

We present a flexible formulation for variable selection in multi-task regression to allow for discrepancies in the estimated sparsity patterns accross the multiple tasks, while leveraging the common structure among them. Our approach is based on

an intuitive decomposition of the regression coefficients into a product between a component that is common to all tasks and another component that captures task-specificity. This decomposition yields the Multi-level Lasso objective that can be solved efficiently via alternating optimization. The analysis of the "orthonormal design" case reveals some interesting insights on the nature of the shrinkage performed by our method, compared to that of related work. Theoretical guarantees are provided on the consistency of Multi-level Lasso. Simulations and empirical study of micro-array data further demonstrate the value of our framework.

Efficient Euclidean Projections onto the Intersection of Norm Balls

Wei YU, Hao Su, Li Fei-Fei

Using sparse-inducing norms to learn robust models has received increasing attention from many fields for its attractive properties. Projection-based methods have been widely applied to learning tasks constrained by such norms. As a key building block of these methods, an efficient operator for Euclidean projection onto the intersection of ℓ_1 and $\ell_{1,q}$ norm balls ($q = 2$ or ∞) is proposed in this paper. We prove that the projection can be reduced to finding the root of an auxiliary function which is piecewise smooth and monotonic. Hence, a bisection algorithm is sufficient to solve the problem. We show that the time complexity of our solution is $O(n + g \log g)$ for $q = 2$ and $O(n \log n)$ for $q = \infty$, where n is the dimensionality of the vector to be projected and g is the number of disjoint groups; we confirm this complexity by experimentation. Empirical study reveals that our method achieves significantly better performance than classical methods in terms of running time and memory usage. We further show that embedded with our efficient projection operator, projection-based algorithms can solve regression problems with composite norm constraints more efficiently than other methods and give superior accuracy.

Learning Efficient Structured Sparse Models

Alex Bronstein, Pablo Sprechmann, Guillermo Sapiro

We present a comprehensive framework for structured sparse coding and modeling extending the recent ideas of using learnable fast regressors to approximate exact sparse codes. For this purpose, we develop a novel block-coordinate proximal splitting method for the iterative solution of hierarchical sparse coding problems, and show an efficient feed forward architecture derived from its iteration. This architecture faithfully approximates the exact structured sparse codes with a fraction of the complexity of the standard optimization methods. We also show that by using different training objective functions, learnable sparse encoders are no longer restricted to be mere approximants of the exact sparse code for a pre-given dictionary, as in earlier formulations, but can be rather used as full-featured sparse encoders or even modelers. A simple implementation shows several orders of magnitude speedup compared

to the state-of-the-art at minimal performance degradation, making the proposed framework suitable for real time and large-scale applications.

5E: Latent-Variable Models and Topic Models, 16:00–17:35, AT LT 3

Max-Margin Nonparametric Latent Feature Models for Link Prediction

Jun Zhu

Recent work on probabilistic latent feature models have shown great promise in predicting unseen links in social network and relational data. We present a max-margin nonparametric latent feature model, which unites the ideas of max-margin learning and Bayesian nonparametrics to discover discriminative latent features for link prediction and automatically infer the unknown latent social dimension. By minimizing a hinge-loss using the linear expectation operator, we can perform posterior inference efficiently without dealing with a highly nonlinear link likelihood function; by using a fully-Bayesian formulation, we can avoid tuning regularization constants. Experimental results on real datasets appear to demonstrate the benefits inherited from max-margin learning and fully-Bayesian nonparametric inference.

Canonical Trends: Detecting Trend Setters in Web Data

Felix Biessmann, Jens-Michalis Papaioannou, Mikio Braun, Andreas Harth

The web offers large amounts of information most of which is just copied or rephrased, a phenomenon that is often called trend. A central problem in the context of web data mining is to detect those web sources that publish information which will give rise to a trend first. Here we present a simple and efficient method for finding trends dominating a pool of web sources and identifying those web sources that publish the information relevant to this trend before other websites do so. We validate our approach on real data collected from the most influential technology news feeds.

Variational Inference in Non-negative Factorial Hidden Markov Models for Efficient Audio Source Separation

Gautham Mysore, Maneesh Sahani

The last decade has seen substantial work on the use of non-negative matrix factorization (NMF) and its probabilistic counterparts for audio source separation. Although able to capture audio spectral structure well, these models neglect the non-stationarity and temporal dynamics that are important properties of audio. The recently proposed non-negative factorial hidden Markov model (N-FHMM) introduces a temporal dimension and improves source separation performance. However, the factorial nature of this model makes the complexity of inference exponential in the number of sound sources. Here, we present a Bayesian variant of the N-FHMM suited to an efficient variational inference algorithm, whose complexity is linear in the number of sound sources. Our algorithm performs comparably to exact inference

in the original NFHMM but is significantly faster. In typical configurations of the N-FHMM, our method achieves around a 30x increase in speed.

Sparse stochastic inference for latent Dirichlet allocation

David Mimno, Matt Hoffman, David Blei

We present a hybrid algorithm for Bayesian topic models that combines the efficiency of sparse Gibbs sampling with the scalability of online stochastic inference. The resulting method scales to a corpus of 1.2 million books comprising 33 billion words with thousands of topics on one CPU. This approach reduces the bias of variational inference and generalizes to many Bayesian hidden-variable models.

Dirichlet Process with Mixed Random Measures: A Nonparametric Topic Model for Labeled Data

Dongwoo Kim, Suin Kim, Alice Oh

We describe a nonparametric topic model for labeled data. The model uses a mixture of random measures (MRM) as a base distribution of the Dirichlet process (DP) of the HDP framework, so we call it the DP-MRM. To model labeled data, we define a DP distributed random measure for each label, and the resulting model generates an unbounded number of topics for each label. We apply DP-MRM on single-labeled and multi-labeled corpora of documents and compare the performance on label prediction with LDA-SVM and Labeled-LDA. We further enhance the model by incorporating ddCRP and modeling multi-labeled images for image segmentation and object labeling, comparing the performance with nCuts and rddCRP.

Rethinking Collapsed Variational Bayes Inference for LDA

Issei Sato, Hiroshi Nakagawa

We provide a unified view for existing inference algorithms of latent Dirichlet allocation by using the alpha-divergence projection. In particular, we explain the collapsed variational Bayes inference with zero-order information, called the CVB0 inference, in terms of the local alpha-divergence projection. We show that the CVB0 inference is composed of two different divergence projections: alpha=1 and -1.

Capturing topical content with frequency and exclusivity

Jonathan Bischof, Edoardo Airoldi

Recent work in text analysis commonly describes topics in terms of their most frequent words, but the exclusivity of words to topics is equally important for communicating their content.We introduce Hierarchical Poisson Convolution (HPC), a model which infers regularized estimates of the differential use of words across topics as well as word frequency within topics. HPC uses known hierarchical structure on human labeled topics to make focused comparisons of differential usage within each branch of the tree. We develop a parallelized Hamiltonian Monte Carlo sampler that allows for fast and scalable computation.

Main Conference Abstracts: Friday, June 29

6A: Semi-supervised learning, 08:40–10:00, AT LT 4

A convex relaxation for weakly supervised classifiers

Armand Joulin, Francis Bach

This paper introduces a general multi-class approach to weakly supervised classification. Inferring the labels and learning the parameters of the model is usually done jointly through a block-coordinate descent algorithm such as expectation-maximization (EM), which may lead to local minima. To avoid this problem, we propose a cost function based on a convex relaxation of the soft-max loss. We then propose an algorithm specifically designed to efficiently solve the corresponding semidefinite program (SDP). Empirically, our method compares favorably to standard ones on different datasets for multiple instance learning and semi-supervised learning as well as on clustering tasks.

Semi-Supervised Learning of Class Balance under Class-Prior Change by Distribution Matching

Marthinus Du Plessis, Masashi Sugiyama

In real-world classification problems, the class balance in the training dataset does not necessarily reflect that of the test dataset, which can cause significant estimation bias. If the class ratio of the test dataset is known, instance re-weighting or resampling allows systematical bias correction. However, learning the class ratio of the test dataset is challenging when no labeled data is available from the test domain. In this paper, we propose to estimate the class ratio in the test dataset by matching probability distributions of training and test input data. We demonstrate the utility of the proposed approach through experiments.

A Simple Algorithm for Semi-supervised Learning with Improved Generalization Error Bound

Ming Ji, Tianbao Yang, Binbin Lin, Rong Jin, Jiawei Han

In this work, we develop a simple algorithm for semi-supervised regression. The key idea is to use the top eigenfunctions of integral operator derived from both labeled and unlabeled examples as the basis functions and learn the prediction function by a simple linear regression. We show that under appropriate assumptions about the integral operator, this approach is able to achieve an improved regression error bound better than existing bounds of supervised learning. We also verify the effectiveness of the proposed algorithm by an empirical study.

Information-theoretic Semi-supervised Metric Learning via Entropy Regular-

ization

Gang Niu, Bo Dai, Makoto Yamada, Masashi Sugiyama

We propose a general information-theoretic approach called Seraph for semi-supervised metric learning that does not rely upon the manifold assumption. It maximizes/minimizes entropies of labeled/unlabeled data pairs in the supervised/unsupervised part, which allows these two parts to be integrated in a natural and meaningful way. Furthermore, it is equipped with the hyper-sparsity: Given a certain probabilistic model parameterized by the learned metric, the posterior distribution and the resultant projection matrix are both 'sparse'. Consequently, the metric learned by Seraph possesses high discriminability even under a noisy environment. The optimization problem of Seraph can be solved efficiently and stably by an EM-like scheme, where the E-Step is analytical and the M-Step is convex and 'smooth'. Experiments demonstrate that Seraph compares favorably with many well-known metric learning methods.

Cross Language Text Classification via Subspace Co-regularized Multi-view Learning

Yuhong Guo, Min Xiao

In many multilingual text classification problems, the documents in different languages often share the same set of categories. To reduce the labeling cost of training a classification model for each individual language, it is important to transfer knowledge gained from the labeled data in one language to another language by conducting cross language classification. In this paper we develop a novel subspace co-regularized multi-view learning method for cross language text classification. This method is built on parallel corpora produced by machine translation. It jointly minimizes the training error of each classifier in each language while penalizing the distance between the subspace representations of parallel documents. Our empirical study on a large set of cross language text classification tasks shows the proposed method consistently outperforms the alternative inductive methods, domain adaptation methods, and multi-view learning methods.

Using CCA to improve CCA: A new spectral method for estimating vector models of words

Paramveer Dhillon, Jordan Rodu, Dean Foster, Lyle Ungar

Unlabeled data is often used to learn representations which can be used to supplement baseline features in a supervised learner. For example, for text applications where the words lie in a very high dimensional space (the size of the vocabulary), one can learn a low rank "dictionary" by an eigen-decomposition of the word co-occurrence matrix (e.g. using PCA or CCA). In this paper, we present a new spectral method based on CCA to learn an *eigenword* dictionary. Our improved procedure computes two

set of CCAs, the first one between the left and right contexts of the given word and the second one between the projections resulting from this CCA and the word itself. We prove theoretically that this two-step procedure has lower sample complexity than the simple single step procedure and also illustrate the empirical efficacy of our approach and the richness of representations learned by our Two Step CCA (TSCCA) procedure on the tasks of POS tagging and sentiment classification.

Semi-Supervised Collective Classification via Hybrid Label Regularization

Luke McDowell, David Aha

Many classification problems involve data instances that are interlinked with each other, such as webpages connected by hyperlinks. Techniques for 'collective classification' (CC) often increase accuracy for such data graphs, but usually require a fully-labeled training graph. In contrast, we examine how to improve the semi-supervised learning of CC models when given only a sparsely-labeled graph, a common situation. We first describe how to use novel combinations of classifiers to exploit the different characteristics of the relational features vs. the non-relational features. We also extend the ideas of 'label regularization' to such hybrid classifiers, enabling them to leverage the unlabeled data to bias the learning process. We find that these techniques, which are efficient and easy to implement, significantly increase accuracy on three real datasets. In addition, our results explain conflicting findings from prior related studies.

6B: Reinforcement learning 3, 08:40–10:00, AT LT 5

Compositional Planning Using Optimal Option Models

David Silver, Kamil Ciosek

In this paper we introduce a framework for option model composition. Option models are temporal abstractions that, like macro-operators in classical planning, jump directly from a start state to an end state. Prior work has focused on constructing option models from primitive actions, by intra-option model learning; or on using option models to construct a value function, by inter-option planning. We present a unified view of intra- and inter-option model learning, based on a major generalisation of the Bellman equation. Our fundamental operation is the recursive composition of option models into other option models. This key idea enables compositional planning over many levels of abstraction. We illustrate our framework using a dynamic programming algorithm that simultaneously constructs optimal option models for multiple subgoals, and also searches over those option models to provide rapid progress towards other subgoals.

Learning Parameterized Skills

Bruno Da Silva, George Konidaris, Andrew Barto

We introduce a method for constructing a single skill capable of solving any tasks drawn from a distribution of parameterized reinforcement learning problems. The method draws example tasks from a distribution of interest and uses the corresponding learned policies to estimate the geometry of the lower-dimensional manifold on which the skill policies lie. This manifold models how policy parameters change as the skill parameters are varied. We then apply non-linear regression to construct a parameterized skill by predicting policy parameters from task parameters. We evaluate our method on an underactuated simulated robotic arm tasked with learning to accurately throw darts at any target location around it.

Safe Exploration in Markov Decision Processes

Teodor Mihai Moldovan, Pieter Abbeel

In unknown or partially unknown environments exploration is necessary to learn how to perform well. Existing reinforcement learning algorithms provide strong exploration guarantees, but they tend to rely on an ergodicity assumption. While ergodicity is a more generally applicable property, it is simplest to state for discrete state spaces, and its essence is that any state is reachable from any other state within some finite amount of time with non-zero probability by following a suitable policy. This assumption allows for exploration algorithms that operate by favoring visiting previously unvisited (or rarely visited) states. For most physical systems this assumption is completely impractical as the systems would break before any reasonable exploration has taken place, i.e., most physical systems don't satisfy the ergodicity assumption. In this paper we address the need for *safe* exploration methods in Markov decision processes. We first propose a general formulation of safety that can be thought of as forcing ergodicity. We show that imposing the safety constraint exactly is NP-hard. We present an efficient approximation algorithm for guaranteed safe, but potentially suboptimal, exploration. At the core of our approach is an optimization formulation in which the constraints restrict attention to guaranteed safe policies. The objective favors exploration policies. Our framework is compatible with the majority of previously proposed exploration methods, which rely on an exploration bonus, which we can replicate in the objective. Our experiments show that our method is able to safely explore state-spaces in which classical exploration methods get stuck.

Modelling transition dynamics in MDPs with RKHS embeddings

Steffen Grunewalder, Guy Lever, Luca Baldassarre, Massi Pontil, Arthur Gretton

We propose a new, nonparametric approach to learning and representing transition dynamics in Markov decision processes (MDPs), which can be combined easily with dynamic programming methods for policy optimisation and value estimation. This approach makes use of a recently developed representation of conditional distribu-

tions as *embeddings* in a reproducing kernel Hilbert space (RKHS). Such representations bypass the need for estimating transition probabilities or densities, and apply to any domain on which kernels can be defined. This avoids the need to calculate intractable integrals, since expectations are represented as RKHS inner products whose computation has linear complexity in the number of points used to represent the embedding. We are able to provide guarantees for the proposed applications in MDPs: in the context of a value iteration algorithm, we prove convergence to either the optimal policy, or to the closest projection of the optimal policy in our model class (an RKHS), under reasonable assumptions. In experiments, we investigate a learning task in a typical classical control setting (the under-actuated pendulum), and on a navigation problem where only images from a sensor are observed. For policy optimisation we compare with least-squares policy iteration where a Gaussian process is used for value function estimation. For value estimation we also compare to the recent NPDP method. Our approach achieves better performance in all experiments.

6C: Applications, 08:40–10:00, AT LT 1

A Joint Model of Language and Perception for Grounded Attribute Learning

Cynthia Matuszek, Nichoals FitzGerald, Luke Zettlemoyer, Liefeng Bo, Dieter Fox

As robots become more ubiquitous and capable, it becomes ever more important to enable untrained users to easily interact with them. Recently, this has led to study of the language grounding problem, where the goal is to extract representations of the meanings of natural language tied to perception and actuation in the physical world. In this paper, we present an approach for joint learning of language and perception models for grounded attribute induction. Our perception model includes attribute classifiers, for example to detect object color and shape, and the language model is based on a probabilistic categorial grammar that enables the construction of rich, compositional meaning representations. The approach is evaluated on the task of interpreting sentences that describe sets of objects in a physical workspace. We demonstrate accurate task performance and effective latent-variable concept induction in physical grounded scenes.

Predicting Manhole Events in New York City

Cynthia Rudin, Rebecca Passonneau, Axinia Radeva, Steve Ierome, Delfina Isaac

We present a knowledge discovery and data mining process developed as part of the Columbia/Con Edison project on manhole event prediction. This process can assist with real-world prioritization problems that involve raw data in the form of noisy documents requiring significant amounts of pre-processing. The documents are linked to a set of instances to be ranked according to prediction criteria. In the case of

manhole event prediction, which is a new application for machine learning, the goal is to rank the electrical grid structures in New York City (manholes and service boxes) according to their vulnerability to serious manhole events such as fires, explosions and smoking manholes. Our ranking results are currently used to help prioritize repair work on the Manhattan, Brooklyn, and Bronx electrical grids.

Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription

Nicolas Boulanger-Lewandowski, Yoshua Bengio, Pascal Vincent

We investigate the problem of modeling symbolic sequences of polyphonic music in a completely general piano-roll representation. We introduce a probabilistic model based on distribution estimators conditioned on a recurrent neural network that is able to discover temporal dependencies in high-dimensional sequences. Our approach outperforms many traditional models of polyphonic music on a variety of realistic datasets. We show how our musical language model can serve as a symbolic prior to improve the accuracy of polyphonic transcription.

Learning Object Arrangements in 3D Scenes using Human Context

Yun Jiang, Marcus Lim, Ashutosh Saxena

We consider the problem of learning object arrangements in a 3D scene. The key idea here is to learn how objects relate to human skeletons based their affordances, ease of use and reachability. We design appropriate density functions based on 3D spatial features to capture this. We then learn the distribution of human poses in a scene using Dirichlet processes. This allows our algorithm to reason about arrangement of the objects in the room. This is in contrast to previous approaches that learn context by learning object - object context such as co-occurrence relationships. We tested our algorithm extensively on 20 different rooms with a total of 47 objects. Our algorithm predicted correct placements with an error of 1.6 meters from the ground truth and received a score of 4.3/5 in arranging five real scenes compared to 3.7 of the best baseline.

6D: Time-Series Analysis, 08:40–10:00, AT LT 2

Learning the Dependence Graph of Time Series with Latent Factors

Ali Jalali, Sujay Sanghavi

This paper considers the problem of learning, from samples, the dependency structure of a system of linear stochastic differential equations, when some of the variables are latent. We observe the time evolution of some variables, and never observe other variables; from this, we would like to find the dependency structure of the observed variables – separating out the spurious interactions caused by the latent variables' time series. We develop a new convex optimization based method to do so in the

case when the number of latent variables is smaller than the number of observed ones. For the case when the dependency structure between the observed variables is sparse, we theoretically establish a high-dimensional scaling result for structure recovery. We verify our theoretical result with both synthetic and real data (from the stock market).

Improved Estimation in Time Varying Models

Doina Precup, Philip Bachman

Locally adapted parametrizations of a model (such as locally weighted regression) are expressive but often suffer from high variance. We describe an approach for reducing the variance, based on the idea of estimating simultaneously a transformed space for the model, as well as locally adapted parameterizations in this new space. We present a new problem formulation that captures this idea and illustrate it in the important context of time-varying models. We develop an algorithm for learning a set of bases for approximating a time-varying sparse network; each learned basis constitutes an archetypal sparse network structure. We also provide an extension for learning task-specific bases. We present empirical results on synthetic data sets, as well as on a BCI EEG classification task.

Bayesian Cointegration

Chris Bracegirdle, David Barber

Cointegration is an important topic for time-series, and describes a relationship between two series in which a linear combination is stationary. Classically, the test for cointegration is based on a two stage process in which first the linear relation between the series is estimated by Ordinary Least Squares. Subsequently a unit root test is performed on the residuals. A well-known deficiency of this classical approach is that is can lead to erroneous conclusions about the presence of cointegration. As an alternative, we present a coherent framework for estimating whether cointegration exists using Bayesian inference which is empirically superior to the classical approach. Finally, we apply our technique to model segmented cointegration in which cointegration may exist only for limited time. In contrast to previous approaches our model makes no restriction on the number of possible cointegration segments.

Sparse-GEV: Sparse Latent Space Model for Multivariate Extreme Value Time Serie Modeling

Yan Liu, Taha Bahadori, Hongfei Li

In many applications of time series models, such as climate analysis or social media analysis, we are often interested in extreme events, such as heatwave, wind gust, and burst of topics. These time series data usually exhibit a heavy-tailed distribution rather than a normal distribution. This poses great challenges to existing approaches due to the significantly different assumptions on the data distributions

and the lack of sufficient past data on extreme events. In this paper, we propose the sparse-GEV model, a latent state model based on the theory of extreme value modeling to automatically learn sparse temporal dependence and make predictions. Our model is theoretically significant because it is among the first models to learn *sparse* temporal dependencies between multivariate extreme value time series. We demonstrate the superior performance of our algorithm compared with state-of-art methods, including Granger causality, copula approach, and transfer entropy, on one synthetic dataset, one climate dataset and one Twitter dataset.

6E: Graph-based learning, 08:40–10:00, AT LT 3

Shortest path distance in k-nearest neighbor graphs

Morteza Alamgir, Ulrike von Luxburg

Consider a weighted or unweighted k-nearest neighbor graph that has been built on n data points drawn randomly according to some density p on R^d. We study the convergence of the shortest path distance in such graphs as the sample size n tends to infinity. We prove that for unweighted kNN graphs, this distance converges to an unpleasant distance function on the underlying space whose properties are detrimental to machine learning. The behavior of weighted kNN graphs depends on the choice of the weights. Our results have important consequences for machine learning applications such as manifold learning or semi-supervised learning.

Submodular Inference of Diffusion Networks from Multiple Trees

Manuel Gomez Rodriguez, Bernhard Schölkopf

Diffusion and propagation of information, influence and diseases take place over increasingly larger networks. We observe when a node copies information, makes a decision or becomes infected but networks are often hidden or unobserved. Since networks are highly dynamic, changing and growing rapidly, we only observe a relatively small set of cascades before a network changes significantly. Scalable network inference based on a small cascade set is then necessary for understanding the rapidly evolving dynamics that govern diffusion. In this article, we develop a scalable approximation algorithm with provable near-optimal performance which achieves a high accuracy in such scenario, solving an open problem first introduced by Gomez-Rodriguez et al (2010). Experiments on synthetic and real diffusion data show that our algorithm in practice achieves an optimal trade-off between accuracy and running time.

Influence Maximization in Continuous Time Diffusion Networks

Manuel Gomez Rodriguez, Bernhard Schölkopf

The problem of finding the optimal set of source nodes in a diffusion network that maximizes the spread of information, influence, and diseases in a limited amount of time depends dramatically on the underlying temporal dynamics of the network.

However, this still remains largely unexplored to date. To this end, given a network and its temporal dynamics, we first describe how continuous time Markov chains allow us to analytically compute the average total number of nodes reached by a diffusion process starting in a set of source nodes. We then show that selecting the set of most influential source nodes in the continuous time influence maximization problem is NP-hard and develop an efficient approximation algorithm with provable near-optimal performance. Experiments on synthetic and real diffusion networks show that our algorithm outperforms other state of the art algorithms by at least 20% and is robust across different network topologies.

Latent Multi-group Membership Graph Model

Myunghwan Kim, Jure Leskovec

We develop the Latent Multi-group Membership Graph (LMMG) model, a model of networks with rich node feature structure. In the LMMG model, each node belongs to multiple groups and each latent group models the occurrence of links as well as the node feature structure. The LMMG can be used to summarize the network structure, to predict links between the nodes, and to predict missing features of a node. We derive efficient inference and learning algorithms and evaluate the predictive performance of the LMMG on several social and document network datasets.

The Most Persistent Soft-Clique in a Set of Sampled Graphs

Novi Quadrianto, Chao Chen, Christoph Lampert

When searching for characteristic subpatterns in potentially noisy graph data, it appears self-evident that having multiple observations would be better than having just one. However, it turns out that the inconsistencies introduced when different graph instances have different edge sets pose a serious challenge. In this work we address this challenge for the problem of finding maximum weighted cliques. We introduce the concept of most persistent soft-clique. This is subset of vertices, that 1) is almost fully or at least densely connected, 2) occurs in all or almost all graph instances, and 3) has the maximum weight. We present a measure of clique-ness, that essentially counts the number of edge missing to make a subset of vertices into a clique. With this measure, we show that the problem of finding the most persistent soft-clique problem can be cast either as: a) a max-min two person game optimization problem, or b) a min-min soft margin optimization problem. Both formulations lead to the same solution when using a partial Lagrangian method to solve the optimization problems. By experiments on synthetic data and on real social network data, we show that the proposed method is able to reliably find soft cliques in graph data, even if that is distorted by random noise or unreliable observations.

Two Manifold Problems with Applications to Nonlinear System Identification

Byron Boots, Geoff Gordon

Recently, there has been much interest in spectral approaches to learning manifolds—so-called kernel eigenmap methods. These methods have had some successes, but their applicability is limited because they are not robust to noise. To address this limitation, we look at two-manifold problems, in which we simultaneously reconstruct two related manifolds, each representing a different view of the same data. By solving these interconnected learning problems together, two-manifold algorithms are able to succeed where a non-integrated approach would fail: each view allows us to suppress noise in the other, reducing bias. We propose a class of algorithms for two-manifold problems, based on spectral decomposition of cross-covariance operators in Hilbert space and discuss when two-manifold problems are useful. Finally, we demonstrate that solving a two-manifold problem can aid in learning a nonlinear dynamical system from limited data.

Incorporating Causal Prior Knowledge as Path-Constraints in Bayesian Networks and Maximal Ancestral Graphs

Giorgos Borboudakis, Ioannis Tsamardinos

We consider the incorporation of causal knowledge about the presence or absence of (possibly indirect) causal relations into a causal model. Such causal relations correspond to directed paths in a causal model. This type of knowledge naturally arises from experimental data, among others. Specifically, we consider the formalisms of Causal Bayesian Networks and Maximal Ancestral Graphs and their Markov equivalence classes: Partially Directed Acyclic Graphs and Partially Oriented Ancestral Graphs. We introduce sound and complete procedures which are able to incorporate causal prior knowledge in such models. In simulated experiments, we show that often considering even a few causal facts leads to a significant number of new inferences. In a case study, we also show how to use real experimental data to infer causal knowledge and incorporate it into a real biological causal network.

7A: Invited Applications, 10:30–12:00, AT LT 4

Conversational Speech Transcription Using Context-Dependent Deep Neural Networks

Dong Yu, Frank Seide, Gang Li

Context-Dependent Deep-Neural-Network HMMs, or CD-DNN-HMMs, combine the classic artificial-neural-network HMMs with traditional context-dependent acoustic modeling and deep-belief-network pre-training. CD-DNN-HMMs greatly outperform conventional CD-GMM (Gaussian mixture model) HMMs: The word error rate is reduced by up to one third on the difficult benchmarking task of speaker-independent single-pass transcription of telephone conversations.

Data-driven Web Design

Ranjitha Kumar, Jerry Talton, Salman Ahmad, Scott Klemmer

This short paper summarizes challenges and opportunities of applying machine learning methods to Web design problems, and describes how structured prediction, deep learning, and probabilistic program induction can enable useful interactions for designers. We intend for these techniques to foster new work in data-driven Web design.

Learning the Central Events and Participants in Unlabeled Text

Nathanael Chambers, Dan Jurafsky

The majority of information on the Internet is expressed in written text. Understanding and extracting this information is crucial to building intelligent systems that can organize this knowledge. Today, most algorithms focus on learning atomic facts and relations. For instance, we can reliably extract facts like 'Annapolis is a City' by observing redundant word patterns across a corpus. However, these facts do not capture richer knowledge like the way detonating a bomb is related to destroying a building, or that the perpetrator who was convicted must have been arrested. A structured model of these events and entities is needed for a deeper understanding of language. This talk describes unsupervised approaches to learning such rich knowledge.

Exemplar-SVMs for Visual Object Detection, Label Transfer and Image Retrieval

Tomasz Malisiewicz, Abhinav Shrivastava, Abhinav Gupta, Alexei Efros

Today's state-of-the-art visual object detection systems are based on three key components: 1) sophisticated features (to encode various visual invariances), 2) a powerful classifier (to build a discriminative object class model), and 3) lots of data (to use in large-scale hard-negative mining). While conventional wisdom tends to attribute the success of such methods to the ability of the classifier to generalize across the positive class instances, here we report on empirical findings suggesting that this might not necessarily be the case. We have experimented with a very simple idea: to learn a separate classifier for each positive object instance in the dataset. In this setup, no generalization across the positive instances is possible by definition, and yet, surprisingly, we did not observe any drastic drop in performance compared to the standard, category-based approaches.

Learning Force Control Policies for Compliant Robotic Manipulation

Mrinal Kalakrishnan, Ludovic Righetti, Peter Pastor, Stefan Schaal

Developing robots capable of fine manipulation skills is of major importance in order to build truly assistive robots. These robots need to be compliant in their actuation and control in order to operate safely in human environments. Manipulation tasks imply complex contact interactions with the external world, and involve reasoning

about the forces and torques to be applied. Planning under contact conditions is usually impractical due to computational complexity, and a lack of precise dynamics models of the environment. We present an approach to acquiring manipulation skills on compliant robots through reinforcement learning. The initial position control policy for manipulation is initialized through kinesthetic demonstration. This policy is augmented with a force/torque profile to be controlled in combination with the position trajectories. The Policy Improvement with Path Integrals (PI^2) algorithm is used to learn these force/torque profiles by optimizing a cost function that measures task success. We introduce a policy representation that ensures trajectory smoothness during exploration and learning. Our approach is demonstrated on the Barrett WAM robot arm equipped with a 6-DOF force/torque sensor on two different manipulation tasks: opening a door with a lever door handle, and picking up a pen off the table. We show that the learnt force control policies allow successful, robust execution of the tasks.

7B: Reinforcement learning 4, 10:30–12:00, AT LT 5

Agnostic System Identification for Model-Based Reinforcement Learning

Stephane Ross, Drew Bagnell

A fundamental problem in control is to learn a model of a system from observations that is useful for controller synthesis. To provide good performance guarantees, existing methods must assume that the real system is in the class of models considered during learning. We present an iterative method with strong guarantees even in the agnostic case where the system is not in the class. In particular, we show that any no-regret online learning algorithm can be used to obtain a near-optimal policy, provided some model achieves low training error and access to a good exploration distribution. Our approach applies to both discrete and continuous domains. We demonstrate its efficacy and scalability on a challenging helicopter domain from the literature.

Greedy Algorithms for Sparse Reinforcement Learning

Christopher Painter-Wakefield, Ronald Parr

Feature selection and regularization are becoming increasingly prominent tools in the efforts of the reinforcement learning (RL) community to expand the reach and applicability of RL. One approach to the problem of feature selection is to impose a sparsity-inducing form of regularization on the learning method. Recent work on L_1 regularization has adapted techniques from the supervised learning literature for use with RL. Another approach that has received renewed attention in the supervised learning community is that of using a simple algorithm that greedily adds new features. Such algorithms have many of the good properties of the L_1 regularization methods, while also being extremely efficient and, in some cases, allowing theoretical

guarantees on recovery of the true form of a sparse target function from sampled data. This paper considers variants of orthogonal matching pursuit (OMP) applied to reinforcement learning. The resulting algorithms are analyzed and compared experimentally with existing L_1 regularized approaches. We demonstrate that perhaps the most natural scenario in which one might hope to achieve sparse recovery fails; however, one variant, OMP-BRM, provides promising theoretical guarantees under certain assumptions on the feature dictionary. Another variant, OMP-TD, empirically outperforms prior methods both in approximation accuracy and efficiency on several benchmark problems.

On the Sample Complexity of Reinforcement Learning with a Generative Model

Mohammad Gheshlaghi Azar, Remi Munos, Bert Kappen

We consider the problem of learning the optimal action-value function in the discounted-reward Markov decision processes (MDPs). We prove a new PAC bound on the sample-complexity of model-based value iteration algorithm in the presence of the generative model, which indicates that for an MDP with N state-action pairs and the discount factor $\gamma \in [0, 1)$ only $O\Big(N \log(N/\delta)/\Big((1-\gamma)^3\epsilon^2\Big)\Big)$ samples are required to find an ϵ-optimal estimation of the action-value function with the probability $1-\delta$. We also prove a matching lower bound of $\Theta\Big(N \log(N/\delta)/\Big((1-\gamma)^3\epsilon^2\Big)\Big)$ on the sample complexity of estimating the optimal action-value function by every RL algorithm. To the best of our knowledge, this is the first matching result on the sample complexity of estimating the optimal (action-)value function in which the upper bound matches the lower bound of RL in terms of N, ϵ, δ and $1-\gamma$. Also, both our lower bound and our upper bound significantly improve on the state-of-the-art in terms of $1/(1-\gamma)$.

Artist Agent: A Reinforcement Learning Approach to Automatic Stroke Generation in Oriental Ink Painting

Ning Xie, Hirotaka Hachiya, Masashi Sugiyama

Oriental ink painting, called Sumi-e, is one of the most appealing painting styles that has attracted artists around the world. Major challenges in computer-based Sumi-e simulation are to abstract complex scene information and draw smooth and natural brush strokes. To automatically find such strokes, we propose to model the brush as a reinforcement learning agent, and learn desired brush-trajectories by maximizing the sum of rewards in the policy search framework. We also provide elaborate design of actions, states, and rewards tailored for a Sumi-e agent. The effectiveness of our proposed approach is demonstrated through simulated Sumi-e experiments.

Path Integral Policy Improvement with Covariance Matrix Adaptation

Freek Stulp, Olivier Sigaud

There has been a recent focus in reinforcement learning (RL) on addressing continuous state and action problems by optimizing parameterized policies. PI2 is a recent example of this approach. It combines a derivation from first principles of stochastic optimal control with tools from statistical estimation theory. In this paper, we consider PI2 as a member of the wider family of methods which share the concept of probability-weighted averaging to iteratively update parameters to optimize a cost function. We compare PI2 to other members of the same family – Cross-Entropy Methods and CMA-ES – at the conceptual level and in terms of performance. The comparison suggests the derivation of a novel algorithm which we call PI2-CMA for "Path Integral Policy Improvement with Covariance Matrix Adaptation". PI2-CMA's main advantage is that it determines the magnitude of the exploration noise automatically.

7C: Clustering 2, 10:30–12:00, AT LT 1

On causal and anticausal learning

Bernhard Schoelkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, Joris Mooij

We consider the problem of function estimation in the case where an underlying causal model can be identified. This has implications for popular scenarios such as covariate shift, concept drift, transfer learning and semi-supervised learning. We argue that causal knowledge may facilitate some approaches for a given problem, and rule out others. In particular, we formulate a hypothesis for when semi-supervised learning can help, and corroborate it with empirical results.

Revisiting k-means: New Algorithms via Bayesian Nonparametrics

Brian Kulis, Michael Jordan

Bayesian models offer great flexibility for clustering applications—Bayesian nonparametrics can be used for modeling infinite mixtures, and hierarchical Bayesian models can be utilized for shared clusters across multiple data sets. For the most part, such flexibility is lacking in classical clustering methods such as k-means. In this paper, we revisit the k-means clustering algorithm from a Bayesian nonparametric viewpoint. Inspired by the asymptotic connection between k-means and mixtures of Gaussians, we show that a Gibbs sampling algorithm for the Dirichlet process mixture approaches a hard clustering algorithm in the limit, and further that the resulting algorithm monotonically minimizes an elegant underlying k-means-like clustering objective that includes a penalty for the number of clusters. We generalize this analysis to the case of clustering multiple data sets through a similar asymptotic argument with the hierarchical Dirichlet process. We also discuss further extensions that highlight the benefits of our analysis: i) a spectral relaxation involving thresholded eigenvectors, and ii) a normalized cut graph clustering algorithm that does not

fix the number of clusters in the graph.

Approximate Principal Direction Trees

Mark McCartin-Lim, Andrew McGregor, Rui Wang

We introduce a new spatial data structure for high dimensional data called the *approximate principal direction tree* (APD tree) that adapts to the intrinsic dimension of the data. Our algorithm ensures vector-quantization accuracy similar to that of computationally-expensive PCA trees with similar time-complexity to that of lower-accuracy RP trees. APD trees use a small number of power-method iterations to find splitting planes for recursively partitioning the data. As such they provide a natural trade-off between the running-time and accuracy achieved by RP and PCA trees. Our theoretical results establish a) strong performance guarantees regardless of the convergence rate of the power-method and b) that $O(\log d)$ iterations suffice to establish the guarantee of PCA trees when the intrinsic dimension is d. We demonstrate this trade-off and the efficacy of our data structure on both the CPU and GPU.

Clustering using Max-norm Constrained Optimization

Ali Jalali, Nathan Srebro

We suggest using the max-norm as a convex surrogate constraint for clustering. We show how this yields a better exact cluster recovery guarantee than previously suggested nuclear-norm relaxation, and study the effectiveness of our method, and other related convex relaxations, compared to other approaches.

Efficient Active Algorithms for Hierarchical Clustering

Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, Aarti Singh

Advances in sensing technologies and the growth of the internet have resulted in an explosion in the size of modern datasets, while storage and processing power continue to lag behind. This motivates the need for algorithms that are efficient, both in terms of the number of measurements needed and running time. To combat the challenges associated with large datasets, we propose a general framework for active hierarchical clustering that repeatedly runs an off-the-shelf clustering algorithm on small subsets of the data and comes with guarantees on performance, measurement complexity and runtime complexity. We instantiate this framework with a simple spectral clustering algorithm and provide concrete results on its performance, showing that, under some assumptions, this algorithm recovers all clusters of size Omega(log n) using O(n log^2 n) similarities and runs in O(n log^3 n) time for a dataset of n objects. Through extensive experimentation we also demonstrate that this framework is practically alluring.

Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients

Iftekhar Naim, Daniel Gildea

The speed of convergence of the Expectation Maximization (EM) algorithm for Gaussian mixture model fitting is known to be dependent on the amount of overlap among the mixture components. In this paper, we study the impact of mixing coefficients on the convergence of EM. We show that when the mixture components exhibit some overlap, the convergence of EM becomes slower as the dynamic range among the mixing coefficients increases. We propose a deterministic anti-annealing algorithm, that significantly improves the speed of convergence of EM for such mixtures with unbalanced mixing coefficients. The proposed algorithm is compared against other standard optimization techniques like LBFGS, Conjugate Gradient, and the traditional EM algorithm. Finally, we propose a similar deterministic anti-annealing based algorithm for the Dirichlet process mixture model fitting and demonstrate its advantages over the conventional variational Bayesian approach.

Groupwise Constrained Reconstruction for Subspace Clustering

Ruijiang Li, Bin Li, Cheng Jin, Xiangyang Xue

Recent proposed subspace clustering methods first compute a self reconstruction matrix for dataset, then converted it to an affinity matrix, before input to a spectral clustering method to obtain the final clustering result. Their success is largely based on the subspace independence assumption, which, however, does not always hold for the applications with increasing number of clusters such as face clustering. In this paper, we proposes a novel reconstruction based subspace clustering method without making the subspace independence assumption. In our model, certain properties of the reconstruction matrix are explicitly characterized using the latent cluster indicators, and the affinity matrix input to the spectral clustering is built from the posterior of the cluster indicators. Evaluation on both synthetic and real-world datasets show that our method can outperform the state-of-the-arts.

Clustering by Low-Rank Doubly Stochastic Matrix Decomposition

Zhirong Yang, Erkki Oja

Clustering analysis by nonnegative low-rank approximations has achieved remarkable progress in the past decade. However, most approximation approaches in this direction are still restricted to matrix factorization. We propose a new low-rank learning method to improve the clustering performance, which is beyond matrix factorization. The approximation is based on a two-step bipartite random walk through virtual cluster nodes, where the approximation is formed by only cluster assigning probabilities. Minimizing the approximation error measured by Kullback-Leibler divergence is equivalent to maximizing the likelihood of a discriminative model, which endows our method with a solid probabilistic interpretation. The optimization is implemented by a relaxed Majorization-Minimization algorithm that is advantageous in finding good local minima. Furthermore, we point out that the regularized algorithm with

Dirichlet prior only serves as initialization. Experimental results show that the new method has strong performance in clustering purity for various datasets, especially for large-scale manifold data.

7D: Supervised learning 2, 10:30–12:00, AT LT 2

Total Variation and Euler's Elastica for Supervised Learning

Tong Lin, Hanlin Xue, Ling Wang, Hongbin Zha

In recent years, total variation (TV) and Euler's elastica (EE) have been successfully applied to image processing tasks such as denoising and inpainting. This paper investigates how to extend TV and EE to the supervised learning settings on high dimensional data. The supervised learning problem can be formulated as an energy functional minimization under Tikhonov regularization scheme, where the energy is composed of a squared loss and a total variation smoothing (or Euler's elastica smoothing). Its solution via variational principles leads to an Euler-Lagrange PDE. However, the PDE is always high-dimensional and cannot be directly solved by common methods. Instead, radial basis functions are utilized to approximate the target function, reducing the problem to finding the linear coefficients of basis functions. We apply the proposed methods to supervised learning tasks (including binary classification, multi-class classification, and regression) on benchmark data sets. Extensive experiments have demonstrated promising results of the proposed methods.

A General Framework for Inferring Latent Task Structures

Alexandre Passos, Piyush Rai, Jacques Wainer, Hal Daume III

When learning multiple related tasks with limited training data per task, it is natural to share parameters across tasks to improve generalization performance. Imposing the correct notion of task relatedness is critical to achieve this goal. However, one rarely knows the correct relatedness notion a priori. We present a generative model that is based on the weight vector of each task being drawn from a nonparametric mixture of nonparametric factor analyzers. The nonparametric aspect of the model makes it broad enough to subsume many types of latent task structures previously considered in the literature as special cases, and also allows it to learn more general task structures, addressing their individual shortcomings. This brings in considerable flexibility as compared to the commonly used multitask learning models that are based on some a priori fixed notion of task relatedness. We also present an efficient variational inference algorithm for our model. Experimental results on synthetic and real-world datasets, on both regression and classification problems, establish the efficacy of the proposed method.

Fast classification using sparse decision DAGs

Robert Busa-Fekete, Djalel Benbouzid, Balazs Kegl

In this paper we propose an algorithm that builds sparse decision DAGs (directed acyclic graphs) out of a list of base classifiers provided by an external learning method such as AdaBoost. The basic idea is to cast the DAG design task as a Markov decision process. Each instance can decide to use or to skip each base classifier, based on the current state of the classifier being built. The result is a sparse decision DAG where the base classifiers are selected in a data-dependent way. The method has a single hyperparameter with a clear semantics of controlling the accuracy/speed trade-off. The algorithm is competitive with state-of-the-art cascade detectors on three object-detection benchmarks, and it clearly outperforms them in the regime of low number of base classifiers. Unlike cascades, it is also readily applicable for multi-class classification. Using the multi-class setup, we show on a benchmark web page ranking data set that we can significantly improve the decision speed without harming the performance of the ranker.

An Efficient Approach to Sparse Linear Discriminant Analysis

Luis Francisco Sánchez Merchante, Yves Grandvalet, Gérrad Govaert

We present the Group-Lasso Optimal Scoring Solver (GLOSS), a novel approach to the formulation and the resolution of sparse Linear Discriminant Analysis (LDA). Our formulation, which is based on penalized Optimal Scoring, preserves an exact equivalence with sparse LDA, contrary to the multi-class approaches based on the regression of class indicator that have been proposed so far. Additionally, the versatility of the implementation allows to impose some particular structure on the within-class covariance matrix. Computationally, the optimization algorithm considers two nested convex problems, the main one being a linear regression regularized by a quadratic penalty implementing the group-lasso penalty. As group-Lasso selects the same features in all discriminant directions, it generates extremely parsimonious models without compromising the prediction performances. Moreover, the resulting sparse discriminant directions are amenable to low-dimensional representations. Our algorithm is highly efficient for medium to large number of variables, and is thus particularly well suited to the analysis of gene expression data.

Sequential Nonparametric Regression

Haijie Gu, John Lafferty

We present algorithms for nonparametric regression in settings where the data are obtained sequentially. While traditional estimators select bandwidths that depend upon the sample size, for sequential data the effective sample size is dynamically changing. We propose a linear time algorithm that adjusts the bandwidth for each new data point, and show that the estimator achieves the optimal minimax rate of convergence. We also propose the use of online expert mixing algorithms to adapt to unknown smoothness of the regression function. We provide simulations that confirm

the theoretical results, and demonstrate the effectiveness of the methods.

The Landmark Selection Method for Multiple Output Prediction

Krishnakumar Balasubramanian, Guy Lebanon

Conditional modeling $x \to y$ is a central problem in machine learning. A substantial research effort is devoted to such modeling when x is high dimensional. We consider, instead, the case of a high dimensional y, where x is either low dimensional or high dimensional. Our approach is based on selecting a small subset y_L of the dimensions of y, and proceed by modeling (i) $x \to y_L$ and (ii) $y_L \to y$. Composing these two models, we obtain a conditional model $x \to y$ that possesses convenient statistical properties. Classification and regression experiments on multiple datasets show that this model outperforms the one vs. all approach as well as several sophisticated multiple output prediction methods.

Ensemble Methods for Convex Regression with Applications to Geometric Programming Based Circuit Design

Lauren Hannah, David Dunson

Convex regression is a promising area for bridging statistical estimation and deterministic convex optimization. We develop a new piecewise linear convex regression method that uses the Convex Adaptive Partitioning (CAP) estimator in an ensemble setting, Ensemble Convex Adaptive Partitioning (E-CAP). The ensembles alleviate some problems associated with convex piecewise linear estimators, such as instability when used to approximate constraints or objective functions for optimization, while maintaining desirable properties, such as consistency and $O(nlog(n)^2)$ computational complexity. We empirically demonstrate that E-CAP outperforms existing convex regression methods both when used for prediction and optimization. We then apply E-CAP to device modeling and constraint approximation for geometric programming based circuit design.

AOSO-LogitBoost: Adaptive One-Vs-One LogitBoost for Multi-Class Problem

Peng Sun, Mark Reid, Jie Zhou

This paper is dedicated to the improvement of model learning in multi-class LogitBoost for classification. Motivated by statistical view, LogitBoost can be seen as additive tree regression. Important facts in such a setting are 1) coupled classifier output as sum-to-zero constraint and 2) dense Hessian matrix arising in tree node split gain and node values fitting. On the one hand, the setting is too complicated for a tractable model learning algorithm; On the other hand, too aggressive simplification of the setting may lead to degraded performance. For example, the original LogitBoost is outperformed by ABC-LogitBoost due to the later's more careful treatment for the above two key points in problem settings. In this paper we propose improved

methods to address the challenge: we adopt 1) vector tree (i.e. node value is vector) that enforces sum-to-zero constraint and 2) adaptive block coordinate descent exploiting dense Hessian when computing tree split gain and node values. Higher classification accuracy and faster convergence rate are observed for a range of public data sets when comparing to both original and ABC LogitBoost.

7E: Probabilistic Models, 10:30–12:00, AT LT 3

Local Loss Optimization in Operator Models: A New Insight into Spectral Learning

Borja Balle, Ariadna Quattoni, Xavier Carreras

This paper re-visits the spectral method for learning latent variable models defined in terms of observable operators. We give a new perspective on the method, showing that operators can be recovered by minimizing a loss defined on a finite subspace of the domain. A non-convex optimization similar to the spectral method is derived. We also propose a regularized convex relaxation of this optimization. We show that in practice the availabilty of a continuous regularization parameter (in contrast with the discrete number of states in the original method) allows a better trade-off between accuracy and model complexity. We also prove that in general, a randomized strategy for choosing the local loss will succeed with high probability. We also prove that in general, a randomized strategy for choosing the local loss will succeed with high probability.

Discriminative Probabilistic Prototype Learning

Edwin Bonilla, Antonio Robles-Kelly

In this paper we propose a simple yet powerful method for learning representations in supervised learning scenarios where each original input datapoint is described by a set of vectors and their associated outputs may be given by soft labels indicating, for example, class probabilities. We represent an input datapoint as a mixture of probabilities over the corresponding set of feature vectors where each probability indicates how likely each vector is to belong to an unknown prototype pattern. We propose a probabilistic model that parameterizes these prototype patterns in terms of hidden variables and therefore it can be trained with conventional approaches based on likelihood maximization. More importantly, both the model parameters and the prototype patterns can be learned from data in a discriminative way. We show that our model can be seen as a probabilistic generalization of learning vector quantization (LVQ). We apply our method to the problems of shape classification, hyperspectral imaging classification and people's work class categorization, showing the superior performance of our method compared to the standard prototype-based classification approach and other competitive benchmark methods.

Isoelastic Agents and Wealth Updates in Machine Learning Markets

Amos Storkey, Jono Millin, Krzysztof Geras

Recently, prediction markets have shown considerable promise for developing flexible mechanisms for machine learning. In this paper agents with isoelastic utilities are considered, and it is shown that the costs associated with homogeneous markets of agents with isoelastic utilities produce equilibrium prices corresponding to alpha-mixtures, with a particular form of mixing component relating to each agent's wealth. We also demonstrate that wealth accumulation for logarithmic and other isoelastic agents (through payoffs on prediction of training targets) can implement both Bayesian model updates and mixture weight updates by imposing different market payoff structures. An efficient variational algorithm is given for market equilibrium computation. We demonstrate that inhomogeneous markets of agents with isoelastic utilities outperform state of the art aggregate classifiers such as random forests, as well as single classifiers (neural networks, decision trees) on a number of machine learning benchmarks, and show that isoelastic combination methods are generally better than using logarithmic agents.

Evaluating Bayesian and L1 Approaches for Sparse Unsupervised Learning

Shakir Mohamed, Katherine Heller, Zoubin Ghahramani

The use of L_1 regularisation for sparse learning has generated immense research interest, with many successful applications in diverse areas such as signal acquisition, image coding, genomics and collaborative filtering. While existing work highlights the many advantages of L_1 methods, in this paper we find that L_1 regularisation often dramatically under-performs in terms of predictive performance when compared with other methods for inferring sparsity. We focus on unsupervised latent variable models, and develop L_1 minimising factor models, Bayesian variants of "L_1", and Bayesian models with a stronger L_0-like sparsity induced through spike-and-slab distributions. These spike-and-slab Bayesian factor models encourage sparsity while accounting for uncertainty in a principled manner, and avoid unnecessary shrinkage of non-zero values. We demonstrate on a number of data sets that in practice spike-and-slab Bayesian methods outperform L_1 minimisation, even on a computational budget. We thus highlight the need to re-assess the wide use of L_1 methods in sparsity-reliant applications, particularly when we care about generalising to previously unseen data, and provide an alternative that, over many varying conditions, provides improved generalisation performance.

Nonparametric variational inference

Samuel Gershman, Matt Hoffman, David Blei

Variational methods are widely used for approximate posterior inference. However, their use is typically limited to families of distributions that enjoy particular conju-

gacy properties. To circumvent this limitation, we propose a family of variational approximations inspired by nonparametric kernel density estimation. The locations of these kernels and their bandwidth are treated as variational parameters and optimized to improve an approximate lower bound on the marginal likelihood of the data. Using multiple kernels allows the approximation to capture multiple modes of the posterior, unlike most other variational approximations. We demonstrate the efficacy of the nonparametric approximation with a hierarchical logistic regression model and a nonlinear matrix factorization model. We obtain predictive performance as good as or better than more specialized variational methods and sample-based approximations. The method is easy to apply to more general graphical models for which standard variational methods are difficult to derive.

Levy Measure Decompositions for the Beta and Gamma Processes

Yingjian Wang, Lawrence Carin

We develop new representations for the Levy measures of the beta and gamma processes. These representations are manifested in terms of an infinite sum of well-behaved (proper) beta and gamma distributions. Further, we demonstrate how these infinite sums may be truncated in practice, and explicitly characterize truncation errors. We also perform an analysis of the characteristics of posterior distributions, based on the proposed decompositions. The decompositions provide new insights into the beta and gamma processes, and we demonstrate how the proposed representation unifies some properties of the two. This paper is meant to provide a rigorous foundation for and new perspectives on Lévy processes, as these are of increasing importance in machine learning.

Copula Mixture Model for Dependency-seeking Clustering

Melanie Rey, Volker Roth

We introduce a copula mixture model to perform dependency-seeking clustering when co-occurring samples from different data sources are available. The model takes advantage of the great flexibility offered by the copulas framework to extend mixtures of Canonical Correlation Analysis to multivariate data with arbitrary continuous marginal densities. We formulate our model as a non-parametric Bayesian mixture, while providing efficient MCMC inference. Experiments on synthetic and real data demonstrate that the increased flexibility of the copula mixture significantly improves the clustering and the interpretability of the results.

Predicting accurate probabilities with a ranking loss

Aditya Menon, Xiaoqian Jiang, Shankar Vembu, Charles Elkan, Lucila Ohno-Machado

In many real-world applications of machine learning classifiers, it is essential to predict the probability of an example belonging to a particular class. This paper proposes a simple technique for predicting probabilities based on optimizing a ranking

loss, followed by isotonic regression. This semi-parametric technique offers both good ranking and regression performance, and models a richer set of probability distributions than statistical workhorses such as logistic regression. We provide experimental results that show the effectiveness of this technique on real-world applications of probability prediction.

8A: Kernel methods 2, 14:00–15:40, AT LT 4

Copula-based Kernel Dependency Measures

Barnabas Poczos, Zoubin Ghahramani, Jeff Schneider

The paper presents a new copula based method for measuring dependence between random variables. Our approach extends the Maximum Mean Discrepancy to the copula of the joint distribution. We prove that this approach has several advantageous properties. Similarly to Shannon mutual information, the proposed dependence measure is invariant to any strictly increasing transformation of the marginal variables. This is important in many applications, for example in feature selection. The estimator is consistent, robust to outliers, and uses rank statistics only. We derive upper bounds on the convergence rate and propose independence tests too. We illustrate the theoretical contributions through a series of experiments in feature selection and low-dimensional embedding of distributions using real and toy datasets.

The Kernelized Stochastic Batch Perceptron

Andrew Cotter, Shai Shalev-Schwartz, Nathan Srebro

We present a novel approach for training kernel Support Vector Machines, establish learning runtime guarantees for our method that are better then those of any other known kernelized SVM optimization approach, and show that our method works well in practice compared to existing alternatives.

Fast Bounded Online Gradient Descent Algorithms for Scalable Kernel-Based Online Learning

Steven C.H. Hoi, Jialei Wang, Peilin Zhao, Rong Jin, Pengcheng Wu

Although kernel-based online learning has shown promising performance in a number of studies, its main shortcoming is the unbounded number of support vectors, making it non-scalable and unsuitable for large-scale datasets. In this work, we study the problem of bounded kernel-based online learning that aims to constrain the number of support vectors by a predefined budget. Although several algorithms have been proposed, they are mostly based on the Perceptron algorithm and only provide bounds for the number of classification mistakes. To address this limitation, we propose a framework for bounded kernel-based online learning based on online gradient descent approach. We propose two efficient algorithms of bounded online gradient descent (BOGD) for bounded kernel-based online learning: (i) BOGD using

uniform sampling and (ii) BOGD++ using non-uniform sampling. We present theoretical analysis of regret bound for both algorithms, show the promising empirical performance of the proposed algorithms by comparing them to the state-of-the-art algorithms for bounded kernel-based online learning.

Distributed Tree Kernels

Fabio Massimo Zanzotto, Lorenzo Dell'Arciprete

In this paper, we propose the distributed tree kernels (DTK) as a novel method to reduce time and space complexity of tree kernels. Using a linear complexity algorithm to compute vectors for trees, we embed feature spaces of tree fragments in low-dimensional spaces where the kernel computation is directly done with dot product. We show that DTKs are faster, correlate with tree kernels, and obtain a statistically similar performance in two natural language processing tasks.

Analysis of Kernel Mean Matching under Covariate Shift

Yaoliang Yu, Csaba Szepesvari

In real supervised learning scenarios, it is not uncommon that the training and test sample follow different probability distributions, thus rendering the necessity to correct the sampling bias. Focusing on a particular covariate shift problem, we derive high probability confidence bounds for the kernel mean matching (KMM) estimator, whose convergence rate turns out to depend on some regularity measure of the regression function and also on some capacity measure of the kernel. By comparing KMM with the natural plug-in estimator, we establish the superiority of the former hence provide concrete evidence/understanding to the effectiveness of KMM under covariate shift.

8B: Active and cost-sensitive learning, 14:00–15:35, AT LT 5

The Greedy Miser: Learning under Test-time Budgets

Zhixiang Xu, Kilian Weinberger, Olivier Chapelle

As machine learning algorithms enter applications in industrial settings, there is increased interest in controlling their cpu-time during testing. The cpu-time consists of the running time of the algorithm and the extraction time of the features. The latter can vary drastically when the feature set is diverse. In this paper, we propose an algorithm, the Greedy Miser, that incorporates the feature extraction cost during training to explicitly minimize the cpu-time during testing. The algorithm is a straightforward extension of stage-wise regression and is equally suitable for regression or multi-class classification. Compared to prior work, it is significantly more cost-effective and scales to larger data sets.

Joint Optimization and Variable Selection of High-dimensional Gaussian

Processes

Bo Chen, Rui Castro, Andreas Krause

Maximizing high-dimensional, non-convex functions through noisy observations is a notoriously hard problem, but one that arises in many applications. In this paper, we tackle this challenge by modeling the unknown function as a sample from a high-dimensional Gaussian process (GP) distribution. Assuming that the unknown function only depends on few relevant variables, we show that it is possible to perform joint variable selection and GP optimization. We provide strong performance guarantees for our algorithm, bounding the sample complexity of variable selection, and as well as providing cumulative regret bounds. We further provide empirical evidence on the effectiveness of our algorithm on several benchmark optimization problems.

Comparison-Based Learning with Rank Nets

Amin Karbasi, Stratis Ioannidis, laurent Massoulie

We consider the problem of search through comparisons, where a user is presented with two candidate objects and reveals which is closer to her intended target. We study adaptive strategies for finding the target, that require knowledge of rank relationships but not actual distances between objects. We propose a new strategy based on rank nets, and show that for target distributions with a bounded *doubling constant*, it finds the target in a number of comparisons close to the entropy of the target distribution and, hence, of the optimum. We extend these results to the case of noisy oracles, and compare this strategy to prior art over multiple datasets.

Bayesian Optimal Active Search and Surveying

Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, Richard Mann

We consider two active binary-classification problems with atypical objectives. In the first, active search, our goal is to actively uncover as many members of a given class as possible. In the second, active surveying, our goal is to actively query points to ultimately predict the class proportion of a given class. Numerous real-world problems can be framed in these terms, and in either case typical model-based concerns such as generalization error are only of secondary importance. We approach these problems via Bayesian decision theory; after choosing natural utility functions, we derive the optimal policies. We provide three contributions. In addition to introducing the active surveying problem, we extend previous work on active search in two ways. First, we prove a novel theoretical result, that less-myopic approximations to the optimal policy can outperform more-myopic approximations by any arbitrary degree. We then derive bounds that for certain models allow us to reduce (in practice dramatically) the exponential search space required by a naïve implementation of the optimal policy, enabling further lookahead while still ensuring the optimal decisions

are always made.

Hybrid Batch Bayesian Optimization

Javad Azimi, Ali Jalali, Xiaoli Zhang-Fern

Bayesian Optimization aims at optimizing an unknown non-convex/concave function that is costly to evaluate. We are interested in application scenarios where concurrent function evaluations are possible. Under such a setting, BO could choose to either sequentially evaluate the function, one input at a time and wait for the output of the function before making the next selection, or evaluate the function at a batch of multiple inputs at once. These two different settings are commonly referred to as the *sequential* and *batch* settings of Bayesian Optimization. In general, the sequential setting leads to better optimization performance as each function evaluation is selected with more information, whereas the batch setting has an advantage in terms of the total experimental time (the number of iterations). In this work, our goal is to combine the strength of both settings. Specifically, we systematically analyze Bayesian optimization using Gaussian process as the posterior estimator and provide a hybrid algorithm that, based on the current state, dynamically switches between a sequential policy and a batch policy with *variable* batch sizes. We provide theoretical justification for our algorithm and present experimental results on eight benchmark BO problems. The results show that our method achieves substantial speedup (up to %78) compared to a pure sequential policy, without suffering any significant performance loss.

Batch Active Learning via Coordinated Matching

Javad Azimi, Alan Fern, Xiaoli Zhang-Fern, Glencora Borradaile , Brent Heeringa

Most prior work on active learning of classifiers has focused on sequentially selecting one unlabeled example at a time to be labeled in order to reduce the overall labeling effort. In many scenarios, however, it is desirable to label an entire batch of examples at once, for example, when labels can be acquired in parallel. This motivates us to study batch active learning, which iteratively selects batches of $k > 1$ examples to be labeled. We propose a novel batch active learning method that leverages the availability of high-quality and efficient sequential active-learning policies by attempting to approximate their behavior when applied for k steps. Specifically, our algorithm first uses Monte-Carlo simulation to estimate the distribution of unlabeled examples selected by a sequential policy over k step executions. The algorithm then attempts to select a set of k examples that best matches this distribution, leading to a combinatorial optimization problem that we term "bounded coordinated matching'. While we show this problem is NP-hard in general, we give an efficient greedy solution, which inherits approximation bounds from supermodular minimization theory. Our experimental results on eight benchmark datasets show that the proposed approach

is highly effective

Bayesian Nonexhaustive Learning for Online Discovery and Modeling of Emerging Classes

Murat Dundar, Ferit Akova, Alan Qi, Bartek Rajwa

In this study we present a framework for online inference in the presence of a nonexhaustively defined set of classes that incorporates supervised classification with class discovery and modeling. A Dirichlet process prior (DPP) model defined over class distributions ensures that both known and unknown class distributions originate according to a common base distribution. In an attempt to automatically discover potentially interesting class formations, the prior model is coupled with a suitably chosen data model, and sequential Monte Carlo sampling is used to perform online inference. Our approach takes into account the rapidly accumulating nature of samples representing emerging classes, which is most evident in the biodetection application considered in this study, where a new class of pathogen may suddenly appear, and the rapid increase in the number of samples originating from this class indicates the onset of an outbreak.

8C: Feature selection and dimensionality reduction, 14:00–15:40, AT LT 1

Robust PCA in High-dimension: A Deterministic Approach

Jiashi Feng, Huan Xu, Shuicheng Yan

We consider principal component analysis for contaminated data-set in the high dimensional regime, where the number of observations is comparable or more than the dimensionality of each observation. We propose a *deterministic* high-dimensional robust PCA algorithm which inherits all theoretical properties of its *randomized* counterpart, i.e., it is tractable, robust to contaminated points, easily kernelizable, asymptotic consistent and achieves maximal robustness - a breakdown point of 50%. More importantly, the proposed method exhibits significantly better computational efficiency, which makes it suitable for large-scale real applications.

Communications Inspired Linear Discriminant Analysis

Minhua Chen, William Carson, Miguel Rodrigues, Lawrence Carin, Robert Calderbank

We study the problem of supervised linear dimensionality reduction, taking an information-theoretic viewpoint. The linear projection matrix is designed by maximizing the mutual information between the projected signal and the class label (based on a Shannon entropy measure). By harnessing a recent theoretical result on the gradient of mutual information, the above optimization problem can be solved directly using gradient descent, without requiring simplification of the objective function. Theoretical analysis and empirical comparison are made between the proposed method and two closely related methods (Linear Discriminant Analysis and Informa-

tion Discriminant Analysis), and comparisons are also made with a method in which Renyi entropy is used to define the mutual information (in this case the gradient may be computed simply, under a special parameter setting). Relative to these alternative approaches, the proposed method achieves promising results on real datasets.

Fast Training of Nonlinear Embedding Algorithms

Max Vladymyrov, Miguel Carreira-Perpinan

Stochastic neighbor embedding and related nonlinear manifold learning algorithms achieve high-quality low-dimensional representations of similarity data, but are notoriously slow to train. We propose a generic formulation of embedding algorithms that includes SNE and other existing algorithms, and study their relation with spectral methods and graph Laplacians. This allows us to define several partial-Hessian optimization strategies, characterize their global and local convergence, and evaluate them empirically. We achieve up to two orders of magnitude speedup over existing training methods with a strategy that adds nearly no overhead to the gradient and yet is simple, scalable and applicable to several existing and future embedding algorithms.

Regularizers versus Losses for Nonlinear Dimensionality Reduction: A Factored View with New Convex Relaxations

James Neufeld, Yaoliang Yu, Xinhua Zhang, Ryan Kiros, Dale Schuurmans

We demonstrate that almost all non-parametric dimensionality reduction methods can be expressed by a simple procedure: regularized loss minimization plus singular value truncation. By distinguishing the role of the loss and regularizer in such a process, we recover a factored perspective that reveals some gaps in the current literature. Beyond identifying a useful new loss for manifold unfolding, a key contribution is to derive new convex regularizers that combine distance maximization with rank reduction. These regularizers can be applied to any loss.

Sparse Support Vector Infinite Push

Alain Rakotomamonjy

In this paper, we address the problem of embedded feature selection for ranking on top of the list problems. We pose this problem as a regularized empirical risk minimization with p-norm push loss function ($p = \infty$) and sparsity inducing regularizers. We leverage the issues related to this challenging optimization problem by considering an alternating direction method of multipliers algorithm which is built upon proximal operators of the loss function and the regularizer. Our main technical contribution is thus to provide a numerical scheme for computing the infinite push loss function proximal operator. Experimental results on toy, DNA microarray and BCI problems show how our novel algorithm compares favorably to competitors for ranking on top while using fewer variables in the scoring function.

Adaptive Canonical Correlation Analysis Based On Matrix Manifolds

Florian Yger, Maxime Berar, Gilles Gasso, Alain Rakotomamonjy

In this paper, we formulate the Canonical Correlation Analysis (CCA) problem on matrix manifolds. This framework provides a natural way for dealing with matrix constraints and tools for building efficient algorithms even in an adaptive setting. Finally, an adaptive CCA algorithm is proposed and applied to a change detection problem in EEG signals.

Fast approximation of matrix coherence and statistical leverage

Michael Mahoney, Petros Drineas, Malik Magdon-Ismail, David Woodruff

The statistical leverage scores of a matrix A are the squared row-norms of the matrix containing its (top) left singular vectors and the coherence is the largest leverage score. These quantities have been of interest in recently-popular problems such as matrix completion and Nystrom-based low-rank matrix approximation; in large-scale statistical data analysis applications more generally; and since they define the key structural nonuniformity that must be dealt with in developing fast randomized matrix algorithms. Our main result is a randomized algorithm that takes as input an arbitrary n × d matrix A, with n » d, and that returns as output relative-error approximations to all n of the statistical leverage scores. The proposed algorithm runs in O(nd log n) time, as opposed to the O(nd2) time required by the naïve algorithm that involves computing an orthogonal basis for the range of A. This resolves an open question from (Drineas et al., 2006b) and (Mohri & Talwalkar, 2011); and our result leads to immediate improvements in coreset-based L2-regression, the estimation of the coherence of a matrix, and several related low-rank matrix problems. Interestingly, to achieve our result we judiciously apply random projections on both sides of A.

Feature Selection via Probabilistic Outputs

Andrea Danyluk, Nicholas Arnosti

This paper investigates two feature-scoring criteria that make use of estimated class probabilities: one method proposed by [**?**] and a complementary approach proposed below. We develop a theoretical framework to analyze each criterion and show that both estimate the spread (across all values of a given feature) of the probability that an example belongs to the positive class. Based on our analysis, we predict when each scoring technique will be advantageous over the other and give empirical results validating those predictions.

8D: Recommendation and Matrix Factorization, 14:00–15:35, AT LT 2

A Combinatorial Algebraic Approach for the Identifiability of Low-Rank Ma-

trix Completion

Franz Király, Ryota Tomioka

In this paper, we review the problem of matrix completion and expose its intimate relations with algebraic geometry and combinatorial graph theory. We present the first necessary and sufficient combinatorial conditions for matrices of arbitrary rank to be identifiable from a set of matrix entries, yielding theoretical constraints and new algorithms for the problem of matrix completion. We conclude with algorithmically evaluating the tightness of the given conditions and algorithms for practically relevant matrix sizes, showing that the algebraic-combinatoric approach can lead to improvements over state-of-the-art matrix completion methods.

Gap Filling in the Plant Kingdom—Trait Prediction Using Hierarchical Probabilistic Matrix Factorization

Hanhuai Shan, Jens Kattge, Peter Reich, Arindam Banerjee, Franziska Schrodt, Markus Reichstein

Plant traits are a key to understand and predict the adaptation of ecosystems to environmental changes, which motivates the TRY project aiming at constructing a global database for plant traits and becoming a standard resource for the ecological community. Despite its unprecedented coverage, a large percentage of missing data substantially constrains joint trait analysis. Meanwhile, the trait data are characterized by the hierarchical phylogenetic structure of the plant kingdom. While factorization based matrix completion techniques have been widely used to address the missing data problem, traditional matrix factorization methods are unable to leverage the phylogenetic structure. We propose hierarchical probabilistic matrix factorization (HPMF), which effectively uses hierarchical phylogenetic information for trait prediction. We demonstrate HPMF's high accuracy, effectiveness of incorporating hierarchical structure and ability to capture trait correlation through experiments.

Stability of matrix factorization for collaborative filtering

Yu-Xiang Wang, Huan Xu

We study the stability vis a vis adversarial noise of matrix factorization algorithm for matrix completion. In particular, our results include: (I) we bound the gap between the solution matrix of the factorization method and the ground truth in terms of root mean square error; (II) we treat the matrix factorization as a subspace fitting problem and analyze the difference between the solution subspace and the ground truth; (III) we analyze the prediction error of individual users based on the subspace stability. We apply these results to the problem of collaborative filtering under manipulator attack, which leads to useful insights and guidelines for collaborative filtering system design.

Latent Collaborative Retrieval

Jason Weston, Chong Wang, Ron Weiss, Adam Berenzweig

Retrieval tasks typically require a ranking of items given a query. Collaborative filtering tasks, on the other hand, learn models comparing users with items. In this paper we study the joint problem of recommending items to a user with respect to a given query, which is a surprisingly common task. This setup differs from the standard collaborative filtering one in that we are given a query $\times$ user $\times$ item tensor for training instead of the more traditional user $\times$ item matrix. Compared to document retrieval we do have a query, but we may or may not have content features (we will consider both cases) and we can also take account of the user's profile. We introduce a factorized model for this new task that optimizes the top ranked items returned for the given query and user. We report empirical results where it outperforms several baselines.

A Bayesian Approach to Approximate Joint Diagonalization of Square Matrices

Mingjun Zhong

We present a fully Bayesian approach to simultaneously approximate diagonalization of several square matrices which are not necessary symmetric. A Gibbs sampler has been derived for simulating the common eigenvectors and the eigenvalues for these matrices. Several data are used to demonstrate the performance of the proposed Gibbs sampler and we then provide comparisons to several other joint diagonalization algorithms, which shows that the Gibbs sampler achieves the state-of-the-art performance. As a byproduct, the output of the Gibbs sampler could be used to estimate the log marginal likelihood by the Bayesian information criterion (BIC) which correctly located the number of common eigenvectors. We then applied the Gibbs sampler to the blind sources separation problem, the common principal component analysis and the common spatial pattern analysis.

Collaborative Topic Regression with Social Matrix Factorization for Recommendation Systems

Sanjay Purushotham, Yan Liu

Social network websites, such as Facebook, YouTube, Lastfm etc, have become a popular platform for users to connect with each other and share content or opinions. They provide rich information to study the influence of user's social circle in their decision process. In this paper, we are interested in examining the effectiveness of social network information to predict the user's ratings of items. We propose a novel hierarchical Bayesian model which jointly incorporates topic modeling and probabilistic matrix factorization of social networks. A major advantage of our model is to automatically infer useful latent topics and social information as well as their impor-

tance to collaborative filtering from the training data. Empirical experiments on two large-scale datasets show that our algorithm provides a more effective recommendation system than the state-of-the art approaches. Our results also reveal interesting insight that the social circles have more influence on people's decisions about the usefulness of information (e.g., bookmarking preference on Delicious) than personal taste (e.g., music preference on Lastfm).

Active Learning for Matching Problems

Laurent Charlin, Rich Zemel, Craig Boutilier

Effective learning of user preferences is critical to easing user burden in various types of matching problems. Equally important is active query selection to further reduce the amount of preference information users must provide. We address the problem of active learning of user preferences for matching problems, introducing a novel method for determining probabilistic matchings, and developing several new active learning strategies that are sensitive to the specific matching objective. Experiments with real-world data sets spanning diverse domains demonstrate that matching-sensitive active learning outperforms standard techniques.

A Graphical Model Formulation of Collaborative Filtering Neighbourhood Methods with Fast Maximum Entropy Training

Aaron Defazio, Tiberio Caetano

Item neighbourhood methods for collaborative filtering learn a weighted graph over the set of items, where each item is connected to those it is most similar to. The prediction of a user's rating on an item is then given by that rating of neighbouring items, weighted by their similarity. This paper presents a new neighbourhood approach which we call item fields, whereby an undirected graphical model is formed over the item graph. The resulting prediction rule is a simple generalization of the classical approaches, which takes into account non-local information in the graph, allowing its best results to be obtained when using drastically fewer edges than other neighbourhood approaches. A fast approximate maximum entropy training method based on the Bethe approximation is presented which utilizes a novel decomposition into tractable sub-problems. When using precomputed sufficient statistics on the Movielens dataset, our method outperforms maximum likelihood approaches by two orders of magnitude.

8E: Graphical models, 14:00–15:40, AT LT 3

Variational Bayesian Inference with Stochastic Search

John Paisley, David Blei, Michael Jordan

Mean-field variational inference is an approximate posterior inference method for Bayesian models. It approximates the full posterior distribution of a model's variables

with a factorized set of distributions by maximizing a lower bound on the marginal likelihood. This requires the ability to integrate the log joint likelihood of the model with respect to the factorized approximation. Often not all integrals are closed-form, which is traditionally handled using lower bound approximations. We present an algorithm based on stochastic optimization that allows for direct optimization of the variational lower bound in all models. This method uses control variates for variance reduction of the stochastic search gradient, in which existing lower bounds can play an important role. We demonstrate the approach on logistic regression and the hierarchical Dirichlet process.

Large Scale Variational Bayesian Inference for Structured Scale Mixture Models

Young Jun Ko, Matthias Seeger

Natural image statistics exhibit hierarchical dependencies across multiple scales. Representing such prior knowledge in non-factorial latent tree models can boost performance of image denoising, inpainting, deconvolution or reconstruction substantially, beyond standard factorial "sparse" methodology. We derive a large scale approximate Bayesian inference algorithm for linear models with non-factorial (latent tree-structured) scale mixture priors. Experimental results on a range of denoising and inpainting problems demonstrate substantially improved performance compared to MAP estimation or to inference with factorial priors.

A Generalized Loop Correction Method for Approximate Inference in Graphical Models

Siamak Ravanbakhsh, Chun-Nam Yu, Russell Greiner

Belief Propagation (BP) is one of the most popular methods for inference in probabilistic graphical models. BP is guaranteed to return the correct answer for tree structures, but can be incorrect or non-convergent for loopy graphical models. Recently, several new approximate inference algorithms based on cavity distribution have been proposed. These methods can account for the effect of loops by incorporating the dependency between BP messages. Alternatively, region-based approximations (that lead to methods such as Generalized Belief Propagation) improve upon BP by considering interactions within small clusters of variables, thus taking small loops within these clusters into account. This paper introduces an approach, Generalized Loop Correction (GLC), that benefits from both of these types of loop correction. We show how GLC relates to these two families of inference methods, then provide empirical evidence that GLC works effectively in general, and can be significantly more accurate than both correction schemes.

Distributed Parameter Estimation via Pseudo-likelihood

Qiang Liu, alexander ihler

Estimating statistical models within sensor networks requires distributed algorithms, in which both data and computation are distributed across the nodes of the network. We propose a general approach for distributed learning based on combining local estimators defined by pseudo-likelihood components, encompassing a number of combination methods, and provide both theoretical and experimental analysis. We show that simple linear combination or max-voting methods, when combined with second-order information, are statistically competitive with more advanced and costly joint optimization. Our algorithms have many attractive properties including low communication and computational cost and "any-time" behavior.

High-Dimensional Covariance Decomposition into Sparse Markov and Independence Domains

Majid Janzamin, Animashree Anandkumar

In this paper, we present a novel framework incorporating a combination of sparse models in different domains. We posit the observed data as generated from a linear combination of a sparse Gaussian Markov model (with a sparse precision matrix) and a sparse Gaussian independence model (with a sparse covariance matrix). We provide efficient methods for decomposition of the data into two domains, viz. Markov and independence domains. We characterize a set of sufficient conditions for identifiability and model consistency. Our decomposition method is based on a simple modification of the popular ℓ_1-penalized maximum-likelihood estimator (ℓ_1-MLE), and is easily implementable. We establish that our estimator is consistent in both the domains, i.e., it successfully recovers the supports of both Markov and independence models, when the number of samples n scales as $n = \Omega(d^2 \log p)$, where p is the number of variables and d is the maximum node degree in the Markov model. Our conditions for recovery are comparable to those of ℓ_1-MLE for consistent estimation of a sparse Markov model, and thus, we guarantee successful high-dimensional estimation of a richer class of models under comparable conditions.

Plenary Session, 17:10–17:40

Machine Learning that Matters

Kiri Wagstaff

Much of current machine learning (ML) research has lost its connection to problems of import to the larger world of science and society. From this perspective, there exist glaring limitations in the data sets we investigate, the metrics we employ for evaluation, and the degree to which results are communicated back to their originating domains. What changes are needed to how we conduct research to increase the impact that ML has? We present six Impact Challenges to explicitly focus the field's energy and attention, and we discuss existing obstacles that must be addressed. We aim to inspire ongoing discussion and focus on ML that matters.

9 COLT Sessions

COLT: Tuesday June 26, 2012

Active Learning

08:30–09:45, AT LT 3. Chair: Jacob Abernethy

08:30–08:55 *Consistency of nearest neighbor classification under selective sampling*
Sanjoy Dasgupta

08:55–09:20 *Active Learning Using Smooth Relative Regret Approximations with Applications*
Nir Ailon, Ron Begleiter, Esther Ezra

09:20–09:45 *Robust Interactive Learning*
Maria-Florina Balcan, Steve Hanneke

Hypothesis Testing and Estimation

10:05–11:30, AT LT 3. Chair: Nina Balcan

10:05–10:30 *Rare Probability Estimation under Regularly Varying Heavy Tails*
Mesrob Ohannessian, Munther Dahleh

10:30–10:55 *Competitive Classification and Closeness Testing*
Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, Ananda Theertha Suresh

10:55–11:20 *Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems*
Magalie Fromont, Béatrice Laurent, Matthieu Lerasle, Patricia Reynaud-Bouret

11:20–11:30 *Bounds on the Bayes Error Given Moments*

Invited COLT Tutorial: Arkadi Nemirovski

12:30–13:30, AT LT 3

Privacy

13:30–14:45, . Chair: Hans Simon

13:30–13:55 *Differentially Private Online Learning*
Prateek Jain, Pravesh Kothari, Abhradeep Thakurta

13:55–14:20 *Differentially Private Convex Optimization for Empirical Risk Minimization with Applications to High-dimensional Regression*
Daniel Kifer, Adam Smith, Abhradeep Thakurta

14:20–14:45 *Distributed Learning, Communication Complexity and Privacy*
Maria-Florina Balcan, Avrim Blum, Shai Fine, Yishay Mansour

Risk Functions

15:05–16:20, AT LT 3. Chair: Ohad Shamir

15:05–15:30 *A Characterization of Scoring Rules for Linear Properties*
Rafael Frongillo, Jacob Abernethy

15:30–15:55 *Divergences and Risks for Multiclass Experiments*
Dario Garcia Garcia, Robert C. Williamson

15:55–16:20 *A Conjugate Property between Loss Functions and Uncertainty Sets in Classification Problems*
Takafumi Kanamori, Akiko Takeda, Taiji Suzuki

Tight Bounds

16:40–17:30, AT LT 3. Chair: Karthik Sridharan

16:40–17:05 *New Bounds for Learning Intervals with Implications for Semi-Supervised Learning*
David Helmbold, Philip Long

17:05–17:30 *Tight Bounds on Proper Equivalence Query Learning of DNF*
Lisa Hellerstein, Devorah Kletenik, Linda Sellie, Rocco Servedio

COLT: Wednesday June 27, 2012

Manifolds and Clustering

10:30–12:10, AT LT 3. Chair: Peter Auer

10:30–10:50 *Distance Preserving Embeddings for General n-Dimensional Manifolds*
Nakul Verma

10:50–11:10 *A Method of Moments for Hidden Markov Models and Multi-view Mixture Models*
Animashree Anandkumar, Daniel Hsu, Sham Kakade

11:10–11:30 *A Correlation Clustering Approach to Link Classification in Signed Networks*
Nicolò Cesa-Bianchi, Claudio Gentile, Fabio Vitale, Giovanni Zappella

11:30–11:50 *Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model*
Kamalika Chaudhuri, Fan Chung, Alexander Tsiatas

11:50–12:10 *Toward understanding complex spaces: graph Laplacians on manifolds with singularities and boundaries*
Mikhail Belkin, Qichao Que, Yusu Wang, Xueyuan Zhou

Online Learning II

16:00–17:40, AT LT 3. Chair: Satyen Kale

16:00–16:20 *Near-Optimal Algorithms for Online Matrix Prediction*
Near-Optimal Algorithms for Online Matrix Prediction

16:20–16:40 *Analysis of Thompson Sampling for the multi-armed bandit problem*
Shipra Agrawal, Navin Goyal

16:40–17:00 *Autonomous Exploration For Navigating In MDPs*
Shiau Hong Lim, Peter Auer

17:00–17:20 *Towards Minimax Policies for Online Linear Optimization with Bandit Feedback*
Sébastien Bubeck, Nicolo Cesa-Bianchi, Sham M. Kakade

17:20–17:40 *The best of both worlds: stochastic and adversarial bandits*
Sébastien Bubeck, Aleksandrs Slivkins

10 ICML Workshops, June 30 - July 1

W01: European Workshop on Reinforcement Learning (EWRL)
June 30 - July 1. Venue: AT LT1.
Chairs: Marc Deisenroth, Csaba Szepesvari and Jan Peters.

W02: Machine Learning for Clinical Data Analysis
June 30 - July 1. Venue: AT LT3.
Chairs: Noemie Elhadad, Milos Hauskrecht, Faisal Farooq, Suchi Saria, Dale Schuurmans, Zeeshan Syed, Csaba Szepesvari, Allan Tucker, Ozlem Uzuner and Shipeng Yu.

W03: Inferning: Interactions between Inference and Learning
June 30. Venue: AT LT5.
Chairs: Michael Wick, David Weiss, Sameer Singh and Andrew McCallum.

W05: Machine Learning in Human Computation & Crowdsourcing
June 30. Venue: DHT FRN.
Chairs: Rajarshi Das and Maja Vukovic.

W06: Music and Machine Learning
June 30. Venue: DHT FRS.
Chairs: Rafael Ramirez, Darrell Conklin and José Manuel Iñesta.

W07: Object, functional and structured data: towards next generation kernel-based methods
June 30. Venue: AT LT2.
Chairs: Florence d'Alché-Buc, Hachem Kadri, Massimiliano Pontil and Alain Rakotomamonjy.

W08: Online Advertising (AdML)
June 30. Venue: IF G07a.
Chairs: Deepak Agarwal, Mikhail Bilenko, Olivier Chapelle, Brendan McMahan and D. Sculley.

W09: Sparsity, Dictionaries and Projections in ML and Signal Processing
June 30. Venue: AT LT4.
Chairs: Rodolphe Jenatton, Michael Davies, Rémi Gribonval and Guillaume Obozinski.

W10: Statistical Relational Learning (SRL)
June 30. Venue: IF G07.
Chairs: James Cussens, Kristian Kersting, Sriraam Natarajan and David Poole.

W11: Teaching ML
June 30. Venue: IF 4.31/33.
Chairs: Kurt Driessens and Elisa Fromont.

W12: Machine Learning in Genetics and Genomics (MLGG)
July 1. Venue: IF 4.31/33.
Chairs: Anna Goldenberg, Pierre Baldi, Sara Mostafavi and Michal Rosen-Zvi.

W13: Markets, Mechanisms and Multi-Agent Models
July 1. Venue: IF G07a.
Chairs: Amos Storkey, Jacob Abernethy and Jenn Wortman Vaughan.

W14: Mining and Learning with Graphs (MLG)
July 1. Venue: AT LT2.
Chairs: Jan Ramon and Hanghang Tong.

W15: New Challenges for Exploration and Exploitation
July 1. Venue: DHT FRN.
Chairs: Jérémie Mary, Aurélien Garivier, Lihong Li, Rémi Munos, Olivier Nicol, Ronald Ortner and Philippe Preux.

W16: Optimization in Machine Learning
July 1. Venue: AT LT4.
Chairs: Katya Scheinberg, Ben Recht and Sasha Rakhlin.

W17: Representation Learning
July 1. Venue: AT LT5.
Chairs: Aaron Courville, Hugo Larochelle, MarcÁurelio Ranzato and Yoshua Bengio.

W18: RKHS and Kernel-based methods
July 1. Venue: IF G07.

Chairs: Arthur Gretton, Zaid Harchaoui and Bharath Sriperumbudur.

W19: Statistics, ML and Neuroscience (STAMLINS)
July 1. Venue: DHT FRS.
Chairs: Iead Rezek, Evangelos Roussos and Christian Beckmann.

W01: European Workshop on Reinforcement Learning (EWRL)

June 30 - July 1. Venue: AT LT1.

Chairs: Marc Deisenroth, Csaba Szepesvari and Jan Peters

June 30

09:00 – 09:15 ***Welcome***

09:15 – 10:10 *Invited Talk: Known Unknowns*
Shie Mannor

10:10 – 10:30 ***Contributed Talks***

10:30 – 11:00 ***COFFEE***

11:00 – 12:30 ***Contributed Talks***

12:30 – 14:00 ***LUNCH***

14:00 – 14:50 *Invited Talk: Neural Architectures for Real World Reinforcement Learning*
Martin Riedmiller

14:50 – 15:30 ***Contributed Talks***

15:30 – 16:00 ***COFFEE***

16:00 – 17:30 ***Poster Session I***

18:30 ***Banquet***

July 1

08:30 – 09:00 ***COFFEE for arrival***

09:00 – 09:50 *Invited Talk: Machine Learning with Multiple Guesses: Contextual Control Libraries*
Drew Bagnell

09:50 – 10:30 ***Contributed Talks***

10:30 – 11:00 ***COFFEE***

11:00 – 12:30 ***Poster Session II***

12:30 – 14:00 ***LUNCH***

14:00 – 14:50 *Invited Talk: Verification in Artificial Intelligence*
Richard Sutton

14:50 – 15:30 ***Contributed Talks***

15:30 – 16:00 ***COFFEE***

16:00 – 17:30 ***Contributed Talks***

17:30 – 17:45 ***Closing Remarks***

W02: Machine Learning for Clinical Data Analysis

June 30 - July 1. Venue: AT LT3.

Chairs: Noemie Elhadad, Milos Hauskrecht, Faisal Farooq, Suchi Saria, Dale Schuurmans, Zeeshan Syed, Csaba Szepesvari, Allan Tucker, Ozlem Uzuner and Shipeng Yu

Clinical and health-care applications have been and continue to be the source of inspiration for many areas of artificial intelligence research. A challenge for machine learning in particular, arises from the wealth and variety of data generated in modern medical and health-care settings. Extensive electronic health records – with free text as well as thousands of fields recording patient conditions, diagnostic tests, treatments, outcomes, and so on – provide an unprecedented source of information that can provide clues leading to potential improvements in disease detection, chronic disease management, design of clinical trials, and other aspects of health-care. The purpose of this workshop is to bring together machine learning and informatics researchers interested in problems and applications in the clinical domain, with the goal of exchanging ideas and perspectives, identifying research bottlenecks and medical applications, bridging the gap between the theory of machine learning, natural language processing, and the needs of the healthcare community, and, in general, raising awareness of potential healthcare applications in the machine learning community.

June 30

9:00-10:00 *Keynote Talk 1*
John Holmes

10:00-10:25 *What We Found on Our Way to Building a Classifier: A Critical Analysis of the AHA Screening Questionnaire*
Quazi Rahman, Sivajothi Kanagalingam, Aurelio Pinheiro, Theodore Abraham, Hagit Shatkay

10:30-11:00 ***Coffee Break***

11:00-11:25 *Kernel-based Characterization of Dynamics in a Heterogeneous Population of Septic Patients Under Therapy*
Kosta Ristovski, Vladan Radosavljevic, Zoran Obradovic

11:25-11:50 *State Space Gaussian Process Prediction*
Zitao Liu, Lei Wu

11:50-12:15 *A Semi-Causal Bayesian Network Approach to Prognosis*
Arjen Hommersom, Peter Lucas, Anne van Altena, Leon Massuger, Lambertus Kiemeney

12:15-12:40 *Non-stationary Clustering Bayesian Networks for Glaucoma*
Stefano Ceccon, Allan Tucker, David Garway-Heath, David Crabb

12:40-14:00 ***Lunch***

14:00-15:00 *Keynote Talk 2*
George Hripcsak

15:00-15:25 *Uplift Modeling for Clinical Trial Data*
Maciej Jaskowski, Szymon Jaroszewicz

15:25-15:50 *Learning Evolving Patient Risk Processes for C. Diff Colonization*
Jenna Wiens, John Guttag, Eric Horvitz

15-50-16:00 ***Coffee Break and Poster Session***

16:00-17:30 ***Poster Session***

July 1

9:00-10:00 *Keynote Talk 3*
Gregory Cooper

10:00-10:25 *A Comparison of Dimensionality Reduction Techniques for Unstructured Clinical Text*
Yoni Halpern, Steven Horng, Larry Nathanson, Nathan Shapiro, David Sontag

10:25-10:50 *Feature Engineering and Selection for Rheumatoid Arthritis Disease Activity Classification Using Electronic Medical Records*
Chen Lin, Timothy Miller, Dmitry Dligach, Robert Plenge, Elizabeth Karlson, Guergana Savova

10:50-11:05 ***Coffee Break***

11:05-11:40 *Early Prediction of Coronary Artery Calcification Levels Using Statistical Relational Learning*
Sriraam Natarajan, Kristian Kersting, Saket Joshi, Santiago Saldana, Edward Ip, David Jacobs, Jeffery Carr

11:30-12:30 *Keynote Talk 4*
Shahram Ebadollahi

12:30-14:00 ***Lunch***

14:00-15:00 ***Panel on Challenges for Clinical Data Analysis***

W03: Inferning: Interactions between Inference and Learning

June 30. Venue: AT LT5.

Chairs: Michael Wick, David Weiss, Sameer Singh and Andrew McCallum

This workshop studies the interactions between algorithms that learn a model, and algorithms that use the resulting model parameters for inference. These interactions are studied from two perspectives.

The first perspective studies how the choice of an inference algorithm influences the parameters the model ultimately learns. For example, many parameter estimation algorithms require inference as a subroutine. Consequently, when we are faced with models for which exact inference is expensive, we must use an approximation instead: MCMC sampling, belief propagation, beam-search, etc. On some problems these approximations yield superior models, yet on others, they fail catastrophically. We invite studies that analyze (both empirically and theoretically) the impact of approximate inference on model learning. How does approximate inference alter the learning objective? Affect generalization? Influence convergence properties? Further, does the behavior of inference change as learning continues to improve the quality of the model?

A second perspective from which we study these interactions is by considering how the learning objective and model parameters can impact both the quality and performance of inference during "test time." These unconventional approaches to learning combine generalization to unseen data with other desiderata such as fast inference. For example, work in structured cascades learns model for which greedy, efficient inference can be performed at test time while still maintaining accuracy guarantees. Similarly, there has been work that learns operators for efficient search-based inference. There has also been work that incorporates resource constraints on running time and memory into the learning objective.

June 30

9:00-9:15 ***Opening Remarks***

9:15-10:00 *Keynote Talk: Learning Tractable but Expressive Models*
Pedro Domingos

10:00-10:15 *On the Mismatch Between Learning and Inference for Single Network Domains*
Rongjing Xiang and Jennifer Neville

10:15-10:30 *A Framework for Tuning Posterior Entropy in Unsupervised Learning*
Rajhans Samdani, Ming-Wei Chang and Dan Roth

10:30-11:15 ***Poster Session (+ Coffee)***

11:15-12:00	*Keynote Talk: Learning to Sample* Max Welling
12:00-12:15	*Cost-sensitive Dynamic Feature Selection* He He, Hal Daume III and Jason Eisner
12:15-14:00	***Lunch***
14:00-14:45	*Keynote Talk: Structured Prediction using Linear Programming Relaxations: Algorithms and Theory* David Sontag
14:45-15:00	*Speeding up MAP with Column Generation and Block Regularization* David Belanger, Alexandre Passos, Sebastian Riedel and Andrew McCallum
15:00-15:15	*Learning Search Based Inference for Object Detection* Peter Gehler and Alain Lehmann
15:15-15:30	*Fast and Accurate CRFs through Evidence-Specific Structures* Veselin Stoyanov and Jason Eisner
15:30-16:15	***Poster Session (+ Coffee)***
16:15-17:00	*Keynote Talk: Learning Approximate Inference Policies for Fast Prediction* Jason Eisner
17:00-17:15	*Approximating Marginals Using Discrete Energy Minimization* Filip Korc, Vladimir Kolmogorov and Christoph H. Lampert
17:15-18:00	*Keynote Talk: Marginalization-Based Parameter Learning in Graphical Models* Justin Domke
18:00-18:15	*Learned Prioritization for Trading Off Accuracy and Speed* Jiarong Jiang, Adam Teichert, Hal Daume III and Jason Eisner
18:15-18:30	***Closing Remarks***

W05: Machine Learning in Human Computation & Crowdsourcing

June 30. Venue: DHT FRN.

Chairs: Rajarshi Das and Maja Vukovic

The ubiquitous access to computing systems has spurred the design of a variety of human computation and crowdsourcing systems, each aiming to harness the power of human computation to tackle problems that cannot be solved by today's computers alone. In the past few years, a number of crowdsourcing efforts have been featured in the popular media including GalaxyZoo, DARPA Red Balloon Challenge, Amazon Mechanical Turk, to name a few. Such systems may involve a large multitude of independent human agents, each with their own level of interest, incentive and expertise for the tasks at hand. In spite of the rapid pace of experimental progress in designing and running human computation and crowdsourcing applications, there is a lack of underlying tools and methods to systematically address issues such as efficiency, robustness and scalability in these systems. Machine learning, with its focus on empirical data, offers the promise of addressing fundamental issues in crowdsourcing such as active learning, expertise inferencing, incentive engineering, response verification and quality control. Indeed, in the last couple of years, a number of papers highlighting novel applications of machine learning in the crowdsourcing domain have appeared in recent editions of leading machine learning conferences such as ICML, AAAI, NIPS and IJCAI, suggesting the emergence of a new area of research in the machine learning community.

The goal of this workshop is to provide a forum for cross-fertilization of ideas among machine learning and crowdsourcing practitioners that could lead to novel machine learning algorithms as well as more robust and efficient crowdsourcing systems. In particular, the workshop will aim to understand how to match the capabilities of new and existing machine learning techniques with practical requirements faced in designing human computation and crowdsourcing systems.

June 30

09:00-09:10 ***Introduction***

09:10-09:40 *A Brief Survey of Machine Learning in Human Computation and Crowdsourcing*
Rajarshi Das and Maja Vukovic

9:40-10:20 *How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing*
Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver

10:30-11:00 ***Coffee***

11:00-11:40	*Get another worker? Active crowdlearning with sequential arrivals* James Zou and David Parkes
11:40-12:05	*Improving Repeated Labeling for Crowdsourced Data Annotation* Sergiu Goschin, Chris Mesterharm, and Haym Hirsh
12:05-12:30	*Aggregating Human-Expert Votes using Stacked Generalization* Evgueni Smirnov, Hua Zhang, and Nikolay Nikolaev
12:30-14:00	***Lunch***
14:00-14:40	*Crowd IQ: Measuring the Intelligence of Crowdsourcing Platforms* Michal Kosinski, Yoram Bachrach, Gjergji Kasneci, Jurgen Van-Gael, and Thore Graepel
14:40-15:05	*Factor-based Regression Models for Forecasting* Chih-Chieh Cheng, Robert Sasseen, Tulay Muezzinoglu, and Richard Rohwer
15:05-15:30	*Dynamic Estimation of Rater Reliability in Regression Tasks using Multi-Armed Bandit Techniques* Alexey Tarasov, Sarah Jane Delany, and Brian Mac Namee
15:30-16:00	***Coffee***
16:00-16:25	*Crowdsourcing Microtasks Using Multiple Crowds* Ittai Abraham, Omar Alonso, Vasilis Kandylas, and Aleksandrs Slivkins
16:25-16:50	*Suggesting Constraints to Interactive Topic Modeling* Yuening Hu and Jordan Boyd-Graber
16:50-17:30	*General Discussion* All Participants

W06: Music and Machine Learning

June 30. Venue: DHT FRS.

Chairs: Rafael Ramirez, Darrell Conklin and José Manuel Iñesta

June 30

08:30-09:00 *Modeling Embellishment: Timing and Energy Expressive Transformations in Jazz Guitar*
Sergio Giraldo, Rafael Ramirez

09:00-09:30 *Generation of Tempo Curve for Performance Rendering of Piano Music with Automatic Phrase Analysis*
Naoya Ito, Satoru Fukayama, Daisuke Saito, Shigeki Sagayama

09:30-10:00 *Music Emotion Classification: Analysis of a Classifier Ensemble Approach*
Renato Panda, Rui Pedro Paiva

10:00-10:30 *Embedding Songs and Tags for Playlist Prediction*
Joshua L. Moore, Shuo Chen, Thorsten Joachims, Douglas Turnbull

10:30-11:00 ***Coffee break***

11:00-11:30 *A Comparison of Two Different Methods for Score-Informed Source Separation*
Joachim Fritsch, Joachim Ganseman, Mark D. Plumbley

11:30-12:00 *Learning content-based metrics for music similarity*
Sander Dieleman, Benjamin Schrauwen

12:00-12:30 *Group Non-negative Basis Pursuit for Automatic Music Transcription*
Ken O'Hanlon, Mark D. Plumbley

12:30-14:00 ***Lunch***

14:00-14:30 *Query Parsing Using Probabilistic Tree Grammars*
José F. Bernabeu, Jorge Calera-Rubio, José Iñesta, David Rizo

14:30-15:00 *Data mining of Basque folk music*
Darrell Conklin

15:00-15:30 *A Mixed Observability Markov Decision Process Model for Musical Pitch*
Pouyan Rafiei Fard, Keyvan Yahya

15:30-16:00 *Mutual Transform of Note Sequence and Melodic Envelope*
Yuichi Tsuchiya, Tetsuro Kitahara

16:00-16:30 ***Coffee break***

16:30-17:00 *Harmonic Analysis of Jazz MIDI Files Using Statistical Parsing*
Mark Granroth-Wilding, Mark Steedman

17:00-17:30 *Four-part Harmonization Using A Bayesian Network*
Syunpei Suzuki, Tetsuro Kitahara

17:30-18:00 *Multi-target pitch tracking of vibrato sources in noise using the GM-PHD filter*
Dan Stowell, Mark D. Plumbley

W07: Object, functional and structured data: towards next generation kernel-based methods

June 30. Venue: AT LT2.

Chairs: Florence d'Alché-Buc, Hachem Kadri, Massimiliano Pontil and Alain Rakotomamonjy,

June 30

08:45 - 09:00 *Welcome*
Organizers

09:00 - 09:45 *Object-oriented data analysis*
Steve Marron (invited)

09:45 - 10:30 *Nonlinear relationship and prediction in high dimensional regression setting*
Frédéric Ferraty (invited)

10:30 - 11:00 ***Posters & Coffe Break***

11:00 - 11:45 *Complex Prediction Problems and Application in Natural Language Processing*
Yasemin Altun (invited)

12:20 - 14:00 *Spotlights + Posters*
Participants

14:00 - 14:45 ***Lunch***

14:45 - 15:30 *TBA*
Charles Micchelli (invited)

15:30 - 16:00 *Kernel for vector-valued functions*
Neil Lawrence (invited)

16:00 - 16:45 *RKHS embeddings of distributions: theory and applications*
Arthur Gretton (invited)

16:45 - 17:30 *Open Problems*
Organizers + Participants

W08: Online Advertising (AdML)

June 30. Venue: IF G07a.

Chairs: Deepak Agarwal, Mikhail Bilenko, Olivier Chapelle, Brendan McMahan and D. Sculley

Online advertising is a multi-billion dollar industry that continues to show double-digit growth rates. The massive success of online advertising has been fueled in large part by the success of machine learning approaches. Yet the field is still very much in its infancy, with a range of exciting challenges still unsolved or only partially solved.

Efficiently serving online advertisements involves a number of prediction tasks that involve all facets of the application: advertisers and inventory, users and context. Examples of such problems include clickthrough and conversion prediction, forecasting yields, adversarial ad filtering, personalized advertisement matching, and relevance estimation. In addition to these, a number of larger-scale learning problems exist outside of the immediate ad serving context. The scope of these range from issues as broad as these include supply and demand forecasting, A/B test management, bandit problems, fraud prevention, and learning from highly biased historical data.

This workshop on machine learning for online advertising is intended to be geared to be useful and interesting to academics, researchers, and practitioners in industry. The workshop will be highlighted by invited talks from leading researchers in the field and talks from selected accepted papers. There will also be a special session on open problems in the field.

Please see the workshop website at `http://sites.google.com/site/admlworkshop` for updates and final schedule information.

June 30

8:30-9:00	***Coffee for arrival***
9:00-9:10	*Welcome* D. Sculley
9:10-9:40	*Invited Talk* Mikhail Bilenko
9:40-10:30	*Contributed Papers* TBD
10:30-11:00	***Coffee break***
11:00-12:00	*Keynote Talk* Leon Bottou
12:00-12:25	*Contributed Paper* TBD
12:25-14:00	***Lunch (on own)***

14:00-15:00	*Keynote Talk* Claudia Perlich
15:00-15:25	*Contributed Paper* TBD
15:25-16:00	***Coffee Break***
16:00-16:50	*Invited Talk* John Langford
16:50-17:30	*Discussion on Open Problems* D. Sculley (moderator)

W09: Sparsity, Dictionaries and Projections in ML and Signal Processing

June 30. Venue: AT LT4.

Chairs: Rodolphe Jenatton, Michael Davies, Rémi Gribonval and Guillaume Obozinski

June 30

9:00-9:45	*Backpropagation rules for dictionary learning* Julien Mairal
9:45-10:30	*Coherence-constrained multilinear identification* Pierre Comon (joint work with L-H.Lim)
10:30-11:00	***coffee***
11:00-11:30	*Learning Incoherent Dictionaries for Sparse Approximation using Iterative Projections and Rotations* Daniele Barchiesi and Mark D. Plumbley
11:30-12:00	*Robust joint reconstruction of misaligned images using semi-parametric dictionaries* Gilles Puy and Pierre Vandergheynst
12:00-12:30	*Probabilistic Subspace Clustering via Sparse Representations* Amir Adler, Michael Elad and Yacov Hel-Or
12:30-14:00	***lunch break***
14:00-14:45	*Sample complexity of dictionary learning through ϵ nets* Daniel Vainsencher (joint work with Shie Mannor and Alfred M. Bruckstein)
14:45-15:30	*TBA* Matthias Seeger
15:30-16:00	***coffee***
16:00-16:30	*Compressive mixture estimation* Anthony Bourrier and Rémi Gribonval
16:30-17:00	*Risk estimation for matrix recovery with spectral regularization* Charles-Alban Deledalle, Samuel Vaiter, Gabriel Peyré, Jalal Fadili and Charles Dossal

17:00-17:30 *Automatic Word Puzzle Generation via Topic Dictionaries*
Balász Pintér, Gyula V
or
os, Zoltán Szabó and L
orincz András

Poster Session

(1) *Learning a Semantically Relevant Multiple Sub-Space Visual Dictionary for Object Recognition*
Ashish Gupta

(2) *Graph Prediction in a Low-Rank and Autoregressive Setting*
Emile Richard, Pierre-André Savalle and Nicolas Vayatis

(3) *Degrees of Freedom of the Group Lasso*
Charles-Alban Deledalle, Samuel Vaiter, Gabriel Peyré, Jalal Fadili and Charles Dossal

(4) *A Block Sparsity Approach to Multiple Dictionary Learning for Audio Modeling*
Gautham J. Mysore

W10: Statistical Relational Learning (SRL)

June 30. Venue: IF G07.

Chairs: James Cussens, Kristian Kersting, Sriraam Natarajan and David Poole

In the tradition of the highly successful SRL workshops and Dagstuhl seminars in the past, SRL-2012 again will be the premier venue for bringing together different sub-disciplines that focus on the probabilistic analysis of relational data. So far, several novel insights into and useful approaches to SRL have been already provided. For instance, several different types of models have been developed. Recently, there has also been a surge in interest in learning the structure of SRL models – an important and challenging task. Still, there are many open problems and challenges lying ahead. How to deal with continuous variables? Or even hybrid models? What about distributed SRL? How do we lift inference? How do we lift learning? Is lifting beneficial for models traditionally considered to be propositional such as Ising and Potts models? How do we deal with evolving environments? What are the next killer applications? Is it computer vision or is it bio-medicine? Which lessons can SRL learn from other, related communities? Motivated by this, we are organizing the Statistical Relational Learning workshop at ICML. We have three exciting invited speakers – Lise Getoor, David Page and Alex Smola. We also have 9 interesting papers that will be presented as posters. The poster sessions and discussions will be jointly held with the MLG workshop.

June 30

08:30 - 09:00 ***Coffee for arrival***

09:00 - 09:10 ***Opening***

09:10 - 10:10 *Invited Talk*
Lise Getoor

10:10 - 10:30 *Poster Highlights*
Two papers

10:30 - 11:00 ***Break***

11:00 - 12:10 *Poster Highlights*
Seven papers

12:10 - 14:00 ***Lunch***

14:00 - 15:00 *Invited Talk*
Alex Smola

15:00 - 16:30 ***Poster Session. coffee break at 15:30***

16:30 - 17:30 *Invited talk*
David Page

17:30 - 18:00 *Open Discussion*
Panel

Poster Session

(1) *Viterbi training in PRISM*
Taisuke Sato and Keiichi Kubota

(2) *Challenge Paper: Marginal Probabilities for Instances and Classes*
Oliver Schulte

(3) *Modelling Relational Statistics With Bayes Nets*
Oliver Schulte and Hassan Khosravi

(4) *A statistical relational learning challenge – extracting proof strategies from exemplar proofs*
Gudmund Grov and Ekaterina Komendantskaya

(5) *Structure Learning with Hidden Data in Relational Domains*
Tushar Khot, Sriraam Natarajan, Kristian Kersting and Jude Shavlik

(6) *Lifted Parameter Learning in Relational Models*
Babak Ahmadi, Kristian Kersting and Sriraam Natarajan

(7) *Learning Multiple Hierarchical Relational Clusterings*
Aniruddh Nath and Pedro Domingos

(8) *Tractable Markov Logic*
Pedro Domingos and Austin Webb

(9) *Learning Recursive PRISM Programs with Observed Outcomes*
Waleed Alsanie and James Cussens

W11: Teaching ML

June 30. Venue: IF 4.31/33.

Chairs: Kurt Driessens and Elisa Fromont

June 30

9:00-9:30	*Opening* Elisa Fromont, Kurt Driessens
9:30:-10:30	*Invited talk : ml-class.org: Teaching machine learning to 100,000 students* Andrew Ng
10:30-11:00	***Coffee***
11:00-12:00	*Invited talk : Some perspectives on teaching machine learning* David Barber
12:00-13:30	***Lunch***
13:30-14:30	*Invited talk : Machine Learning: Unity in Diversity (Preliminary title)* Peter Flach
14:30-15:30	*Position talks* tba
15:30-16:00	***Coffee***
16:00-17:00	*Poster session for student assignments* tba
17:00-17:30	*Debate and open questions* Elisa Fromont, Kurt Driessens

W12: Machine Learning in Genetics and Genomics (MLGG)

July 1. Venue: IF 4.31/33.

Chairs: Anna Goldenberg, Pierre Baldi, Sara Mostafavi and Michal Rosen-Zvia

The field of computational biology has seen a dramatic growth over the past few years—not only in terms of new available data, but also in new scientific questions, and new challenges for learning and inference. For instance, multiple types of large-scale (often genome-wide) datasets, such as gene expression, genotyping data, whole-genome sequences, protein-protein interactions, and protein abundance measurements, are widely available for multiple model organisms and across multiple conditions, and some of these data types are also available for large patient cohorts. Computational approaches for analyzing and learning from these data are faced with major challenges including scalability, data heterogeneity, missing data and confounding factors to name a few. The goal of this workshop is to present emerging genomics research questions and machine learning techniques that can address some of the challenges on the way to answering fundamental biological questions and refining our understanding of the genesis and progression of diseases.

July 1

9:00-9:50	*Invited talk by Yves Moreau (title TBA)* Yves Moreau
9:50-10:10	*Identifying differentially expressed transcripts from RNA-seq data with biological variation* Peter Glaus, Antti Honkela and Magnus Rattray
10:10-10:30	*Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data* Richard Savage, David Wild, Paul Kirk, Zoubin Ghahramani and Jim Griffin
1:30-11:00	***COFFEE BREAK***
11:00-11:50	*Invited talk by Florence d'Alché-Buc (title TBA)* Florence d'Alché-Buc
11:50-12:10	*The role of indirect connections in gene networks in predicting function* Jesse Gillis and Paul Pavlidis
12:10-12:30	*Using Prior Biological Knowledge when Constructing Regulatory Networks* Sara Mostafavi and Daphne Koller

12:30-14:00 ***LUNCH BREAK***

14:00-14:50 *Invited talk by Samuel Kaski : Data translation in genomics with multi-view machine learning*
Samuel Kaski

14:50-15:10 *Combining multiple biological domains using patient network fusion*
Bo Wang, Aziz M. Mezlini and Anna Goldenberg

15:10-15:30 *Using machine learning techniques for selecting optimal antiretroviral treatment for HIV*
Michal Rosen-Zvi et al.

15:30-16:00 ***COFFEE BREAK***

16:00-17:00 *Panel discussion: Role of machine learning in future medicine*
Mark Girolami, Pierre Baldi, Sam Kaski, Michal Rosen-Zvi

W13: Markets, Mechanisms and Multi-Agent Models

July 1. Venue: IF G07a.

Chairs: Amos Storkey, Jacob Abernethy and Jenn Wortman Vaughan

Many of society's greatest accomplishments are in large part due to the facility of markets. Markets and other allocation mechanisms have become necessary tools of the modern age, and they have been key to facilitating the development of complex structures, advanced engineering, and a range of other improvements to our collective capabilities. Much work in economics has been done to demonstrate that markets can, in aggregate, function very well even when the individual participants are noisy, irrational or myopic.

In terms of aims and benefits, the design of machine learning techniques has much in common with the development of market mechanisms: information aggregation, maximal efficiency, scalability, and, more recently, decentralization. Current machine learning algorithms are often single goal methods, built from simple homogeneous units by one person or individual groups. Perhaps looking to the organisations of economies may help in moving beyond the current centralised design of most machine learning methods. Allowing agents with different opinions, approaches or methods to enter and leave the market, to interact, and to adapt to changes can have many benefits. For example it may enable us to develop methods that provide continuous improvement on complex problems, reuse results by improving on previous outcomes rather than building bigger models from scratch, and adapting to changes.

There are many relationships between machine learning methods, Bayesian decision theory, risk minimisation, economics, statistical physics and information theory that have been known for some time. There are also many open questions regarding the full nature and impact of these connections. This workshop will explore these connections from many different directions.

July 1

09.00-09.20 *Introduction: Markets, Mechanisms and Multi-agent Models*
Amos Storkey

09:20-10:00 *Invited Speaker: The Artificial Prediction Market*
Adrian Barbu

10:00-10:30 *Interpreting prediction markets: a stochastic approach*
Rafael M. Frongillo,Nicolas Della Penna ,Mark D. Reid

10:30-11:00 ***Coffee Break***

11:00-11:30 *Dynamic Pricing with Limited Supply*
Moshe Babaioff,Shaddin Dughmi,Robert Kleinberg,Aleksandrs Slivkins

11:30-12:00 *Approximating Equilibria in Sequential Auctions with Incomplete Information and Multi-Unit Demand*
Amy Greenwald,Jiacui Li,Eric Sodomka

12:00-12:30 *Towards auction-based multi-agent collision-avoidance under continuous stochastic dynamics*
Jan-Peter Calliess,Michael A. Osborne,Stephen Roberts

12:30-14:00 ***Lunch***

14:00-14:20 ***Interim Summary and Discussion***

14:20-15:00 *Invited Speaker: A Tractable Combinatorial Market Maker using Constraint Generation*
Sebastien Lahaie

15:00-15:30 *Designing Duality-Based Market Makers with Adaptive Liquidity*
Xiaolong Li,Jennifer Wortman Vaughan

15:30-16:00 ***Coffee and discussion***

16:00-16:30 *Risk Aversion in Multi armed Bandits*
Amir Sani,Alessandro Lazaric,Remi Munos

16:30-17:00 *The Weighted Majority Algorithm does not Converge in Nearly Zero-sum Games*
Maria-Florina Balcan,Florin Constantin,Ruta Mehta

17:00-17:30 ***Discussion and Conclusion***

W14: Mining and Learning with Graphs (MLG)

July 1. Venue: AT LT2.

Chairs: Jan Ramon and Hanghang Tong

The workshop on Mining and Learning with Graphs (MLG) is concerned with analyzing data that is represented as a graph. Examples include the WWW, social networks, biological networks, communication networks, food webs, and many others. A number of subareas have been working with graph structured data, including graph mining, learning from structured data, statistical relational learning, inductive logic programming, and, moving beyond sub-disciplines in computer science, social network analysis, and, more broadly network science.

The objective of this workshop is to bring together researchers from a variety of areas, to discuss commonality and differences in challenges faced, summarize some of the different approaches, share the preliminary results, new ideas and methodologies, and provide a forum to present and learn about the preliminary result, new ideas, and some of the most cutting edge research in this area.

July 1

8:30-9:00	***Coffee***
9:00-9:45	*Invited talk: Link prediction* Deepayan Chakrabarti (Facebook)
9:45-10:07	*A Scalable Multiclass Algorithm for Node Classification* Giovanni Zappella
10:07-10:29	*Entropic Regularization of Mixed-membership Network Models using Pseudo-observations* Ramnath Balasubramanyan and William Cohen
10:30-11:00	***Coffee***
11:00-12:30	***Poster Session and discussion***
12:30-14:00	***Lunch***
14:00-14:45	*Invited talk: A Linear Time Active Learning Algorithm for Link Classification* Claudio Gentille
14:45-15:07	*Query-driven Active Surveying for Collective Classification* Galileo Mark Namata, Ben London, Lise Getoor and Bert Huang

14:07-15:29 *Low-distortion Inference of Latent Similarities from a Multiplex Social Network*
Ittai Abraham, Shiri Chechik, David Kempe and Aleksandrs Slivkins

15:30-16:00 ***Coffee***

16:00-16:22 *Active Sampling of Networks*
Joseph Pfeiffer, Jennifer Neville and Paul Bennett

16:22-16:44 *Propagation Kernels for Partially Labeled Graphs*
Marion Neumann, Roman Garnett, Plinio Moreno, Novi Patricia and Kristian Kersting

16:44-17:06 *Attribute-based Decision Graphs for Data Classification*
João Bertini, Maria Nicoletti and Liang Zhao

17:06-17:20 ***Closing***

W15: New Challenges for Exploration and Exploitation

July 1. Venue: DHT FRN.

Chairs: Jérémie Mary, Aurélien Garivier, Lihong Li, Rémi Munos, Olivier Nicol, Ronald Ortner and Philippe Preux

The workshop will be about trading exploration and exploitation in learning problems, and the techniques used in practice to achieve a good balance. In particular, the workshop will focus on web-related applications. It will provide a space to present and discuss the approaches and techniques used by the participants of the Exploration and Exploitation Challenge 2012 (`http://explochallenge.inria.fr/`), which is running on a dataset collected by Yahoo!.

July 1

9:00-10:00 *Causal reasoning and Learning Systems.*
Leon Bottou, Jonas Peters, D. Max Chickering, Patrice Simard

10:00-10:30 *On Bayesian Upper Confidence Bounds for Bandits Problems and Optimality of Thomson sampling*
Émilie Kaufmann

10:30-11:00 ***Coffee break***

11:00-12:00 *Two basic problems in finite stochastic optimization*
Sebastien Bubeck

12:00-12:10 *Dynamic Pricing with Limited Supply*
Moshe Babaioff,Shaddin Dughmi, Robert Kleinberg, Aleksandrs Slivkins

12:15-12:30 *Learning Hurdles for Sleeping Experts*
Varun Kanade, Thomas Steinke

12:30-12:45 *Contextual Bandits: Approximated Linear Bayes for Large Contexts*
Jose Antonio Martin H.

12:45-14:00 ***Lunch***

14:00-15:00 *The knowledge gradient for optimal learning*
Warren B. Powell

15:00-15:15 *Introduction to the Exploration & Exploitation Challenge*
Jérémie Mary

15:15-15:35 *Second team of the Phase 1 presentation*

15:35-15:45 *Winner of phase 1 presentation*

15:45-16:00 ***Coffee break***

16:00-16:10 *Annoucement of winners of Phase 2*

16:10-17:10 *Open discussion on future of exploration & exploitation*

W16: Optimization in Machine Learning

July 1. Venue: AT LT4.

Chairs: Katya Scheinberg, Ben Recht and Sasha Rakhlin

The focus on optimization algorithms in machine learning community in the past decade has expanded at a high rate. The number of papers at each major conference dedicated to optimization is growing from one year to another. By the same token, several major areas of optimization have enjoyed significant growth due to the applications arising in machine learning. Large scale convex optimization and stochastic optimization are two such areas. Our one-day workshop consists of several invited tutorial-style talks by prominent researchers in the area. The talks by the invited speakers are intended to cover a variety of important topics on optimization in ML. In addition several posters will be presented and introduced during the poster spotlight session.

July 1

9:45 - 10:30 *TBC*
Peter Richtárik (University of Edinburgh)

10:30 - 11:00 ***Coffee Break***

11:00- 11:45 *TBC*
Anatoli Juditsky (Université Joseph Fourier)

11:45 - 12:15 ***Poster Spotlights***

12:15 - 14:00 ***Lunch***

14:00 - 14:45 *TBC*
Ali Jadbabaie (University of Pennsylvania)

14:45- 15:30 *TBC*
Karthik Sridharan (University of Pennsylvania)

15:30 - 16:00 ***Coffee Break***

16:00- 16:45 *TBC*
Michael Mahoney (Stanford University)

W17: Representation Learning

July 1. Venue: AT LT5.

Chairs: Aaron Courville, Hugo Larochelle, MarcÁurelio Ranzato and Yoshua Bengio

In this workshop we consider the question of how we can learn meaningful and useful representations of the data. There has been a great deal of recent work on this topic, much of it emerging from researchers interested in training deep architectures. Deep learning methods such as deep belief networks, sparse coding-based methods, convolutional networks, and deep Boltzmann machines, have shown promise as a means of learning invariant representations of data and have already been successfully applied to a variety of tasks in computer vision, audio processing, natural language processing, information retrieval, and robotics. Bayesian nonparametric methods and other hierarchical graphical model-based approaches have also been recently shown the ability to learn rich representations of data. By bringing together researchers with diverse expertise and perspectives but who are all interested in the question of how to learn data representations, we will explore the challenges and promising directions for future research in this area.

July 1

9:00-9:20	*Overview of representation learning methods and the challenges we face* Yoshua Bengio
9:20-10:00	*Invited talk: Hierarchical Bayesian Multi-scale Representations for Imagery* Lawrence Carin
10:00-10:40	*Invited talk: Deconvolutional Networks* Rob Fergus
10:40-11:00	***Coffee Break***
11:00-11:40	*Invited talk: Challenges and Progress in Learning Representations* Ryan Adams
11:40-12:00	*Learning a selectivity-invariance-selectivity feature extraction architecture for images* Michael Gutmann and Aapo Hyvarinen
12:00-12:30	***Poster Session***
12:30-14:00	***Lunch***

14:00-14:40 *Invited talk: Scaling Deep Learning*
Jeff Dean

14:40-15:20 *Invited talk: Semi-supervised Deep Inference Machines for Multi-modal Perception*
Drew Bagnell

15:20-15:40 *Sequence Transduction with Recurrent Neural Networks*
Alex Graves

15:40-16:00 ***Coffee Break***

16:00-16:40 *Invited talk: Some thoughts on learning representations for text, images, music and knowledge*
Jason Weston

16:40-17:20 *Invited talk: Exploiting compositionality to explore a large space of model structures*
Roger Grosse

17:20-17:40 *On Linear Embeddings and Unsupervised Feature Learning*
Ryan Kiros and Csaba Szepesvari

17:40-18:30 ***Poster Session***

W18: RKHS and Kernel-based methods

July 1. Venue: IF G07.

Chairs: Arthur Gretton, Zaid Harchaoui and Bharath Sriperumbudur

July 1

9:00-9:45 *Distances and Kernels on Discrete Structures*
Marco Cuturi

9:45-10:30 *Theory of reproducing kernels and its general applications*
Saburou Saitoh

10:30 - 10:45 ***Short coffee break***

10:45-11:30 *Kernels for learning solutions of PDEs*
Robert Schaback

11:30-12:15 ***Spotlights***

12:15-13:45 ***Lunch***

13:45-14:30 *Reproducing kernel Hilbert spaces and smoothing splines*
Paul Eggermont

14:30 - 15:15 ***Contributed talk***

15:15-16:00 ***Poster session and coffee***

16:00-16:45 ***Contributed talk***

16:45-17:30 *Learning geometry with kernels*
Lorenzo Rosasco

W19: Statistics, ML and Neuroscience (STAMLINS)

July 1. Venue: DHT FRS.

Chairs: Iead Rezek, Evangelos Roussos and Christian Beckmann

Neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), have revolutionised the field of neuroscience by providing non-invasive measurements of the brain activity. The mathematical models, linking the observations and domain-specific variables of interest and algorithms for making inferences given the data and the models, fall into two difference strands. The first and classical statistical approach, which is largely based on known models and highly-constrained experimental conditions, offers good interpretability (because of simpler models), the possibility to examine strength of the analysis using well-established hypothesis tests, and a means to determine the functional specialisation of the human brain. The second machine learning approach learns the model itself, which is typically abstract and generic, to widen the model space. Also, the learning algorithms are generally tuned to obtain optimal prediction performance, automation and speed. Both these strands represent an aspect of the scientific method: hypothesis generation, represented by machine learning, on the one hand, and hypothesis validation, represented by classical statistics, on the other. The aim of this workshop is to debate the strengths/weaknesses of classical statistical vs. machine learning methods, and establish the parameters that would bring together hypothesis and data driven approaches. This workshop is to be driven by the practical needs of neuroscience research. Simultaneously, we expect that the need for improved methods will spark more fundamental research and substantial contributions to be made towards Predictive Medicine, Translational Medicine and Interpretational Models of Disease.

July 1

08:30-09:00 ***Coffee for arrival***

09:00-09:10 *Welcome Note*
I. Rezek, E. Roussos, C. Beckmann

09:10-09:40 *Nonlinear fMRI time series analysis*
William Penny

09:40-10:10 *Limitations and strengths of classical statistical methods of fMRI analysis.*
Thomas E. Nichols

10:10-10:30 *Discussion*
All participants

10:30-11:00 ***Coffee***

11:00-11:30 *Decoding of cognitive states: from patterns to mechanisms*
Marcel van Gerven

11:30-12:00 *Inference in brain imaging: sources of uncertainty*
Eugene Duff

12:00:12:30 *Discussion*
All participants

12:30-14:00 ***Lunch***

14:00-14:30 ***Spotlight Presentations***

14:30-15:30 ***Poster Presentations***

15:30-16:00 ***Coffee***

16:00-17:30 *Discussion/Round-up*
All participants

Index of Names

– C –

– D –

– E –

– F –

– I –

– J –

– K –

– L –

– M –

– N –

– O –

– P –

– Q –

– T –

– U –

– V –

– W –

– X –

– Y –

– Z –

12 Organizing Committee

General Chair

Andrew McCallum (University of Massachusetts Amherst)

Program Chairs

John Langford (Yahoo! Research)
Joelle Pineau (McGill University)

Local Chair

Charles Sutton (University of Edinburgh)

Workshop Chairs

Francis Bach (INRIA)
Irina Rish (IBM Research)

Tutorial Chairs

Rob Schapire (Princeton University)
Olivier Chapelle (Yahoo! Research)

Publication Chairs

Kilian Weinberger (Washington University, St. Louis)
Amir Globerson (Hebrew University of Jerusalem)

Publicity Chair

Jingrui He (IBM Research)

Scholarship Chair

Jesse Davis (KU Leuven)

Funding Chair

S V N Vishwanathan (Purdue University)

Volunteers and Registration Chair

Iain Murray (University of Edinburgh)

Local Organizing Committee

Chris Williams (University of Edinburgh)
Amos Storkey (University of Edinburgh)
Guido Sanguinetti (University of Edinburgh)
Sethu Vijayakumar (University of Edinburgh)

Workflow Manager

Mahdi Milani Fard (McGill University)

Webmaster

Francesco Figari (University of Edinburgh)

Area Chairs

Alan Fern (Oregon State University)
Alina Beygelzimer (IBM Research)
Andreas Krause (California Institute of Technology)
Arindam Banerjee (University of Minnesota)
Arthur Gretton (University College London)
Ben Taskar (University of Pennsylvania)
Charles Elkan (University of California)
Claire Monteleoni (Columbia University)
Csaba Szepesvari (University of Alberta)
Dale Schuurmans (University of Alberta)
Daniel Hsu (Microsoft Research)
David Blei (Princeton University)
David McAllester (University of Chicago)
Drew Bagnell (Carnegie Mellon University)
Elad Hazan (Israel Institute of Technology)
Filip Radlinski (Microsoft Research)
Frank McSherry (Microsoft Research)
Geoff Gordon (Carnegie Mellon)
MAPS Hal Daume (University of Maryland)
Hoifung Poon (Microsoft Research)

Inderjit Dhillon (University of Texas)
Jennifer Wortman Vaughan (University of California, Los Angeles)
Jerry Zhu (University of Wisconsin-Madison)
Jure Leskovec (Stanford University)
Kilian Weinberger (Washington University)
Kristian Kersting (University of Bonn)
Lawrence Carin (Duke University)
Lihong Li (Yahoo! Research)
Marc'Aurelio Ranzato (University of Toronto)
Mario Marchand (Université Laval)
Michael Bowling (University of Alberta)
Miroslav Dudik (Yahoo! Research)
Naoki Abe (IBM Research)
Olivier Chapelle (Yahoo! Research)
Pedro Domingos University of Washington)
Peter Stone (University of Texas)
Pradeep Ravikumar (University of Texas)
Rich Caruana (Microsoft Research)
Rich Zemel (University of Toronto)
Ruslan Salakhutdinov (Massachusetts Institute of Technology)
S V N Vishwanathan (Purdue University)
Samy Bengio (Google)
Sanjoy Dasgupta (University of California, San Diego)
Satyen Kale (IBM Research)
Thorsten Joachims (Cornell University)
Tobias Scheffer (University of Potsdam)
Tong Zhang (Rutgers University)
Yee Whye Teh (University College London)
Yoshua Bengio (Université de Montréal)
Zoubin Ghahramani (University of Cambridge)

Appleton Tower (AT)