Causal Inference and Stable Learning

Peng Cui
Tsinghua University

Tong Zhang
Hong Kong University of Science and Technology
ML techniques are impacting our life

• A day in our life with ML techniques
Now we are stepping into risk-sensitive areas

Shifting from Performance Driven to Risk Sensitive
Problems of today’s ML - *Explainability*

Most machine learning models are black-box models.

Unexplainable

Human in the loop

Health  Military  Finance  Industry
Problems of today’s ML - *Stability*

Most ML methods are developed under I.I.D hypothesis
Problems of today’s ML - Stability
Problems of today’s ML - *Stability*

- Cancer survival rate prediction

**Training Data**

*City Hospital*
Higher income, higher survival rate.

**Predictive Model**

**Testing Data**

*City Hospital*

- Higher income, higher survival rate.

*University Hospital*

- Survival rate is not so correlated with income.
A plausible reason: **Correlation**

Correlation is the very basics of machine learning.
Correlation is not explainable
Correlation is ‘unstable’
It's not the fault of correlation, but the way we use it

• Three sources of correlation:
  • Causation
    • Causal mechanism
    • Stable and explainable
  • Confounding
    • Ignoring X
    • Spurious Correlation
  • Sample Selection Bias
    • Conditional on S
    • Spurious Correlation
A Practical Definition of Causality

Definition: T causes Y if and only if changing T leads to a change in Y, while keeping everything else constant.

Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

*Interventionist definition [http://plato.stanford.edu/entries/causation-mani/]
The benefits of bringing causality into learning

Causal Framework

Grass—Label: Strong correlation
Weak causation

Dog nose—Label: Strong correlation
Strong causation

More Explainable and More Stable
The *gap* between causality and learning

- How to evaluate the outcome?
- Wild environments
  - High-dimensional
  - Highly noisy
  - Little prior knowledge (model specification, confounding structures)
- Targeting problems
  - Understanding v.s. Prediction
  - Depth v.s. Scale and Performance

How to bridge the gap between *causality* and *(stable) learning*?
Outline

- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Conclusions
Paradigms - Structural Causal Model

A graphical model to describe the causal mechanisms of a system

• Causal Identification with back door criterion
• Causal Estimation with do calculus

How to discover the causal structure?
Paradigms – Structural Causal Model

- Causal Discovery
  - Constraint-based: conditional independence
  - Functional causal model based

A generative model with strong expressive power. But it induces high complexity.
Paradigms - Potential Outcome Framework

• A simpler setting
  • Suppose the confounders of T are known a priori

• The computational complexity is affordable
  • Under stronger assumptions
  • E.g. all confounders need to be observed

More like a **discriminative** way to estimate treatment’s partial effect on outcome.
Causal Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Treated Group ($T = 1$) and Control Group ($T = 0$)
- Potential Outcome: $Y(T = 1)$ and $Y(T = 0)$
- Average Causal Effect of Treatment (ATE):

\[
ATE = E[Y(T = 1) - Y(T = 0)]
\]
Counterfactual Problem

- Two key points for causal effect estimation
  - Changing $T$
  - Keeping everything else constant

- For each person, observe only one: either $Y_{T=1}$ or $Y_{T=0}$
- For different group ($T=1$ and $T=0$), something else are not constant

<table>
<thead>
<tr>
<th>Person</th>
<th>$T$</th>
<th>$Y_{T=1}$</th>
<th>$Y_{T=0}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>0.4</td>
<td>?</td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>?</td>
<td>0.6</td>
</tr>
<tr>
<td>P3</td>
<td>1</td>
<td>0.3</td>
<td>?</td>
</tr>
<tr>
<td>P4</td>
<td>0</td>
<td>?</td>
<td>0.1</td>
</tr>
<tr>
<td>P5</td>
<td>1</td>
<td>0.5</td>
<td>?</td>
</tr>
<tr>
<td>P6</td>
<td>0</td>
<td>?</td>
<td>0.5</td>
</tr>
<tr>
<td>P7</td>
<td>0</td>
<td>?</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Ideal Solution: Counterfactual World

- Reason about a world that does not exist
- Everything in the counterfactual world is the same as the real world, except the treatment

\[ Y(T = 1) \]  
\[ Y(T = 0) \]
Randomized Experiments are the “Gold Standard”

- Drawbacks:
  - Cost
  - Unethical
  - Unrealistic
Causal Inference with Observational Data

• Counterfactual Problem:

\[ Y(T = 1) \quad \text{or} \quad Y(T = 0) \]

• Can we estimate ATE by directly comparing the average outcome between treated and control groups?
  • Yes with randomized experiments (\( X \) are the same)
  • No with observational data (\( X \) might be different)
Confounding Effect

Balancing Confounders’ Distribution
Methods for Causal Inference

- Matching
- Propensity Score
- Directly Confounder Balancing
Matching

$T = 0$

$T = 1$
Matching
Matching

• Identify pairs of treated (T=1) and control (T=0) units whose confounders $X$ are similar or even identical to each other

$\text{Distance}(X_i, X_j) \leq \epsilon$

• Paired units guarantee that the everything else (Confounders) approximate constant

• Small $\epsilon$: less bias, but higher variance

• Fit for low-dimensional settings

• But in high-dimensional settings, there will be few exact matches
Methods for Causal Inference

• Matching

• Propensity Score

• Directly Confounder Balancing
Propensity Score Based Methods

- Propensity score \( e(X) \) is the probability of a unit to get treated
  \[
  e(X) = P(T = 1|X)
  \]

- Then, Donald Rubin shows that the propensity score is sufficient to control or summarize the information of confounders
  \[
  T \perp X | e(X) \quad \Rightarrow \quad T \perp (Y(1), Y(0)) | e(X)
  \]

- Propensity scores cannot be observed, need to be estimated
Propensity Score Matching

• Estimating propensity score: \( \hat{e}(X) = P(T = 1|X) \)
  
  • **Supervised learning**: predicting a known label \( T \) based on observed covariates \( X \).
  
  • Conventionally, use logistic regression

• Matching pairs by distance between propensity score:

\[
Distance(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|
\]

• High dimensional challenge: from matching to PS estimation

---

Inverse of Propensity Weighting (IPW)

- Why weighting with inverse of propensity score?
  - Propensity score induces the distribution bias on confounders $X$

$$e(X) = P(T = 1|X)$$

Reweighting by inverse of propensity score:

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

Inverse of Propensity Weighting (IPW)

• Estimating ATE by IPW [1]:

\[ ATE_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i)Y_i}{1 - \hat{e}(X_i)} \]

\[ w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i} \]

• Interpretation: IPW creates a pseudo-population where the confounders are the same between treated and control groups.
• But requires correct model specification for propensity score
• High variance when \( e \) is close to 0 or 1

Non-parametric solution

- Model specification problem is inevitable
- Can we directly learn sample weights that can balance confounders’ distribution between treated and control groups?
Methods for Causal Inference

• Matching

• Propensity Score

• Directly Confounder Balancing
Directly Confounder Balancing

**Motivation**: The collection of all the moments of variables uniquely determine their distributions.

**Methods**: Learning sample weights by directly balancing confounders’ moments as follows (ATT problem)

\[
\min_W \, \left\| \overline{X}_t - X_c^T W \right\|_2^2
\]

With moments, the sample weights can be learned without any model specification.

Entropy Balancing

\[
\begin{align*}
\min_{W} & \quad W \log(W) \\
\text{s.t.} & \quad \| \mathbf{X}_t - \mathbf{X}_c^T W \|_2^2 = 0 \\
& \quad \sum_{i=1}^{n} W_i = 1, W \geq 0
\end{align*}
\]

- Directly confounder balancing by sample weights \( W \)
- Minimize the entropy of sample weights \( W \)

Either know confounders a priori or regard all variables as confounders. All confounders are balanced equally.

Differentiated Confounder Balancing

**Idea**: Different confounders make different confounding bias

- Simultaneously learn *confounder weights* $\beta$ and *sample weights* $W$.

$$
\min \left( \beta^T \cdot (\bar{X}_t - X_c^T W) \right)^2
$$

- *Confounder weights* determine which variable is confounder and its contribution on confounding bias.
- *Sample weights* are designed for confounder balancing.

Differentiated Confounder Balancing

- General relationship among $X$, $T$, and $Y$:

  \[ Y = f(X) + T \cdot g(X) + \epsilon \]

  \[ AT\bar{T} = E\left(g(X_t)\right) \]

  \[ Y(0) = f(X) + \epsilon \]

  \[ f(X) = a_1 X + \sum_{ij} a_{ij} X_i X_j + \sum_{ijk} a_{ijk} X_i X_j X_k + \cdots + R_n(X) \]

  \[ = \alpha M. \]

  \[ M = (X, X_i X_j, X_i X_j X_k, \cdots). \]

  Confounder weights

  Confounding bias

  \[ \widehat{AT\bar{T}} = AT\bar{T} + \sum_{k=1}^{p} \alpha_k \left( \sum_{i:T_i=1} \frac{1}{n_t} M_{i,k} - \sum_{j:T_j=0} W_j M_{j,k} \right) + \phi(\epsilon). \]

  If $\alpha_k = 0$, then $M_k$ is not confounder, no need to balance.

  Different confounders have different confounding weights.

Differentiated Confounder Balancing

• **Ideas**: simultaneously learn *confounder weights* $\beta$ and *sample weighs* $W$.

$$\min \left( \beta^T \cdot (M_t - M_c^T W) \right)^2 + \lambda \sum_{j: T_j = 0} (1 + W_j) \cdot (Y_j - M_j \cdot \beta)^2,$$

s.t. $\|W\|_2^2 \leq \delta$, $\|\beta\|_2^2 \leq \mu$, $\|\beta\|_1 \leq \nu$, $1^T W = 1$ and $W \succeq 0$

• **Confounder weights** determine which variable is confounder and its contribution on confounding bias.

• **Sample weights** are designed for confounder balancing.

• The ENT algorithm is a special case of DCB algorithm by setting the confounder weights as unit vector.

Experiments

Assumptions of Causal Inference

• **A1: Stable Unit Treatment Value (SUTV):** The effect of treatment on a unit is independent of the treatment assignment of other units

\[ P(Y_i|T_i, T_j, X_i) = P(Y_i|T_i, X_i) \]

• **A2: Unconfounderness:** The distribution of treatment is independent of potential outcome when given the observed variables

\[ T \perp (Y(0), Y(1)) | X \]

No unmeasured confounders

• **A3: Overlap:** Each unit has nonzero probability to receive either treatment status when given the observed variables

\[ 0 < P(T = 1|X = x) < 1 \]
Sectional Summary

- Progress has been made to draw causality from big data.
- From single to group
- From binary to continuous
- Weak assumptions

Ready for Learning?
Outline

- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Future Directions and Conclusions
Stability and Prediction

Prediction Performance

Learning Process

True Model

Bin Yu (2016), Three Principles of Data Science: predictability, computability, stability
Stable Learning

Model

Distribution 1

Distribution 2

Distribution 3

Distribution n

Accuracy 1

Accuracy 2

Accuracy 3

Accuracy n

Training

Testing

I.I.D. Learning

VAR (Acc)

Stable Learning

Transfer Learning
Stability and Robustness

• Robustness
  • More on prediction performance over data perturbations
  • Prediction performance-driven

• Stability
  • More on the true model
  • Lay more emphasis on Bias
  • Sufficient for robustness

Stable learning is a (intrinsic?) way to realize robust prediction
Stability

• Statistical stability holds if statistical conclusions are robust to appropriate perturbations to data.
  • Prediction Stability
  • Estimation Stability

*Bernoulli* 19(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

Departments of Statistics and EECS, University of California at Berkeley, Berkeley, CA 94720, USA.
E-mail: binyu@stat.berkeley.edu
Prediction Stability

- **Lasso**

\[ \hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \]

- **Prediction Stability by Cross-Validation**
  - n data units are randomly partitioned into V blocks, each block has \( d = \lfloor n/V \rfloor \) units.
  - Leave one out: training on \( (n-d) \) units, validating on \( d \) units.
  - CV does not provide a good interpretable model because Lasso+CV is unstable.
Estimation Stability

- Estimation Stability:
  - Mean regression function:
  \[ \hat{m}(\tau) = \frac{1}{V} \sum_{v} X \hat{\beta}_v(\tau), \]
  
  - Variance of function \( m \):
  \[ \hat{T}(\tau) = \frac{n - d}{d} \frac{1}{V} \sum_{v} (\| X \hat{\beta}_v(\tau) - \hat{m}(\tau) \| ^2). \]
  
  - Estimation Stability:
  \[ ES(\tau) = \frac{1/V \sum_v \| X \hat{\beta}_v(\tau) - \hat{m}(\tau) \| ^2}{\hat{m}^2(\tau)} = \frac{d}{n - d} \frac{\hat{T}(\tau)}{\hat{m}^2(\tau)} \]

ES+CV is better than Lasso+CV
Domain Generalization / Invariant Learning

- Given data from different observed environments $e \in \mathcal{E}$:
  
  $$(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$$

- The task is to predict $Y$ given $X$ such that the prediction works well (is “robust”) for “all possible” (including unseen) environments
Domain Generalization

- **Assumption**: the conditional probability $P(Y|X)$ is stable or invariant across different environments.
- **Idea**: taking knowledge acquired from a number of related domains and applying it to previously unseen domains.
- **Theorem**: Under reasonable technical assumptions. Then with probability at least $1 - \delta$

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \mathbb{E}_{\mathcal{D}}^{*} \mathbb{E}_{\mathcal{P}} \ell(f(\tilde{X}_{ij}), Y_i) - \mathbb{E}_{\mathcal{P}} \ell(f(\tilde{X}_{ij}), Y_i) \right|^2$$

$$\leq c_1 \cdot \mathbb{V}_{\mathcal{H}}(\mathbb{P}^1, \mathbb{P}^2, \ldots, \mathbb{P}^N) + c_2 \frac{N \cdot (\log \delta^{-1} + 2 \log N)}{n} + c_3 \frac{\log \delta^{-1}}{N} + \frac{c_4}{N}$$

$distributional$ variance $vanish$ as $N, n \to \infty$

Invariant Prediction

• **Invariant Assumption**: There exists a subset $S \in X$ is causal for the prediction of $Y$, and the conditional distribution $P(Y|S)$ is stable across all environments.

  for all $e \in \mathcal{E}$, $X^e$ has an arbitrary distribution and

  $$Y^e = g(X^e_{S^*}, \varepsilon^e), \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \Perp X^e_{S^*}$$

• **Idea: Linking to causality**
  
  • Structural Causal Model (Pearl 2009):
  
  • The parent variables of $Y$ in SCM satisfies Invariant Assumption
  
  • The causal variables lead to invariance w.r.t. “all” possible environments

From Variable Selection to Sample Reweighting

Typical Causal Framework

Directly Confounder Balancing

Given a feature T

Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Sample reweighting can make a variable independent of other variables.
Global Balancing: Decorrelating Variables

Typical Causal Framework

**Global Balancing**

Given ANY feature T

Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Partial effect can be regarded as causal effect. Predicting with causal variables is stable across different environments.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. *KDD*, 2018.
Theoretical Guarantee

**Proposition 3.3.** If $0 < \hat{P}(X_i = x) < 1$ for all $x$, where $\hat{P}(X_i = x) = \frac{1}{n} \sum_i \mathbb{I}(X_i = x)$, there exists a solution $W^*$ satisfies equation (4) equals 0 and variables in $X$ are independent after balancing by $W^*$.

$$\sum_{j=1}^{p} \left\| \frac{X_T^{T,j} \left( W \odot X_{.,j} \right)}{W^T \cdot X_{.,j}} - \frac{X_T^{T,j} \left( W \odot (1-X_{.,j}) \right)}{W^T \cdot (1-X_{.,j})} \right\|_2^2, \quad (4)$$

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. *KDD*, 2018.
Causal Regularizer

Set feature $j$ as treatment variable

$$\sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|^2,$$

- **All features excluding treatment $j$**
- **Sample Weights**
- **Indicator of treatment status**

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. *ACM MM*, 2018.
Causally Regularized Logistic Regression

\[
\begin{align*}
\min & \quad \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (x_i \beta))) \\
\text{s.t.} & \quad \sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2 \leq \lambda_1, \\
& \quad W \geq 0, \quad \|W\|_2^2 \leq \lambda_2, \quad \|\beta\|_2^2 \leq \lambda_3, \quad \|\beta\|_1 \leq \lambda_4, \\
& \quad (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_5,
\end{align*}
\]

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. *ACM MM*, 2018.
From Shallow to Deep - DGBR

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. *KDD*, 2018.
Experiment 1 – non-i.i.d. image classification

- Source: **YFCC100M**
- Type: high-resolution and multi-tags
- Scale: 10-category, each with nearly 1000 images
- Method: select 5 *context tags* which are frequently co-occurred with the *major tag* (category label)
Experimental Result - insights
Experimental Result - insights
Experiment 2 – online advertising

- Environments generating:
  - Separate the whole dataset into 4 environments by users’ age, including $Age \in [20,30)$, $Age \in [30,40)$, $Age \in [40,50)$, and $Age \in [50,100)$.

Fig. 15: Prediction across environments separated by age. The models are trained on dataset where users’ $Age \in [20,30)$, but tested on various datasets with different users’ age range.

Fig. 16: $Average\_Error$ and $Stability\_Error$ of all algorithms across environments after fixing $P(Y)$ as the same with its value on global dataset.
From *Causal* problem to *Learning* problem

- Previous logic:

  ![Diagram showing the previous logic]

- More direct logic:

  ![Diagram showing the more direct logic]
Thinking from the **Learning** end

**Problem 1. (Stable Learning)**: Given the target $y$ and $p$ input variables $x = [x_1, \ldots, x_p] \in \mathbb{R}^p$, the task is to learn a predictive model which can achieve uniformly small error on any data point.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)
Stable Learning of Linear Models

• Consider the linear regression with misspecification bias

\[ y = x^\top \bar{\beta}_{1:p} + \bar{\beta}_0 + b(x) + \epsilon \]

Goes to infinity when perfect collinearity exists!

Bias term with bound \( b(x) \leq \delta \)

• By accurately estimating \( \bar{\beta} \) with the property that \( b(x) \) is uniformly small for all \( x \), we can achieve stable learning.

• However, the estimation error caused by misspecification term can be as bad as \( \| \hat{\beta} - \bar{\beta} \|_2 \leq 2(\delta / \gamma) + \delta \), where \( \gamma^2 \) is the smallest eigenvalue of centered covariance matrix.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)
Toy Example

• Assume the design matrix $X$ consists of two variables $X_1, X_2$, generated from a multivariate normal distribution:

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

• By changing $\rho$, we can simulate different extent of collinearity.
• To induce bias related to collinearity, we generate bias term $b(X)$ with $b(X) = Xv$, where $v$ is the eigenvector of centered covariance matrix corresponding to its smallest eigenvalue $\gamma^2$.
• The bias term is sensitive to collinearity.
Simulation Results

large variance in different distributions

large error (estimation bias)

increase collinearity

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)
Reducing collinearity by sample reweighting

Idea: Learn a new set of sample weights $w(x)$ to decorrelate the input variables and increase the smallest eigenvalue

- Weighted Least Square Estimation

$$\hat{\beta} = \arg\min_{\beta} \mathbb{E}_{(x) \sim D} w(x) \left( x^\top \beta_{1:p} + \beta_0 - y \right)^2$$

which is equivalent to

$$\hat{\beta} = \arg\min_{\beta} \mathbb{E}_{(x) \sim \tilde{D}} \left( x^\top \beta_{1:p} + \beta_0 - y \right)^2$$

So, how to find an “oracle” distribution $\tilde{D}$ which holds the desired property?

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)
Sample Reweighted Decorrelation Operator (cont.)

\[
\mathbf{X} = \begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1p} \\
  x_{21} & x_{22} & \cdots & x_{2p} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\]

\[
\tilde{\mathbf{X}} = \begin{pmatrix}
  x_{i1} & \cdots & x_{ri} & \cdots \\
  x_{j1} & \cdots & x_{sl} & \cdots \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{k1} & \cdots & x_{tl} & \cdots
\end{pmatrix}
\]

where \(i, j, k, r, s, t\) are drawn from \(1 \ldots n\) at random

- By treating the different columns independently while performing random resampling, we can obtain a column-decorrelated design matrix with the same marginal as before.
- Then we can use density ratio estimation to get \(w(x)\).
Experimental Results

• Simulation Study

(a) Estimation error

(b) Prediction error over different test environments

(c) Average prediction error & stability environments

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)
Experimental Results

- Regression
- Classification

(a) AUC over different test environments. (b) Average AUC of all the environments and stability.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)
Disentanglement Representation Learning

- Learning Multiple Levels of Abstraction
  - The big payoff of deep learning is to allow learning higher levels of abstraction
  - Higher-level abstractions disentangle the factor of variation, which allows much easier generalization and transfer

Disentanglement for Causality

• Causal / mechanism independence
  • Independently Controllable Factors (Thomas, Bengio et al., 2017)

A policy $\pi_k$ selectively change correspond to value

A representation $f_k$

$$sel(s, a, k) = \mathbb{E}_{s' \sim P^a_{ss'}} \left[ \frac{|f_k(s') - f_k(s)|}{\sum_{k'} |f_{k'}(s') - f_{k'}(s)|} \right]$$

• Optimize both $\pi_k$ and $f_k$ to minimize

$$\mathbb{E}_s \left[ \frac{1}{2} ||s - g(f(s))||^2_2 \right] - \lambda \sum_k \sum_a \pi_k(a|s) sel(s, a, k) \] + \mathcal{L}_{ace} the reconstruction error + \mathcal{L}_{sel} the disentanglement objective

Require subtle design on the policy set to guarantee causality.
Sectional Summary

- Causal inference provide valuable insights for stable learning
- Complete causal structure means data generation process, necessarily leading to stable prediction
- Stable learning can also help to advance causal inference
- Performance driven and practical applications

Benchmark is important!
Outline

- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Future Directions and Conclusions
Non-I.I.D. Image Classification

- Non I.I.D. Image Classification

\[ \psi(D_{\text{train}} = (X_{\text{train}}, Y_{\text{train}})) \neq \psi(D_{\text{test}} = (X_{\text{test}}, Y_{\text{test}})) \]

- Two tasks
  - Targeted Non-I.I.D. Image Classification
    - Have prior knowledge on testing data
    - e.g. transfer learning, domain adaptation
  - General Non-I.I.D. Image Classification
    - Testing is unknown, no prior
    - more practical & realistic
Existence of Non-I.I.Dness

• One metric (NI) for Non-I.I.Dness

**Definition 1 Non-I.I.D. Index (NI)** Given a feature extractor $g_{\phi}(\cdot)$ and a class $C$, the degree of **distribution shift** between training data $D_{\text{train}}^C$ and testing data $D_{\text{test}}^C$ is defined as:

$$NI(C) = \frac{\mathbb{E}_{x \sim D_{\text{train}}^C} [g_{\phi}(x)] - \mathbb{E}_{x \sim D_{\text{test}}^C} [g_{\phi}(x)]}{\sigma(g_{\phi}(D_{\text{train}}^C \cup D_{\text{test}}^C))},$$

**Distribution shift**

For normalization

• Existence of Non-I.I.Dness on Dataset consisted of 10 subclasses from ImageNet

• For each class
  • Training data
  • Testing data
  • CNN for prediction
Related Datasets

- DatasetA & DatasetB & DatasetC
  - NI is ubiquitous, but small on these datasets
  - NI is Uncontrollable, not friendly for Non I.I.D setting

A dataset for Non-I.I.D. image classification is demanded.
NICO - Non-I.I.D. Image Dataset with Contexts

- **NICO** Datasets:
- Object label: e.g. dog
- Contextual labels (Contexts)
  - the background or scene of a object, e.g. grass/water
- Structure of NICO

![Diagram showing the structure of NICO](image)

- 2 Superclasses per
- 10 Classes per
- 10 Contexts

Overlapping

Diverse & Meaningful
NICO - Non-I.I.D. Image Dataset with Contexts

- Data size of each class in NICO
  - Sample size: thousands for each class
  - Each superclass: 10,000 images
  - Sufficient for some basic neural networks (CNN)

- Samples with contexts in NICO
Controlling NI on NICO Dataset

- Minimum Bias (comparing with ImageNet)
- Proportional Bias (controllable)
  - Number of samples in each context
- Compositional Bias (controllable)
  - Number of contexts that observed
Minimum Bias

• In this setting, the way of random sampling leads to minimum distribution shift between training and testing distributions in dataset, which simulates a nearly i.i.d. scenario.

• 8000 samples for training and 2000 samples for testing in each superclass (ConvNet)

<table>
<thead>
<tr>
<th>Average NI</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>3.85</td>
</tr>
<tr>
<td>Vehicle</td>
<td>3.20</td>
</tr>
</tbody>
</table>

Average NI on ImageNet: 2.7

Images in NICO are with rich contextual information

Our NICO data is more Non-iid, more challenging for image classification
Proportional Bias

• Given a class, when sampling positive samples, we use all contexts for both training and testing, but the percentage of each context is different between training and testing dataset.

\[ \text{Dominant Ratio} = \frac{N_{\text{dominant}}}{N_{\text{minor}}} \]

We can control NI by varying dominate ratio
Compositional Bias

- Given a class, the observed contexts are different between training and testing data.

\[
Dominant Ratio = \frac{N_{dominant}}{N_{minor}}
\]

- **Moderate setting** (Overlap)
  - Training: 4.34
  - Testing: 4.44

- **Radical setting** (No Overlap & Dominant ratio)
  - Training: 4.0
  - Testing: 5:1
NICO - Non-I.I.D. Image Dataset with Contexts

• Large and controllable NI
NICO - Non-I.I.D. Image Dataset with Contexts

• The dataset can be downloaded from (temporary address):

  https://www.dropbox.com/sh/8mouawi5guaupyb/AAD4fdySrA6fn3PgsmhKwFgva?dl=0

• Please refer to the following paper for details:

Outline

- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Conclusions
Conclusions

• Predictive modeling is not only about Accuracy.
• **Stability** is critical for us to trust a predictive model.
• Causality has been demonstrated to be useful in stable prediction.
• How to marry causality with predictive modeling effectively and efficiently is still an open problem.
Conclusions

Debiasing

Causal Inference

Propensity Score

Direct Confounder Balancing

Global Balancing

Linear Stable Learning

Disentangled Learning

Stable Learning

Prediction
Reference


Reference

• Austin P C. An introduction to propensity score methods for reducing the effects of confounding in observational studies[J]. Multivariate behavioral research, 2011, 46(3): 399-424.
Reference

Acknowledgement

Kun Kuang
Tsinghua U

Zheyuan Shen
Tsinghua U

Hao Zou
Tsinghua U

Yue He
Tsinghua U

Bo Li
Tsinghua U

Susan Athey
Stanford U
Thanks!

Peng Cui
cuip@tsinghua.edu.cn
http://pengcui.thumedialab.com