Convergence of Uncertainty Sampling for Active Learning

Anant Raj [1,2] Francis Bach [1]

[1]Inria, Ecole Normale Supérieure
PSL Research University, Paris, France.
[2]Coordinated Science Laboratory,
University of Illinois, Urbana-Champaign.
anant.raj@inria.fr

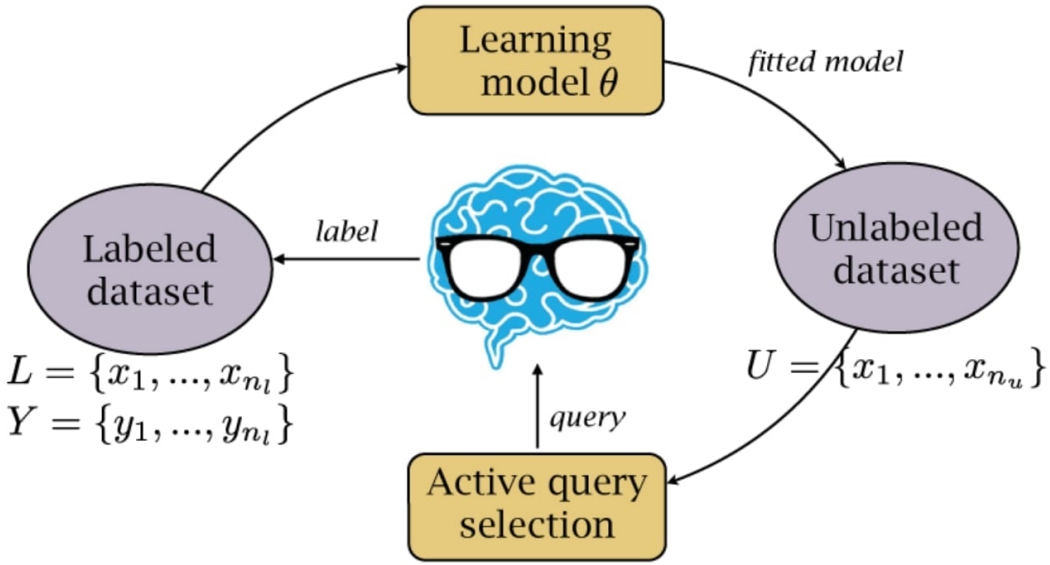
ICML 2022

Data in Machine Learning

- The premise of modern machine learning is based on the availability of large scale data.
- Data annotation is a time consuming and an expensive process.

Smart Data Annotation Technique

- Not all the labels are important to learn a machine learning model.
- Smarter choices (by an algorithm) can be made while data annotation.
- Active Learning: A learning algorithm that interactively queries a user (or some other information source) to label new data points with the desired outputs.



Active Learning

- **Membership Query Synthesis**: Learner generates its own instance from an underlying natural distribution.
- **Pool-Based Sampling**: Instances are drawn from the entire data pool and assigned a confidence score.
- **Stream-Based Selective Sampling**: In this work, we study the case of data streaming.

Active Learning Techniques

- Expected model change [Settles et al., 2007].
- Expected error reduction [Roy and McCallum, 2001].
- Expected variance reduction [Wang et al., 2015].
- Query by committee [Seung et al., 1992].
- etc..
- **Uncertainty sampling** [Balcan and Long, 2013, Wang and Singh, 2016, Cesa-Bianchi et al., 2009, Dekel et al., 2010, Orabona and Cesa-Bianchi, 2011, Cavallanti et al., 2011, Agarwal, 2013, Mussmann and Liang, 2018].
  - Hard to extended to multi-class classification and are computationally expensive.
  - Convergence under strong distributional assumption.
  - A proper convergence analysis is missing.

Max-Margin Classification

- **Binary Classification**

$$\ell(x, y, \theta) = \max\{0, 1 - y\theta^\top x\} \Rightarrow P(y\theta^\top x \leqslant 0) \leqslant \mathbb{E}(\ell(x, y, \theta)),$$

- **Multi-class Classification**

$$\ell(x, y, \theta) = \max\left[0, 1 - \theta^\top \left(\phi(x, y) - \phi(x, y^\star(\theta, x, y))\right)\right],$$

$$\text{where} \quad y^\star(\theta, x, y) = \arg\max_{z \in \mathcal{Y} \setminus y} \theta^\top \phi(x, z).$$

$$\Rightarrow P(\theta^\top \phi(x, y) - \theta^\top \phi(x, z^\star) \leqslant 0) \leqslant \mathbb{E}(\ell(x, y, \theta)).$$

Uncertainty Sampling for Max-Margin Classification

Require a sampling function σ, which maps prediction score to probability of querying a label.

- Compute probability $p_u(x_t, \theta_t) = \sigma(\theta_t, x_t)$.
- Sample Bernoulli random variable z_t with $p = p_u(x_t, \theta_t)$.
- Compute $\ell_t(x_t, y_t, \theta_t) \leftarrow \max(0, 1 - y_t(\theta_t^\top x_t))$.
- Update $\theta_{t+1} \leftarrow \theta_t + \gamma z_t (y_t x_t) \ell_t(x_t, y_t, \theta_t)$.

Computational cheap update per iteration ($O(d)$).
Same algorithm can be extended for the multiclass classification setting.

Theoretical Results

Major Task: Choose a nice sampling function σ whose properties can be exploited in showing the convergence.

Answer: We found one such candidate function for σ which can be written as,

$$\sigma(\theta, x) = \frac{1}{1 + \mu\theta^\top x}.$$

We state next results for above written candidate function.
We make no distributional assumption to show the convergence of the algorithm.

Separable Case

Convergence for Noiseless Case

Consider a set of n *i.i.d.* samples (x_i, y_i) from $\mathcal{P}$, and $y_i \in \{-1, 1\}$ for all $i = 1, \ldots, n$ then, if there exists a $\theta_\star$ for which $y(\theta_\star^\top x) \geq \rho^\star$, we have for some $C > 0$:

$$\mathbb{E}(1 - y\bar{\theta}_n^\top x)_+ \leq \frac{C \max\left\{1, \frac{1}{\mu}\right\}}{\min\left\{\frac{1}{\mu}, \frac{\rho^\star - 1}{1+\mu}\right\}^2 n}. \tag{1}$$

Expected # of Lables Queried

$$\#_n = \sum_{t=0}^{n-1} \sigma(\theta_t, x_t) = \sum_{t=0}^{n-1} \frac{1}{1 + \mu|\theta_t^\top x_t|}.$$

Inseparable Case

For any fixed $\theta^\star$,

$$P(y\theta_\star^\top x|(x,y) \leq \rho^\star) \leq \eta.$$

Convergence for Noisy Case

Consider a set of n *i.i.d.* samples (x_i, y_i), $y_i \in \{-1, 1\}$ for all $i = 1, \ldots, n$. Then under the above noise assumption for all (x, y) pair, we have:

- For small noise parameter η,
$$\mathbb{E}(1 - y\bar{\theta}_n^\top x)_+ = O\left(\frac{1}{n}\right).$$
- For large η,
$$\mathbb{E}(1 - y\bar{\theta}_n^\top x)_+ = O\left(\frac{1}{n} + \eta\right).$$

Look at the paper for multiclass results.

See you at the Poster!

Thank you!

References I

Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228, 2013.

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.

Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 83(1):71–102, 2011.

Nicolo Cesa-Bianchi, Claudio Gentile, and Francesco Orabona. Robust bounds for classification via selective sampling. In *Proceedings of the International Conference on Machine Learning*, pages 121–128, 2009.

References II

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Robust selective sampling from single and multiple teachers. In *COLT*, pages 346–358, 2010.

Stephen Mussmann and Percy S Liang. Uncertainty sampling is preconditioned stochastic gradient descent on zero-one loss. In *Advances in Neural Information Processing Systems*, pages 6955–6964, 2018.

Francesco Orabona and Nicolo Cesa-Bianchi. Better algorithms for selective sampling. In *International Conference on Machine Learning*, pages 433–440. Omnipress, 2011.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML*, 2:441–448, 2001.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in Neural Information Processing Systems*, 20: 1289–1296, 2007.

References III

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

Ran Wang, Chi-Yin Chow, and Sam Kwong. Ambiguity-based multiclass active learning. *IEEE Transactions on Fuzzy Systems*, 24(1):242–248, 2015.

Yining Wang and Aarti Singh. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *AAAI Conference on Artificial Intelligence*, 2016.