Sample and Computationally Efficient Active Learning

Maria-Florina Balcan
Carnegie Mellon University
Two Minute Version

Modern applications: **massive amounts** of raw data.

Only **a tiny fraction** can be annotated by human experts.

Protein sequences  
Billions of webpages  
Images

**Active Learning**: utilize data, minimize expert intervention.
Two Minute Version

**Active Learning**: technique for best utilizing data while minimizing need for human intervention.

**This talk**: the power of aggressive localization for label efficient, noise tolerant, poly time algo for learning linear separators [Awasthi-Balcan-Long JACM'17]

[Awasthi-Balcan-Haghtalab-Urner COLT'15] [Balcan-Long COLT'13]

- Much better noise tolerance than previously known for classic passive learning via poly time algos. [KKMS'05] [KLS'09]

- Solve an adaptive sequence of convex optimization pbs on smaller & smaller bands around current guess for target.
Passive and Active Learning
Supervised Learning

- E.g., which emails are spam and which are important.

- E.g., classify objects as chairs vs non chairs.
Statistical / PAC learning model

- **Data Source**: Distribution $D$ on $X$
- **Expert / Oracle**: $c^* : X \rightarrow \{0,1\}$
- **Labeled Examples**: $(x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m))$
- **Algorithm**: $h : X \rightarrow \{0,1\}$

- Algo sees $(x_1, c^*(x_1)), \ldots, (x_m, c^*(x_m))$, $x_i$ i.i.d. from $D$
- Does optimization over $S$, finds hypothesis $h \in C$.
- Goal: $h$ has small error, $\text{err}(h) = \Pr_{x \in D}(h(x) \neq c^*(x))$
- $c^*$ in $C$, realizable case; else agnostic
Two Main Aspects in Classic Machine Learning

**Algorithm Design. How to optimize?**

Automatically generate rules that do well on observed data.

Running time: \( \text{poly}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}\right) \)

**Generalization Guarantees, Sample Complexity**

Confidence for rule effectiveness on future data.

\[
0\left(\frac{1}{\epsilon}\left(\text{VCdim}(C) \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)
\]

\(C=\text{linear separators in } \mathbb{R}^d: 0\left(\frac{1}{\epsilon}\left(d \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)\)
Modern ML: New Learning Approaches

Modern applications: **massive amounts** of raw data.

Only a tiny fraction can be annotated by human experts.

- Protein sequences
- Billions of webpages
- Images
Active Learning

Data Source

Unlabeled examples

请求标签的示例

算法输出分类器

- 学习者可以选择特定的示例进行标注。
- **目标：**使用更少的标注示例。**[选择有信息的示例进行标注]**。
Active Learning in Practice

- **Text classification: active SVM** (Tong & Koller, ICML2000).
  - e.g., request label of the example closest to current separator.

- **Video Segmentation** (Fathi-Balcan-Ren-Regh, BMVC 11).
Can adaptive querying help? [CAL92, Dasgupta04]

- Threshold fns on the real line: \( h_w(x) = 1(x \geq w) \), \( C = \{h_w : w \in \mathbb{R}\} \)

Active Algorithm

- Get \( N = O(1/\epsilon) \) unlabeled examples
- How can we recover the correct labels with \( \ll N \) queries?
- Do binary search! Just need \( O(\log N) \) labels!

Passive supervised: \( \Omega(1/\epsilon) \) labels to find an \( \epsilon \)-accurate threshold.
Active: only \( O(\log 1/\epsilon) \) labels. Exponential improvement.
Active learning, provable guarantees

Lots of exciting results on sample complexity. E.g.,

• “Disagreement based” algorithms

Pick a few points at random from the current region of disagreement (uncertainty), query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

[BalcanBeygelzimerLangford'06, Hanneke07, DasguptaHsuMontleoni'07, Wang'09, Fridman'09, Koltchinskii10, BHW'08, BeygelzimerHsuLangfordZhang'10, Hsu'10, Ailon'12, ...]

Generic (any class), adversarial label noise.

• suboptimal in label complexity
• computationally prohibitive.
Poly Time, Noise Tolerant/Agnostic, Label Optimal AL Algos.
Margin Based Active Learning

Margin based algo for learning linear separators

- Realizable: exponential improvement, only $O(d \log \frac{1}{\epsilon})$ labels to find $w$ error $\epsilon$ when $D$ logconcave. [Balcan-Long COLT 2013]

- Agnostic & malicious noise: poly-time AL algo outputs $w$ with $\text{err}(w) = O(\eta)$, $\eta = \text{err}(\text{best lin. sep})$. [Awasthi-Balcan-Long JACM 2017]

- First poly time AL algo in noisy scenarios!

- Improves on noise tolerance of previous best passive [KKMS'05], [KLS'09] algos too!
Margin Based Active-Learning, Realizable Case

Draw $m_1$ unlabeled examples, label them, add them to $W(1)$.

Iterate $k = 2, ..., s$

- Find a hypothesis $w_{k-1}$ consistent with $W(k-1)$.
- $W(k) = W(k-1)$.
- Sample $m_k$ unlabeled samples $x$ satisfying $|w_{k-1} \cdot x| \leq \gamma_{k-1}$
- Label them and add them to $W(k)$.
Margin Based Active-Learning, Realizable Case

Log-concave distributions: log of density fnc concave.

- wide class: uniform distr. over any convex set, Gaussian, etc.

\[ f(\lambda x_1 + (1 - \lambda x_2)) \geq f(x_1)^\lambda f(x_2)^{1-\lambda} \]

**Theorem** D log-concave in \( \mathbb{R}^d \). If \( \gamma_k = O\left(\frac{1}{2^k}\right) \) then \( \text{err}(w_s) \leq \varepsilon \) after \( s = \log\left(\frac{1}{\varepsilon}\right) \) rounds using \( \tilde{O}(d) \) labels per round.

**Active learning**

- \( O(d \log \left(\frac{1}{\varepsilon}\right)) \) label requests
- \( \Theta\left(\frac{d}{\varepsilon}\right) \) unlabeled examples

**Passive learning**

- \( \Theta\left(\frac{d}{\varepsilon}\right) \) label requests
Analysis: Aggressive Localization

Induction: all $w$ consistent with $W(k)$, $err(w) \leq 1/2^k$
Analysis: Aggressive Localization

Induction: all $w$ consistent with $W(k)$, $\text{err}(w) \leq 1/2^k$

Suboptimal
Analysis: Aggressive Localization

Induction: all \( w \) consistent with \( W(k) \), \( \text{err}(w) \leq 1/2^k \)

\[
\text{err}(w) = \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1}) + \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq 1/2^{k+1}
\]
Analysis: Aggressive Localization

Induction: all $w$ consistent with $W(k)$, $\text{err}(w) \leq 1/2^k$

\[
\text{err}(w) = \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1}) + \Pr(w \text{ errs on } x | |w_{k-1} \cdot x| \leq \gamma_{k-1}) \Pr(|w_{k-1} \cdot x| \leq \gamma_{k-1})
\]

Enough to ensure $\Pr(w \text{ errs on } x | |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C$

Need only $m_k = \tilde{O}(d)$ labels in round $k$.

Key point: localize aggressively, while maintaining correctness.
Margin Based Active-Learning, Agnostic Case

Draw $m_1$ unlabeled examples, label them, add them to $W$.

Iterate $k=2, ..., s$

- Find $w_{k-1}$ in $B(w_{k-1}, r_{k-1})$ of small $\tau_{k-1}$ hinge loss wrt $W$.
  - Clear working set.
  - Sample $m_k$ unlabeled samples $x$ satisfying $|w_{k-1} \cdot x| \leq \gamma_{k-1}$;
  - Label them and add them to $W$.

End iterate

Analysis, key idea:

- Pick $\tau_k \approx \gamma_k$
- Localization & variance analysis control the gap between hinge loss and 0/1 loss (only a constant).
**Implements over Passive Learning too!**

<table>
<thead>
<tr>
<th></th>
<th>Passive Learning</th>
<th>Prior Work</th>
<th>Our Work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Malicious</strong></td>
<td></td>
<td>( \text{err}(w) = O(\eta d^{1/4}) )(^{[\text{KKMS'05}]} )</td>
<td>( \text{err}(w) = O(\eta) ) Info theoretic optimal (^{[\text{Awasthi-Balcan-Long'17}] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \text{err}(w) = O(\sqrt{\eta \log(d/\eta)}) ) (^{[\text{KLS'09}] )</td>
<td></td>
</tr>
<tr>
<td><strong>Agnostic</strong></td>
<td></td>
<td>( \text{err}(w) = O(\eta \sqrt{\log(1/\eta)}) ) (^{[\text{KKMS'05}] )</td>
<td>( \text{err}(w) = O(\eta) ) (^{[\text{Awasthi-Balcan-Long'17}] )</td>
</tr>
<tr>
<td><strong>Bounded Noise</strong></td>
<td></td>
<td>( \text{NA} )</td>
<td>( \eta + \epsilon ) (^{[\text{Awasthi-Balcan-Haghtalab-Urner'15}] )</td>
</tr>
<tr>
<td><strong>Active Learning</strong></td>
<td>[agnostic/malicious/bounded]</td>
<td>( \text{NA} )</td>
<td>same as above! Info theoretic optimal (^{[\text{Awasthi-Balcan-Long'14}] )</td>
</tr>
</tbody>
</table>

Slightly better results for the uniform distribution case.
Localization both algorithmic and analysis tool!

Useful for active and passive learning!
Discussion, Open Directions

- Active learning: important modern learning paradigm.
- First poly time, label efficient AL algo for agnostic learning in high dimensional cases.
- Also leads to much better noise tolerant algos for passive learning of linear separators!

Open Directions

- More general distributions, other concept spaces.
- Exploit localization insights in other settings (e.g., online convex optimization with adversarial noise).