The power of localization for learning with noise

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

This talk: label efficient, noise tolerant, poly time algo for learning linear separators
[Balcan-Long COLT’13] [Awasthi-Balcan-Long STOC’14]

• Our algo operates in an online selective sampling model.

• Much better noise tolerance than previously known for classic passive learning via poly time algos.
Passive and Active Learning
Supervised Learning

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

  Not spam

  spam

- E.g., classify objects as chairs vs non chairs.

  Not chair

  chair
Statistical / PAC learning model

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

- **Goal**: $h$ has small error, $\text{err}(h) = \Pr_{x \in D}(h(x) \neq c^*(x))$

- **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 $h \in C$.

- **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.
E.g., Boosting, SVM, etc.

Generalization Guarantees, Sample Complexity
Confidence for rule effectiveness on future data.

$$O\left(\frac{1}{\epsilon} (VC\text{dim}(C) \log \left(\frac{1}{\epsilon}\right) + \log \left(\frac{1}{\delta}\right))\right)$$
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

- Learner can choose specific examples to be labeled.
- **Goal**: use fewer labeled examples [pick informative examples to be labeled].

- **Selective sampling AL**: stream of unlabeled examples, when each arrives make a decision to ask for label or not.
Active learning, provable guarantees

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

- DasguptaKalaiMonteleoni’05, CastroNowak’07, CavallantiCesa-BianchiGentile’10

- “Disagreement based” algs [query pts from current region of disagreement, throw out hypotheses when 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 complex & 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 $O(d \log 1/\epsilon)$ labels to find $w$ error $\epsilon$ when $D$ logconcave. [Balcan-Long COLT 2013]

- Agnostic & malicious noise: poly-time AL algo outputs $w$ with $\text{err}(w) = O(\eta)$, $\eta = \text{err( best lin. sep)}$. [Awasthi-Balcan-Long STOC 2014]

  - First poly time AL algo in noisy scenarios!
  - First for malicious noise [Val85] (features corrupted too).

- Improves on noise tolerance of previous best passive [KKMS'05], [KLS'09] algos too!
Draw $m_1$ unlabeled examples, label them, add them to $W(1)$.

iterate $k = 2, \ldots, s$

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

Passive learning
\[ \Theta\left(\frac{d}{\varepsilon}\right) \] label requests
Linear Separators, Log-Concave Distributions

Fact 1

\[ d(u, v) \approx \frac{\theta(u, v)}{\pi} \]

Proof idea:
- project the region of disagreement in the space given by \( u \) and \( v \)
- use properties of log-concave distributions in 2 dimensions.

Fact 2

\[ \Pr_x [ |v \cdot x| \leq \gamma ] \leq \gamma. \]
Linear Separators, Log-Concave Distributions

**Fact 3** If $\theta(u, v) = \beta$ and $\gamma = C\beta$

$$\Pr_x[(u \cdot x)(v \cdot x) \leq 0, |v \cdot x| \geq \gamma] \leq \frac{\beta}{4}.$$
Margin Based Active-Learning, Realizable Case

**Proof Idea**

Induction: all $w$ consistent with $W(k)$ have error $\leq 1/2^k$; so, $w_k$ has error $\leq 1/2^k$.

For $\gamma_k = O \left( \frac{c}{2^k} \right)$

$$\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1}) + \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq 1/2^{k+1}$$

iterate $k=2, \ldots, s$

- 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 $W(k)$. 

$$w_k - 1 \overline{\gamma_k} - 1 w^*$$
Proof Idea

Under logconcave distr. for $\gamma_k = \Theta \left( \frac{c}{2^k} \right)$

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

\[ \leq 1/2^{k+1} \]
Proof Idea

Under logconcave distr. for $\gamma_k = O\left(\frac{c}{2^k}\right)$

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

Enough to ensure 

$$\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C_1$$ 

Can do with only 

$$m_k = O(d + \log \log(1/\epsilon))$$ labels.
Margin Based Active-Learning, Agnostic Case

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

Iterate $k=2, \ldots, s$

- 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
Margin Based Active-Learning, Agnostic Case

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

iterate $k=2, \ldots, s$

• find $w_{k-1}$ in $B(w_{k-1}, r_{k-1})$ of small $	au_{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

Localization in concept space.

Localization in instance space.
Analysis: the Agnostic Case

\begin{align*}
\text{Theorem} & \quad D \text{ log-concave in } \mathbb{R}^d. \quad \eta = \Omega(\epsilon / \log^2(1/\epsilon)) \\
\text{If} & \quad \gamma_k = O\left(\frac{c}{2^k}\right), \quad \tau_k = O\left(\frac{c}{2^k}\right), \quad r_k = O\left(\frac{c}{2^k}\right), \quad s = \log\left(\frac{1}{\epsilon}\right), \quad \text{err}(w_s) \leq \epsilon.
\end{align*}

Key ideas:

- As before need \( \Pr(w \text{ errs on } x \mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C \)

- For \( w \) in \( B(w_{k-1}, r_{k-1}) \) we have \( l(w, x) \)

  \[
  \text{err}(w_k) \leq \mathbb{E}[l(w_k, x, y)] \leq \tau_k / \gamma_k + \eta 2^k \sqrt{d} \leq C
  \]

- sufficient to set \( \eta \leq \epsilon / \sqrt{d} \)

- Careful variance analysis leads \( \eta = \Omega(\epsilon / \log^2(1/\epsilon)) \)
Analysis: Malicious Noise

**Theorem** \( D \) log-concave in \( \mathbb{R}^d \). \( \eta = \Omega(\epsilon / \log^2(1/\epsilon)) \)

If \( \gamma_k = O\left(\frac{c}{2^k}\right), \tau_k = O\left(\frac{c}{2^k}\right), r_k = O\left(\frac{c}{2^k}\right), s = \log\left(\frac{1}{\epsilon}\right), \text{err}(w_s) \leq \epsilon. \)

The adversary can corrupt both the label and the feature part.

**Key ideas:**

- As before need \( \Pr(w \text{ errs on } x \mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C \)

- Soft localized outlier removal and careful variance analysis.
Improves over Passive Learning too!

<table>
<thead>
<tr>
<th>Passive Learning</th>
<th>Prior Work</th>
<th>Our Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malicious</td>
<td>$\eta = \Omega\left(\frac{\varepsilon^3}{\log^2(d/\varepsilon)}\right)$ [KLS'09]</td>
<td>$\eta = \Omega(\varepsilon/\log^2(1/\varepsilon))$</td>
</tr>
<tr>
<td>Agnostic</td>
<td>$\eta = \Omega\left(\frac{\varepsilon^3}{\log(1/\varepsilon)}\right)$ [KLS'09]</td>
<td>$\eta = \Omega(\varepsilon/\log^2(1/\varepsilon))$</td>
</tr>
<tr>
<td>Active Learning [agnostic/malicious]</td>
<td>NA</td>
<td>$\eta = \Omega(\varepsilon/\log^2(1/\varepsilon))$</td>
</tr>
</tbody>
</table>
## Improves over Passive Learning too!

<table>
<thead>
<tr>
<th>Passive Learning</th>
<th>Prior Work</th>
<th>Our Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malicious</td>
<td>$\eta = \Omega \left( \frac{\epsilon}{d^{1/4}} \right)$ [KKMS’05]</td>
<td>$\eta = \Omega(\epsilon)$</td>
</tr>
<tr>
<td></td>
<td>$\eta = \Omega \left( \frac{\epsilon^2}{\log(d/\epsilon)} \right)$ [KLS’09]</td>
<td>Info theoretic optimal</td>
</tr>
<tr>
<td>Agnostic</td>
<td>$\eta = \Omega \left( \frac{\epsilon}{\sqrt{\log(1/\epsilon)}} \right)$ [KKMS’05]</td>
<td>$\eta = \Omega(\epsilon)$</td>
</tr>
<tr>
<td>Active Learning [agnostic/malicious]</td>
<td>NA</td>
<td>$\eta = \Omega(\epsilon)$</td>
</tr>
</tbody>
</table>

Info theoretic optimal

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

Useful for active and passive learning!
Discussion, Open Directions

• First poly time, label efficient AL algo for agnostic learning in high dimensional cases.

• Better noise tolerant algos for passive learning of linear separators.

• A PTAS for uniform, agnostic, Daniely'14.

Open Directions

• More general distributions, other concept spaces.

• Exploit localization insights in other settings (e.g., online convex optimization with adversarial noise).