Active Learning of Linear Separators

Maria-Florina Balcan

11/23/2015
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

Request for the Label of an Example

A Label for that Example

Request for the Label of an Example

A Label for that Example

Algorithm outputs a classifier

- Learner can choose specific examples to be labeled.
- Goal: use fewer labeled examples [pick informative examples to be labeled].
Active learning, provable guarantees

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

- DasguptaKalaiMonteleoni’05, CastroNowak’07, CavallantiCesa-BianchiGentile’10
- “Disagreement based” algos [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 $O(d \log \frac{1}{\varepsilon})$ labels to find $w$ error $\varepsilon$ 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 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] algs 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 $W(k)$. 

![Diagram of margin based active learning, realizable case with lines and points illustrating the process.](image-url)
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 \( \gamma_k = \mathcal{O}\left(\frac{1}{2^k}\right) \) 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

\( \mathcal{O}\left(d \log \left(\frac{1}{\varepsilon}\right)\right) \) 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
\[ \Pr_x [|v \cdot x| \leq \gamma] \leq \gamma. \]
Fact 3: If $\theta(u, v) = \beta$ and $\gamma = C\beta$

$$\Pr_{x}[(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma] \leq \frac{\beta}{4}.$$
**Fact 3** If $\theta(u, v) = \beta$ and $\gamma = C\beta$

$$\Pr_x [(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma] \leq \frac{\beta}{4}.$$  

**Proof idea:**

- project the region of disagreement in the space given by $u$ and $v$
- Note that each $x$ in $E$ has $||x|| \geq \gamma/\beta = c_2$

$$\Pr_x [x \in E] = \sum_{i=1}^{\infty} \Pr [E \cap (B((i + 1)c_2) - B(ic_2))] \leq C\beta(i + 1)^2 \exp[-Ci]$$
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 \( \gamma_k = O\left(\frac{c}{2^k}\right) \)

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

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

\[
\text{err}(w) = \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1}) + \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq \frac{1}{2^{k+1}}
\]
Proof Idea

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

\[
\begin{align*}
\text{err}(w) &= \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1}) + \\
&\quad \Pr(w \text{ errs on } x \mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \Pr(|w_{k-1} \cdot x| \leq \gamma_{k-1}) \\
&\quad \leq C \gamma_{k-1}.
\end{align*}
\]

Enough to ensure

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

Can do with only $m_k = O\left(d + \log \log(1/\epsilon)\right)$ labels.
Margin Based Active-Learning, Agnostic Case

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

**iterate** $k=2, \ldots, s$

- find $w_k$ 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.

Margin Based AL, Summary

• Extensions to nearly log-concave distributions, noisy settings. Matching Lower Bounds.

• General class of pbs for which AL provides exponential improvement in $1/\varepsilon$ (without additional increase on d).

• Can be made differentially private!

• Cool implications to passive learning, distributed learning.

• Zhang-Chaudhuri’ NIPS14 extensions to general concept spaces --- not computationally efficient.