Noise-adaptive Margin-based Active Learning, and Lower Bounds

Yining Wang, Aarti Singh
Carnegie Mellon University
Machine Learning: the setup

- The machine learning problem
  - Each data point \((x_i, y_i)\) consists of data \(x_i\) and label \(y_i\)
  - Access to training data \((x_1, y_1), \cdots, (x_n, y_n)\)
  - Goal: train classifier \(\hat{f}\) to predict \(y\) based on \(x\)
  - Example: Classification
    \[
    x_i \in \mathbb{R}^d, \quad y_i \in \{+1, -1\}
    \]
Machine learning: passive vs. active

- Classical framework: passive learning
  - \textit{I.I.D.} training data \((x_i, y_i) \text{i.i.d. } D\)
  - Evaluation: \textit{generalization error} \(\Pr_D \left[ y \neq \hat{f}(x) \right]\)
- An active learning framework
  - Data are cheap, but labels are expensive!
  - \textit{Example}: medical data (labels require domain knowledge)
  - Active learning: minimize \textit{label requests}
Active Learning

- Pool-based active learning

- The learner $A$ has access to unlabeled data stream $x_1, x_2, \ldots \sim \text{i.i.d. } D$

- For each $x_i$, the learner decides whether to query; if label requested, $A$ obtains $y_i$

- Minimize number of requests, while scanning through polynomial number of unlabeled data.
Active Learning

❖ Example: learning homogeneous linear classifier

\[
y_i = \text{sgn}(w^\top x_i) + \text{noise}
\]

❖ Basic (passive) approach: empirical risk minimization (ERM)

\[
\hat{w} \in \arg\min_{\|w\|_2=1} \sum_{i=1}^{n} I[y_i \neq \text{sgn}(w^\top x_i)]
\]

❖ How about active learning?
Margin-based Active Learning

BALCAN, BRODER and ZHANG, COLT’07

- Data dimension $d$, query budget $T$, no. of iterations $E$
- At each iteration $k \in \{1, \cdots, E\}$
  - Determine parameters $b_{k-1}, \beta_{k-1}$
  - Find $n = T/E$ samples in $\{x \in \mathbb{R}^d : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$
  - Constrained ERM: $\hat{w}_k = \min_{\theta(w, \hat{w}_{k-1}) \leq \beta_{k-1}} L(\{x_i, y_i\}_{i=1}^n; w)$
- Final output: $\hat{w}_E$
Tsybakov Noise Condition

- There exist constants $\mu > 0, \alpha \in (0, 1)$ such that
  \[ \mu \cdot \theta(w, w^*)^{1/(1-\alpha)} \leq \text{err}(w) - \text{err}(w^*) \]

- $\alpha \in (0, 1)$: key noise magnitude parameter in TNC

- Which one is harder?
Margin-based Active Learning

- Main Theorem [BBZ07]: when \( D \) is the uniform distribution, the margin-based algorithm achieves

\[
\text{err}(\hat{w}) - \text{err}(w^*) = \tilde{O}_P \left\{ \frac{d}{T} \right\}^{1/2\alpha}.
\]

Passive Learning:

\[
O((d/T)^{\frac{1-\alpha}{2\alpha}}).
\]
Proof outline

❖ At each iteration $k$, perform restricted ERM over within-margin data

\[ \hat{w}_k = \arg\min_{\theta(w, \hat{w}_{k-1}) \leq \beta_{k-1}} \hat{\text{err}}(w|S_1), \]

\[ S_1 = \{ x : |x^\top \hat{w}_{k-1}| \leq b_{k-1} \} \]
Proof outline

❖ Key fact: if \( \theta(\hat{w}_{k-1}, w^*) \leq \beta_{k-1} \) and \( b_k = \tilde{\Theta}(\beta_k / \sqrt{d}) \) then
\[
err(\hat{w}_k) - err(w^*) = \tilde{O} \left( \beta_{k-1} \sqrt{d/T} \right)
\]
❖ Proof idea: decompose the excess error into two terms
\[
\left[ err(\hat{w}_k \mid S_1) - err(w^* \mid S_1) \right] Pr[x \in S_1] = \tilde{O}(\sqrt{d/T})
\]
\[
\left[ err(\hat{w}_k \mid S_1^c) - err(w^* \mid S_1^c) \right] Pr[x \in S_1^c] = \tilde{O}(b_{k-1} \sqrt{d})
\]
❖ Must ensure \( w^* \) is always within reach!
\[
\beta_k = 2^{\alpha - 1} \beta_{k-1}
\]
Problem

- What if $\alpha$ is not known? How to set key parameters $b_k, \beta_k$

- If the true parameter is $\alpha$ but the algorithm is run with $\alpha' > \alpha$

- The convergence is $\alpha'$ instead of $\alpha$!
Noise-adaptive Algorithm

- Agnostic parameter settings
  \[ E = \frac{1}{2} \log T, \beta_k = 2^{-k} \pi, b_k = \frac{2\beta_k}{\sqrt{d}} \sqrt{2E} \]
- Main analysis: two-phase behaviors
  - "Tipping point": \( k^* \in \{1, \cdots, E\} \), depending on \( \alpha \)
  - Phase I: \( k \leq k^* \), we have that \( \theta(\hat{w}_k, w^*) \leq \beta_k \)
  - Phase II: \( k > k^* \), we have that
    \[ \text{err}(\hat{w}_{k+1}) - \text{err}(\hat{w}_k) \leq \beta_k \cdot \tilde{O}(\sqrt{d/T}) \]
Noise-Adaptive Analysis

- Main theorem: for all $\alpha \in (0, 1/2)$
  \[ \text{err}(\hat{w}) - \text{err}(w^*) = \tilde{O}_P \left\{ \left( \frac{d}{T} \right)^{1/2\alpha} \right\}. \]

- Matching the upper bound in [BBZ07]

- ... and also a lower bound (this paper)
Lower Bound

- Is there any active learning algorithm that can do better than the $\tilde{O}_P((d/T)^{1/2\alpha})$ sample complexity?
- In general, no [Henneke, 2015]. But the data distribution $D$ is quite contrived in the negative example.
- We show that $\tilde{O}_P((d/T)^{1/2\alpha})$ is tight even if $D$ is as simple as the uniform distribution over unit sphere.
Lower Bound

- The “Membership Query Synthesis” (QS) setting
  - The algorithm $A$ picks an arbitrary data point $x_i$
  - The algorithm receives its label $y_i$
  - Repeat the procedure $T$ times, with $T$ the budget
- QS is more powerful than pool-based setting when $D$ has density bounded away from below.
- We prove lower bounds for the QS setting, which implies lower bounds in the pool-based setting.
Tsybakov’s Main Theorem

TSYBAKOV and ZAIATS, *Introduction to Nonparametric Estimation*

- Let $\mathcal{F}_0 = \{f_0, \cdots, f_M\}$ be a set of models. Suppose
  - **Separation:** $D(f_j, f_k) \geq 2\rho$, $\forall j, k \in \{1, \cdots, M\}, j \neq k$
  - **Closeness:** $\frac{1}{M} \sum_{j=1}^{M} \text{KL}(P_{f_j} \parallel P_{f_0}) \leq \gamma \log M$
  - **Regularity:** $P_{f_j} \ll P_{f_0}$, $\forall j \in \{1, \cdots, M\}$

- Then the following bound holds

$$
\inf_{\hat{f}} \sup_{f \in \mathcal{F}_0} \Pr \left[ D(\hat{f}, f) \geq \rho \right] \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left( 1 - 2\gamma - 2\sqrt{\frac{\gamma}{\log M}} \right).
$$
Negative Example Construction

• **Separation:** $D(f_j, f_k) \geq 2\rho, \forall j, k \in \{1, \cdots, M\}, j \neq k$

• Find hypothesis class $\mathcal{W} = \{w_1, \cdots, w_m\}$ such that

\[ t \leq \theta(w_i, w_j) \leq 6.5t, \quad \forall i \neq j \]

• … can be done for all $t \in (0, 1/4)$, using constant weight coding

• … can guarantee that $\log |\mathcal{W}| = \Omega(d)$
Negative Example Construction
Negative Example Construction

- **Closeness:** 
  \[
  \frac{1}{M} \sum_{j=1}^{M} \text{KL}(P_{f,j} \| P_{f_0}) \leq \gamma \log M
  \]

  \[
  \text{KL}(P_{i,T} \| P_{j,T}) = \mathbb{E}_i \left[ \log \frac{P_{X,Y,T}^{(i)}(x_1, y_1, \cdots, x_T, y_T)}{P_{X,Y,T}^{(j)}(x_1, y_1, \cdots, x_T, y_T)} \right]
  \]

  \[
  = \mathbb{E}_i \left[ \log \frac{\prod_{t=1}^{T} P_{X,Y,T}^{(i)}(y_t | x_t) P_{X,Y,T}^{(j)}(X_1, Y_1, \cdots, X_{t-1}, Y_{t-1})}{\prod_{t=1}^{T} P_{X,Y,T}^{(j)}(y_t | x_t) P_{X,Y,T}^{(i)}(X_1, Y_1, \cdots, X_{t-1}, Y_{t-1})} \right]
  \]

  \[
  = \mathbb{E}_i \left[ \log \frac{\prod_{t=1}^{T} P_{Y|X}^{(i)}(y_t | x_t)}{\prod_{t=1}^{T} P_{Y|X}^{(j)}(y_t | x_t)} \right]
  \]

  \[
  = \sum_{t=1}^{T} \mathbb{E}_i \left[ \mathbb{E}_i \left[ \log \frac{P_{Y|X}^{(i)}(y_t | x_t)}{P_{Y|X}^{(j)}(y_t | x_t)} \middle| X_1 = x_1, \cdots, X_T = x_T \right] \right]
  \]

  \[
  \leq T \cdot \sup KL(P_{Y|X}^{(i)}(\cdot|x) \| P_{Y|X}^{(j)}(\cdot|x)).
  \]
Lower Bound

TSYBAKOV and ZAIATS, *Introduction to Nonparametric Estimation*

- Let $\mathcal{F}_0 = \{f_0, \cdots, f_M\}$ be a set of models. Suppose
  - **Separation:** $D(f_j, f_k) \geq 2\rho, \forall j, k \in \{1, \cdots, M\}, j \neq k$
  - **Closeness:** $\frac{1}{M} \sum_{j=1}^{M} \text{KL}(P_{f_j} \| P_{f_0}) \leq \gamma \log M$
  - **Regularity:** $P_{f_j} \ll P_{f_0}, \forall j \in \{1, \cdots, M\}$
  - Take $\rho = \Theta(t) = \Theta((d/T)^{(1-\alpha)/2\alpha})$, $\log M = \Theta(d)$
  - We have that
    $$\inf \sup \Pr \left[ \theta(\hat{w}, w^*) \geq \frac{t}{2} \right] = \Omega(1)$$
Lower Bound

- Suppose \( D \) has density bounded away from below and fix \( \mu > 0, \alpha \in (0, 1) \). Let \( \mathcal{P}_{Y|X} \) be class of distributions satisfying \( (\mu, \alpha) \)-TNC. Then we have that

\[
\inf_{\mathcal{A}} \sup_{P \in \mathcal{P}_{Y|X}} \mathbb{E}_P [\text{err}(\hat{w}) - \text{err}(w^*)] \geq \Omega \left[ \left( \frac{d}{T} \right)^{1/2\alpha} \right].
\]
Extension: “Proactive” learning

- Suppose there are $m$ different users (labelers) who share the same classifier $w^*$ but with different TNC parameters $\alpha_1, \cdots, \alpha_m$.
- The TNC parameters are not known.
- At each iteration, the algorithm picks a data point $x$ and also a user $j$, and observes $f(x;j)$.
- The goal is to estimate the Bayes classifier $w^*$. 
Extension: “Proactive” learning

- Algorithm framework:
  - Operate in $E = O(\log T)$ iterations.
  - At each iteration, use conventional Bandit algorithms to address exploration-exploitation tradeoff.

- Key property: search space $\{\beta_k\}$ and margin $\{b_k\}$ does not depend on unknown TNC parameters.

- Many interesting extensions: what if multiple labelers can be involved each time?
Thanks! Questions?