Advanced Machine Learning

Active Learning

MEHRYAR MOHRI
MOHRI@
COURANT INSTITUTE & GOOGLE RESEARCH
Active Learning Setup

- Passive learning:
  - IID sample \( ((x_1, y_1), \ldots, (x_m, y_m)) \sim D^m \) is drawn.
  - learner receives full labeled sample.

- Active learning:
  - IID sample \( ((x_1, y_1), \ldots, (x_m, y_m)) \sim D^m \) is drawn.
  - learner has access to \((x_1, \ldots, x_m)\).
  - learner can request the label \( y_i \) of point \( x_i \).
  - objective: fewer label requests than in passive learning.
Key Active Learning Problem

Tension:

- requesting label of new point to gain more information.
- sample bias induced by the label queries.
Favorable Example

- Binary classification problem in $\mathbb{R}$:
  - $H$: threshold functions.
  - data assumed separable.

- Sample complexity for determining $\theta^*$ within $\epsilon$:
  - supervised learner needs $O\left(\frac{1}{\epsilon}\right)$ samples since at least one point is needed in $[\theta^* - \epsilon, \theta^* + \epsilon]$.
  - active learner needs only $O\left(\log\frac{1}{\epsilon}\right)$ using binary search.

---

exponential improvement!
Negative Result

Non-realizable case:

- stochastic or deterministic labels.
- if Bayes error is $\beta > 0$, the sample complexity of any active learning algorithm is at least
  $$\Omega \left( \frac{\beta^2}{\epsilon^2} \right).$$
- thus, lower bound matches passive learning upper bound $O \left( \frac{1}{\epsilon^2} \right)$. 

(Kääriäinen, 2006)
CAL Algorithm

Assume realizable case with hypothesis set $H$.

CAL($H$)

1. $H_1 \leftarrow H$
2. for $t \leftarrow 1$ to $T$ do
3.  if $(\exists h, h' \in H_t : h(x_t) \neq h'(x_t))$ then
4.    $y_t \leftarrow \text{QUERYLABEL}(x_t)$
5.    $H_{t+1} \leftarrow \{h \in H_t : h(x_t) = y_t\}$
6.  else $H_{t+1} \leftarrow H_t$
7. return $H_{T+1}$
CAL Algorithm

Simple algorithm, but:
- Computational cost of maintaining $H_t$s.
- Separability requirement.
Definitions

- Region of disagreement:
  \[
  \text{DIS}(H) = \{ x \in X \mid \exists h, h' \in H : h(x) \neq h'(x) \}. 
  \]

- Disagreement metric:
  \[
  d(h, h') = \Pr_{x \sim D}[h(x) \neq h'(x)].
  \]

- Disagreement ball:
  \[
  B(h, r) = \left\{ h' \in H : d(h, h') \leq r \right\}.
  \]

- Disagreement coefficient (rate of disagreement decrease):
  \[
  \theta = \limsup_{r \to 0} \frac{\text{DIS}(B(h^*, r))}{r}.
  \]

(Hanneke, 2009)
Disagreement Coefficient

Property: for all \( r > 0 \), \( \text{DIS}(B(h^*, r)) \leq \theta r \).

Examples:

- threshold functions: \( \theta \leq 2 \).
  
  - let \( t, t' \in B(t^*, r) \), then, \( t, t' \in [t^* - \epsilon, t^* + \epsilon'] \) where
    \[
    \epsilon = \underset{\epsilon > 0}{\text{argmax}} \{ \Pr([t^* - \epsilon, t^*]) \leq r \} \quad \epsilon' = \underset{\epsilon > 0}{\text{argmax}} \{ \Pr([t^*, t^* + \epsilon]) \leq r \}.
    
    - thus, \( \text{DIS}(B(h^*, r)) \leq 2r \).
  
  - finite hypothesis sets: \( \theta \leq |H| \).

- linear separators going through the origin and uniform distribution: \( \theta \leq \pi \sqrt{N} \).
CAL Guarantees

Theorem: let $H$ be a hypothesis set with $\text{VCdim}(H) = d$ and assume that the data is separable with disagreement coefficient $\theta$. Then, the label complexity of CAL is bounded by

$$\tilde{O}(\theta d \log \frac{1}{\epsilon}).$$
DHM Algorithm

(According to Dasgupta, Hsu, and Monteleoni, 2007)

\[ \mathcal{A}(S, T) \text{ returns hypothesis in } H \text{ consistent with } S \text{ with minimum error on } T \text{ when it exists, } \text{NIL otherwise.} \]

DHM((x_1, \ldots, x_T))

1. \( S \leftarrow \emptyset \) \quad \triangledown \text{ labels inferred}
2. \( T \leftarrow \emptyset \) \quad \triangledown \text{ labels queried}
3. \textbf{for} \ t \leftarrow 1 \textbf{ to } T \textbf{ do}
4. \quad h_+ \leftarrow A(S \cup (x_t, +1), T)
5. \quad h_- \leftarrow A(S \cup (x_t, -1), T)
6. \quad \textbf{if} \ (h_+ = \text{NIL}) \textbf{ then}
7. \quad \quad S \leftarrow S \cup \{(x_t, -1)\}
8. \quad \textbf{elseif} \ (h_- = \text{NIL}) \textbf{ then}
9. \quad \quad S \leftarrow S \cup \{(x_t, +1)\}
10. \quad \textbf{elseif} \ \hat{R}_{S\cup T}(h_+) - \hat{R}_{S\cup T}(h_-) > \Delta_t \textbf{ then}
11. \quad \quad S \leftarrow S \cup \{(x_t, -1)\}
12. \quad \textbf{elseif} \ \hat{R}_{S\cup T}(h_-) - \hat{R}_{S\cup T}(h_+) > \Delta_t \textbf{ then}
13. \quad \quad S \leftarrow S \cup \{(x_t, +1)\}
14. \quad \textbf{else} \ y_t \leftarrow \text{QUERYLABEL}(x_t)
15. \quad \quad T \leftarrow T \cup \{(x_t, y_t)\}
16. \quad \textbf{return} \ H_{T+1}
$S \cup T$ not an i.i.d. sample drawn according to $D$.

$\Delta_t$ is defined by $\Delta_t = \beta_t^2 + \beta_t \left( \sqrt{\hat{R}_t(h_+)} + \sqrt{\hat{R}_t(h_-)} \right)$,

with $\beta_t = 2 \sqrt{\frac{\log \left((8t^2 + t)\Pi_{2t}(H)\right) + \log \frac{1}{\delta}}{t}} = \tilde{O} \left( \sqrt{\frac{d \log t}{t}} \right)$. 
Theorem: let $H$ be a hypothesis set with $VC_{\text{dim}}(H) = d$ and disagreement coefficient $\theta$. Then, the label complexity of DHM is bounded by

$$\tilde{O}\left(\theta \left( d \log^2 \frac{1}{\epsilon} + \frac{d\nu^2}{\epsilon^2} \right) \right),$$

where $\nu = R(h^*)$. 
Heuristics

(see for example (Tong and Koller, 2002))

Idea:

- select points close to the decision surface.
- poor theory: no guarantee.
- experiments: often effective.
Recent Algorithms

  - uniformly distributed linear classifiers.
  - log-concave distributions.

- Confidence-rated predictors (Zhang and K. Chaudhuri, 2014):
  - better sample complexity than disagreement-based ones (term better than dis. coeff.).
  - more general than margin-based techniques.
  - however, computationally inefficient.
Empirical Results

(Guyon, Cawley, Dror and Lemaire, 2011)

Active learning challenge (2011):

- algorithms allowed to query labels with a budget.
- performance measured in terms of AUC.
- disappointing results compared to baseline passive learning algorithms.
References


References

- D. Cohn, L. Atlas, and R. Ladner (1994) Improving generalization with active learning, Machine Learning, 15, 201–221.


