Advanced Machine Learning

Active Learning

MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Active Learning Setup

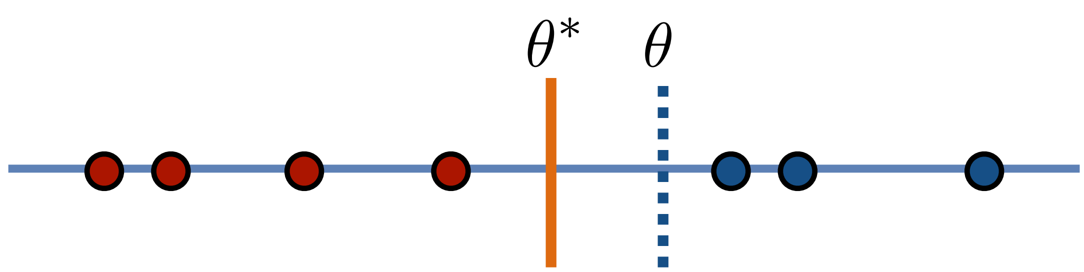
- Passive learning:
  - IID sample $((x_1, y_1), \ldots, (x_m, y_m)) \sim D^m$ is drawn.
  - learner receives full labeled sample.
- Active learning:
  - IID sample $((x_1, y_1), \ldots, (x_m, y_m)) \sim D^m$ is drawn.
  - learner has access to $(x_1, \ldots, x_m)$.
  - learner can request the label 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 θ^* within ϵ:
  - supervised learner needs $O(\frac{1}{\epsilon})$ samples since at least one point is needed in $[\theta^* - \epsilon, \theta^* + \epsilon]$.
  - active learner needs only $O(\log \frac{1}{\epsilon})$ using binary search.

$\longrightarrow$ exponential improvement!

Negative Result

(Kääriäinen, 2006)

- 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 $O\left(\frac{1}{\epsilon^2}\right)$.

CAL Algorithm

(Cohn, Atlas, and Ladner, 1994)

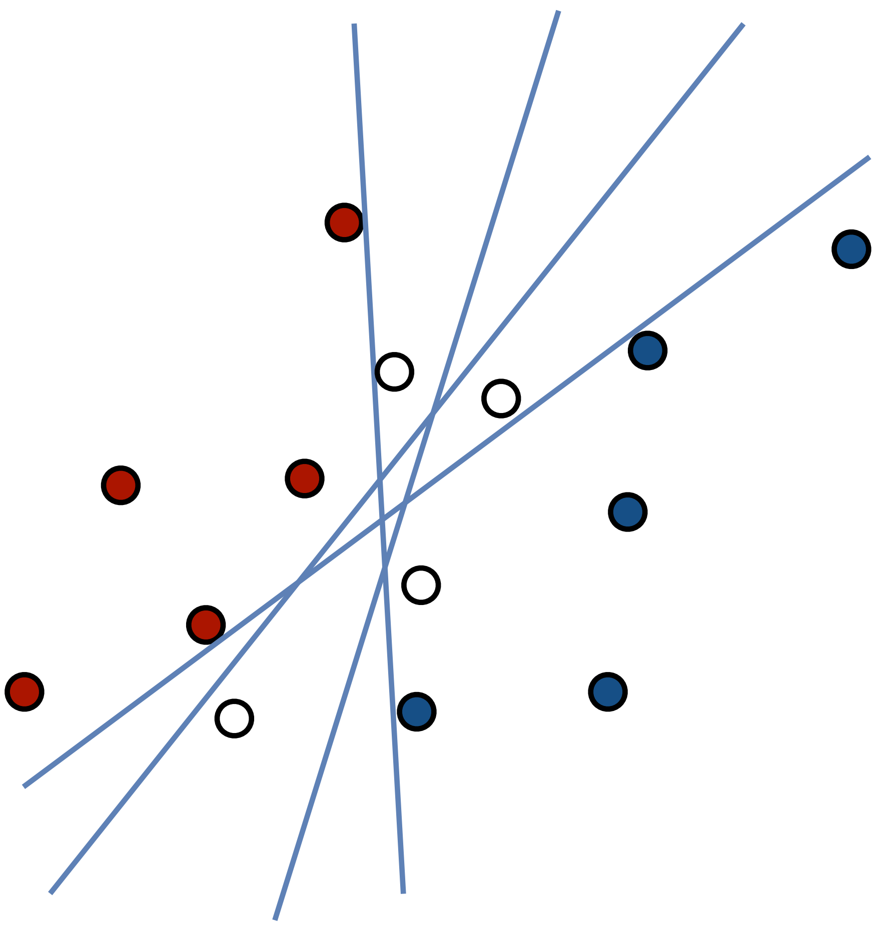
- Assume realizable case with hypothesis set H.

$\text{CAL}(H)$

$H_1 \leftarrow H$
for $t \leftarrow 1$ **to** T **do**
    if $(\exists h, h' \in H_t \colon h(x_t) \neq h'(x_t))$ **then**
        $y_t \leftarrow \textsc{QueryLabel}(x_t)$
        $H_{t+1} \leftarrow \{h \in H_t \colon h(x_t) = y_t\}$
    else $H_{t+1} \leftarrow H_t$
return H_{T+1}

CAL Algorithm

- Simple algorithm, but:
  - Computational cost of maintaining H_ts.
  - Separability requirement.



Definitions

(Hanneke, 2009)

- Region of disagreement:

$$\mathrm{DIS}(H) = \{x \in X \mid \exists h, h' \in H \colon h(x) \neq h'(x)\}.$$

- Disagreement metric:

$$d(h, h') = \Pr_{x \sim D}[h(x) \neq h'(x)].$$

- Disagreement ball:

$$B(h, r) = \Big\{h' \in H \colon d(h, h') \leq r\Big\}.$$

- Disagreement coefficient (rate of disagreement decrease):

$$\theta = \limsup_{r \to 0} \frac{\Pr\Big(\mathrm{DIS}\big(B(h^*, r)\big)\Big)}{r}.$$

Disagreement Coefficient

(Hanneke, 2009)

- Property: for all $r > 0$, $\mathrm{DIS}\big(B(h^*, r)\big) \leq \theta r$.
- Examples:
  - threshold functions: $\theta \leq 2$.
    - let $t \in B(t^*, r)$, then $t \in [t^* - \epsilon, t^* + \epsilon']$ where

      $$\epsilon = \operatorname*{argmax}_{\epsilon > 0} \{\Pr([t^* - \epsilon, t^*]) \leq r\} \quad \epsilon' = \operatorname*{argmax}_{\epsilon > 0} \{\Pr([t^*, t^* + \epsilon]) \leq r\}.$$
    - thus, $\mathrm{DIS}\big(B(h^*, r)\big) \leq 2r$.
  - 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 $\mathrm{VCdim}(H) = d$ and assume that the data is separable with disagreement coefficient θ. Then, the label complexity of CAL is bounded by

$$\widetilde{O}\left(\theta d \log \frac{1}{\epsilon}\right).$$

DHM Algorithm

(Dasgupta, Hsu, and Monteleoni, 2007)

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

DHM$((x_1, \ldots, x_T))$

$S \leftarrow \emptyset$ ◁ labels inferred
$T \leftarrow \emptyset$ ◁ labels queried
for $t \leftarrow 1$ **to** T **do**
    $h_+ \leftarrow \mathcal{A}(S \cup (x_t, +1), T)$
    $h_- \leftarrow \mathcal{A}(S \cup (x_t, -1), T)$
    if $(h_+ = \text{NIL})$ **then**
        $S \leftarrow S \cup \{(x_t, -1)\}$
    elseif $(h_- = \text{NIL})$ **then**
        $S \leftarrow S \cup \{(x_t, +1)\}$
    elseif $\widehat{R}_{S \cup T}(h_+) - \widehat{R}_{S \cup T}(h_-) > \Delta_t$ **then**
        $S \leftarrow S \cup \{(x_t, -1)\}$
    elseif $\widehat{R}_{S \cup T}(h_-) - \widehat{R}_{S \cup T}(h_+) > \Delta_t$ **then**
        $S \leftarrow S \cup \{(x_t, +1)\}$
    else $y_t \leftarrow$ QueryLabel(x_t)
        $T \leftarrow T \cup \{(x_t, y_t)\}$
return H_{T+1}

Notes

- $S \cup T$ not an i.i.d. labeled sample drawn according to D.
- Δ_t is defined by $\Delta_t = \beta_t^2 + \beta_t\left(\sqrt{\widehat{R}_t(h_+)} + \sqrt{\widehat{R}_t(h_-)}\right)$,

with $\beta_t = 2\sqrt{\frac{\log\big((8t^2+t)\Pi_{2t}^2(H)\big) + \log\frac{1}{\delta}}{t}} = \widetilde{O}\left(\sqrt{\frac{d\log t}{t}}\right)$.

DHM Guarantees

- **Theorem**: let H be a hypothesis set with $\mathrm{VCdim}(H) = d$ and disagreement coefficient θ. Then, the label complexity of DHM is bounded by

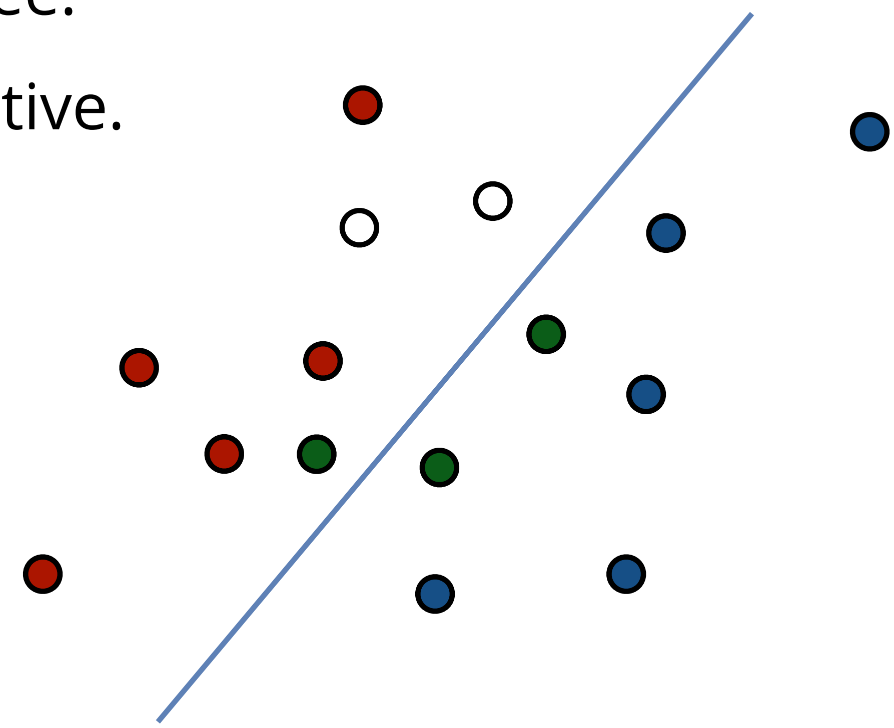
$$\widetilde{O}\left(\theta\left(d\log^2\frac{1}{\epsilon} + \frac{d\nu^2}{\epsilon^2}\right)\right),$$

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

- 'Margin-based active learning' (Balcan, Broder, and Zhang, 2007; Balcan and Long, 2013; Awasthi, Balcan, and Long, 2014): improvement over disagreement-based for
  - 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

- P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In STOC, 2014.
- M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In ICML, 2006.
- M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In COLT, 2007.
- M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In COLT, 2013.
- A. Beygelzimer, S. Dasgupta, and J. Langford (2009) Importance-weighted active learning, Twenty-Sixth International Conference on Machine Learning (ICML).

References

- D. Cohn, L. Atlas, and R. Ladner (1994) Improving generalization with active learning, Machine Learning, 15, 201–221.
- S. Dasgupta (2011) Two faces of active learning, Theoretical Computer Science, 412, 1767–1781.
- S. Dasgupta and D.J. Hsu. Hierarchical sampling for active learning, in ICML, 2008.
- S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In NIPS, 2007.
- S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning, Journal of Machine Learning Research, 10, 281–299, 2009.

References

- I. Guyon, G. Cawley, G. Dror and V. Lemaire: Results of the Active Learning Challenge. Active Learning and Experimental Design at AISTATS, 19-45. 2011.
- S. Hanneke. Theoretical Foundations of Active Learning, Ph.D. Thesis, CMU Machine Learning Department, 2009.
- M. Kääriäinen. Active learning in the non-realizable case. In COLT, 2006.
- V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. Journal of Machine Learning, 11:2457–2485, 2010.
- S. Tong and D. Koller. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2:45–66, 2002.
- C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In NIPS, 2014.