Machine learning theory

Active learning

Hamid Beigy

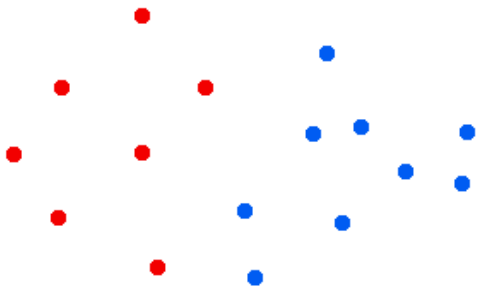
Sharif university of technology

June 6, 2022

Table of contents

Introduction

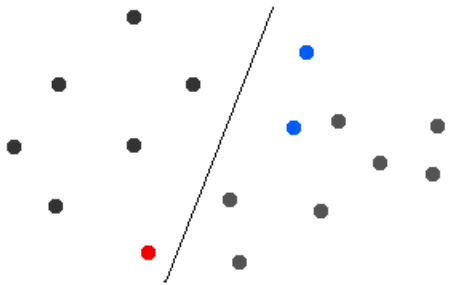
1. We have studied the passive supervised learning methods.
2. Given access to a labeled sample of size m (drawn iid from an unknown distribution $\mathcal{D}$), we want to learn a classifier $h \in H$ such that $\mathbf{R}(h) \leq \epsilon$ with probability higher than $(1 - \delta)$.



3. We need m to be roughly $\frac{VC(H)}{\epsilon}$ in realizable case and $\frac{VC(H)}{\epsilon^2}$ in urealizable case.
4. In many applications such as web-page classification, there are a lot of unlabeled examples but obtaining their labels is a costly process.

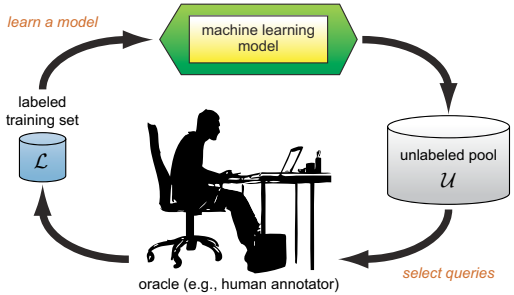
Active learning

1. In many applications **unlabeled data is cheap and easy to collect**, but **labeling it is very expensive** (e.g., requires a hired human).
2. Considering the problem of web page classification.
   - A basic web crawler can very quickly collect millions of web pages, which can serve as the unlabeled pool for this learning problem.
   - In contrast, obtaining labels typically requires a human to read the text on these pages to determine its label.
   - Thus, the time-bottleneck in the data-gathering process is the time spent by the human labeler.
3. The idea is to let the classifier/regressor pick which examples it wants labeled.

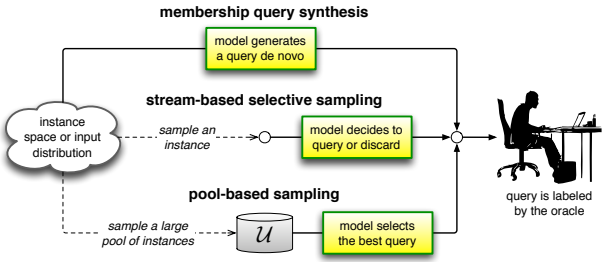


1. The hope is that by directing the labeling process, we can pick a good classifier at low cost.
2. It is therefore desirable to minimize the number of labels required to obtain an accurate classifier.
3. In passive supervised learning setting, we have
   - There is a set $\mathcal{X}$ called the **instance space**.
   - There is a set $\mathcal{Y}$ called the **label space**.
   - There is a distribution $\mathcal{D}$ called the **target distribution**.
   - Given a training sample $S \subset \mathcal{X} \times \mathcal{Y}$, the goal is to find a classifier $h : \mathcal{X} \mapsto \mathcal{Y}$ with acceptable error rate $\mathbf{R}(h) = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(x) \neq y]$.

1. In active learning, we have
   - There is a set $\mathcal{X}$ called the **instance space**.
   - There is a set $\mathcal{Y}$ called the **label space**.
   - There is a distribution $\mathcal{D}$ called the **target distribution**.
   - The learner have access to sample $S_X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\infty\} \subset \mathcal{X}$.
   - There is an oracle that labels each instant $\mathbf{x}$.
   - There is a budget m.
   - The learner chooses an instant and gives it to the oracle and receives its label.
   - After a number of these label requests not exceeding the budget m, the algorithm halts and returns a classifier h.



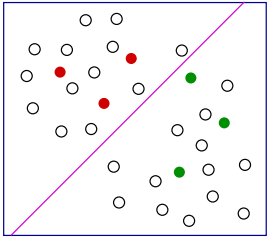
1. There are three main scenarios where active learning has been studied.



2. In all scenarios, at each iteration a model is fitted to the current labeled set and that model is used to decide which unlabeled example we should label next.
3. In membership query synthesis, the active learner is expected to produce an example that it would like us to label.
4. In stream based selective sampling, the learner gets a stream of examples from the data distribution and decides if a given instance should be labeled or not.
5. In pool-based sampling, the learner has access to a large pool of unlabeled examples and chooses an example to be labeled from that pool. This scenario is most useful when gathering data is simple, but the labeling process is expensive.

Typical heuristics for active learning(Dasgupta 2011)

1: Start with a pool of unlabeled data.
2: Pick a few points at random and get their labels.
3: **repeat**
4: Fit a classifier to the labels seen so far.
5: Query the unlabeled point that is closest to the boundary (or most uncertain, or most likely to decrease overall uncertainty,...)
6: **until** forever

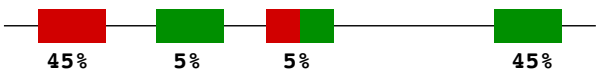


Biased sampling: the labeled points are not representative of the underlying distribution!

Typical heuristics for active learning

1: Start with a pool of unlabeled data.
2: Pick a few points at random and get their labels.
3: **repeat**
4: Fit a classifier to the labels seen so far.
5: Query the unlabeled point that is closest to the boundary (or most uncertain, or most likely to decrease overall uncertainty,...)
6: **until** forever

Example (Sampling bias)



Even with infinitely many labels, converges to a classifier with 5% error instead of the best achievable, 2.5%. Not consistent!

Can adaptive querying really help?

There are two distinct narratives for explaining how adaptive querying can help

1. Efficient search through hypothesis space
2. Exploiting (cluster) structure in data

Efficient search through hypothesis space

1. Ideal case is when each query cuts the version space in two subsets.
2. Then perhaps we need just $\log|H|$ labels to get a perfect hypothesis!

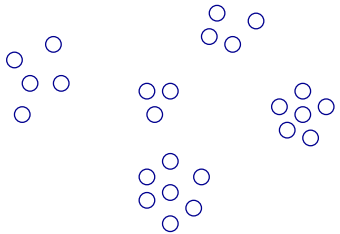
In general, the efficient search through hypothesis space has the following challenges

1. Do there always exist queries that will cut off a good portion of the version space?
2. If so, how can these queries be found?
3. What happens in the non-separable case?

Can adaptive querying really help?

Exploiting (cluster) structure in data

1. Suppose the unlabeled data looks like this



2. Then perhaps we just need five labels!

In general, the cluster structure has the following challenges

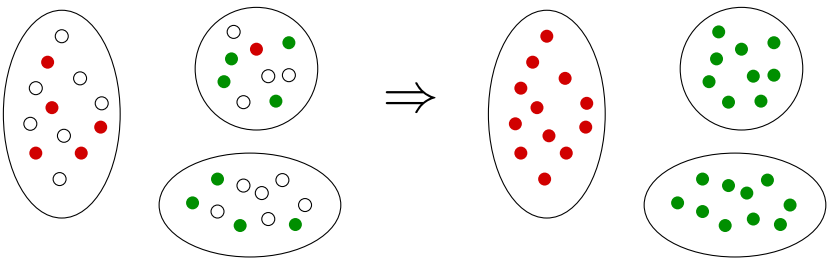
1. It is not so clearly defined
2. There exists at many levels of granularity.

The clusters themselves might not be pure in their labels.

How to exploit whatever structure happens to exist?

Exploiting cluster structure in data (Dasgupta and Hsu 2008)

1. Find a clustering of the data
2. Sample a few randomly-chosen points in each cluster
3. Assign each cluster its majority label
4. Now use this fully labeled data set to build a classifier



1. Threshold functions on the real line: $H = \{h_w \mid w \in \mathbb{R}\}$ and $h_w(x) = \mathbb{I}[x \geq w]$.

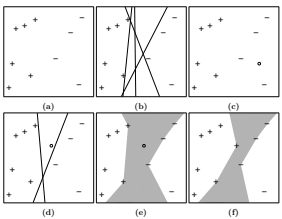
2. **Passive learning:** we need $\Omega\left(\frac{1}{\epsilon}\right)$ labeled points to have $\mathbf{R}(h_w) \leq \epsilon$.
3. **Active learning:** start with $\frac{1}{\epsilon}$ unlabeled points.

4. **Binary search:** need just $\log \frac{1}{\epsilon}$ labels, from which the rest can be inferred.
   Exponential improvement in label complexity!
5. **Challenges:**
   5.1 Nonseparable data?
   5.2 Other hypothesis classes?

Algorithm CAL (Cohn, Atlas, and Ladner 1994)

1: Let $h : \mathcal{X} \mapsto \{-1, +1\}$ and $h^* \in H$.
2: Initialize $i = 1$ and $H_1 = H$.
3: **while** $(|H_i| > 1)$ **do**
4: Select $\mathbf{x}_i \in \{\mathbf{x} \mid h \in H_1 \text{ disagrees}\}$. ▷ Region of disagreement
5: Query with $\mathbf{x}_i$ to obtain $y_i = h^*(\mathbf{x}_i)$. ▷ Query the oracle
6: Set $H_{i+1} \leftarrow \{h \in H_i \mid h(\mathbf{x}_i) = y_i\}$. ▷ Version space
7: Set $i \leftarrow i + 1$.
8: **end while**

CAL example



Problems

1. intractable to maintain H_i
2. nonseparable data

Definition (Label complexity(Hanneke 2014a,b))

Active learning algorithm A achieves label complexity m_A if, for every $\epsilon \geq 0$ and $\delta \in [0,1]$, every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and every integer m higher than $m_A(\epsilon, \delta, \mathcal{D})$, if h is the classifier produced by running A with budget m, then with probability at least $(1-\delta)$, we have $\mathbf{R}(h) \leq \epsilon$.

Definition (Disagreement coefficient (separable case)(Hanneke 2014a,b))

Let $\mathcal{D}_{\mathcal{X}}$ be the underlying probability distribution on input space $\mathcal{X}$. Let H_ϵ be all hypotheses in H with error less than ϵ. Then,

1. disagreement region is defined as

$$DIS(H_\epsilon) = \left\{ \mathbf{x} \;\middle|\; \exists h, h' \in H_\epsilon \text{ such that } h(\mathbf{x}) \neq h'(\mathbf{x}) \right\}.$$

2. Then, disagreement coefficient is defined as

$$\theta = \sup_{\epsilon} \frac{\mathcal{D}_{\mathcal{X}}(DIS(H_\epsilon))}{\epsilon}.$$

Example (Threshold classifier)

Let H be the set of all threshold functions in real line $\mathbb{R}$. Show that $\theta = 2$.

Algorithm CAL (Cohn, Atlas, and Ladner 1994)

1: Let $h : \mathcal{X} \mapsto \{-1, +1\}$ and $h^* \in H$.
2: Initialize $i = 1$ and $H_1 = H$.
3: **while** $(|H_i| > 1)$ **do**
4: Select $\mathbf{x}_i \in \{\mathbf{x} \mid h \in H_1 \text{ disagrees}\}$. ▷ Region of disagreement
5: Query with $\mathbf{x}_i$ to obtain $y_i = h^*(\mathbf{x}_i)$. ▷ Query the oracle
6: Set $H_{i+1} \leftarrow \{h \in H_i \mid h(\mathbf{x}_i) = y_i\}$. ▷ Version space
7: Set $i \leftarrow i + 1$.
8: **end while**

1. The label complexity of CAL can be captured by $VC(H) = d$ and disagreement coefficient θ.
   - For realizable case, label complexity of CAL equals to

$$\theta d \log(1/\epsilon).$$

   - For unrealizable case, label complexity of CAL equals to (If best achievable error rate is v)

$$\theta \left(d \log^2 \frac{1}{\epsilon} + \frac{dv^2}{\epsilon^2} \right).$$

Summary

Summary

1. We considered active learning problems:
2. There are different scenarios of active learning.
3. We defined two different measures of label complexity and disagreement coefficient.
4. We showed that the label complexity is characterized by $VC(H)$ of hypothesis space and disagreement coefficient θ.
5. It was shown that active learning decreases the label complexity in an exponential improvement over passive learning.

Readings

Readings

1. Read the papers given in the references.

- Cohn, David, Les Atlas, and Richard Ladner (May 1994). "Improving Generalization with Active Learning". In: *Machine Learning* 15.2, pp. 201–221.
- Dasgupta, Sanjoy (Apr. 2011). "Two faces of active learning". In: *Theoretical Computer Science* 412.19, pp. 1767–1781.
- Dasgupta, Sanjoy and Daniel J. Hsu (2008). "Hierarchical sampling for active learning". In: *Proceedings of the 25 International Conference on Machine Learning (ICML)*. Vol. 307, pp. 208–215.
- Hanneke, Steve (2014a). *Theory of Active Learning*. Tech. rep. Pennsylvania State University.
- – (2014b). "Theory of Disagreement-Based Active Learning". In: *Foundations and Trends in Machine Learning* 7.2-3, pp. 131–309.
- Settles, Burr (2012). *Active Learning*. Morgan & Claypool Publishers.

Questions?