COMS 6998-4 Fall 2017 Presenter: Geelon So
October 9, 2017 Scribe: Yuemei Zhang

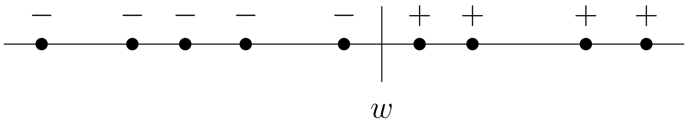
# Splitting Index

## 1 Introduction

In the setting of active learning, the data comes unlabeled and querying the label of a data point is expensive. The goal of an active learner is to reduce the number of labels needed and output a hypothesis with error rate $\leq \epsilon$. Recall that the usual sample complexity of supervised learning is $\Omega(1/\epsilon)$. The motivation for defining splitting index is to characterize the sample complexity of active learning. In the following parts of this section, we give some examples to show that the label complexity depends on the underlying distribution $\mathbb{P}$ and the target hypothesis $h^*$.

### 1.1 Motivating examples

**Good example: linear separators in $\mathbb{R}^1$**



Suppose the data lie on the real line, and the hypothesis class contains all the thresholding functions, that is, $\mathcal{H} = \{h_w : w \in \mathbb{R}\}$, where

$$h_w(x) = \begin{cases} 1 & \text{if } x \geq w, \\ 0 & \text{if } x < w. \end{cases}$$

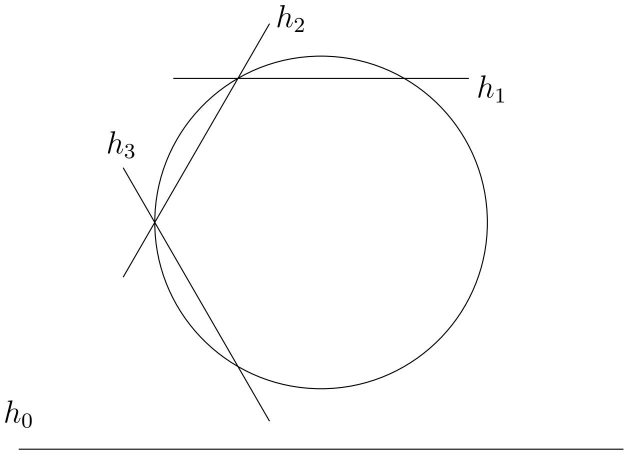
To learn the linear separator in $\mathbb{R}^1$ with error rate less than $\epsilon$, we only need to find two consecutive points with different labels, and require the gap between them to be less than $\epsilon$. The sample complexity of supervised learning is $O(1/\epsilon)$. For active learning, we instead draw $m = O(1/\epsilon)$ unlabeled points from $\mathbb{P}$. Then a simple binary search over these unlabeled points needs only $\log m = O(\log 1/\epsilon)$ labels. This example shows that compared with the sample complexity $O(1/\epsilon)$ of regular supervised learning, active learning gives an exponential improvement in label complexity over supervised learning.

**Bad example: linear separators in $\mathbb{R}^2$**

In this example we want to learn a linear separator in $\mathbb{R}^2$, and the underlying distribution $\mathbb{P}$ is supported on the unit circle. Consider the case where the target hypotheses is one of the following $n + 1$ hypotheses ($n = 1/\epsilon$):

$h_0$: all points are positive.

$h_i$, $i = 1, \ldots, n$: each $B_i = \{x : h_i(x) = 0\}$ lies on an arc of probability mass $\epsilon$, and all the $B_i$ are distinct.



Each time we ask for a point $x$, if $x$ turns out to be positive, its label eliminates at most one hypothesis in $h_1, \ldots, h_n$. In order to distinguish between $h_1, \ldots, h_n$, we need at least $\Omega(1/\epsilon)$ labels. This example shows that there are cases where active learning makes little improvement in the number of labels needed.

# 2 Preliminaries

## 2.1 The setting

Let $\mathcal{X}$ be an instance space with underlying distribution $\mathbb{P}$. The hypothesis class $\mathcal{H}$ is a set of functions from $\mathcal{X}$ to $\{0, 1\}$ with finite VC dimension. We focus on the realizable case of active learning where the target hypothesis $h^* \in \mathcal{H}$, with the non-Bayesian setting, that is, we have no prior on the space $\mathcal{H}$. To measure the distance between hypotheses in $\mathcal{H}$, we introduce a pseudometric induced by $\mathbb{P}$:

$$d(h, h') = \mathbb{P}\{x : h(x) \neq h'(x)\}.$$

Likewise, the notion of neighborhood is defined as $B(h, r) = \{h' \in \mathcal{H} : d(h, h') < r\}$. The error rate of a hypothesis $h$ is thus its distance to $h^*$. With this pseudometric we can measure the volume of a version space $\mathcal{S} \subset \mathcal{H}$ with its diameter:

$$\text{diam}(\mathcal{S}) = \sup_{h, h' \in \mathcal{S}} d(h, h').$$

## 2.2 Basic definitions

The goal of an active learning algorithm is to output a hypothesis $h \in \mathcal{H}$ with $d(h, h') < \epsilon$. To do this, it is sufficient to reduce the diameter of the version space to at most $\epsilon$, and output any hypothesis in the version space.

What we care about is how to quantify the amount by which a point $x \in \mathcal{X}$ reduces the diameter of the version space $\mathcal{S}$. Towards this end we imagine a graph with $\mathcal{H}$ as vertices and $\{h, h'\} \in \binom{\mathcal{H}}{2}$ as edges.

For any finite $Q \in \binom{\mathcal{H}}{2}$, a point $x \in \mathcal{X}$ is said to *$\rho$-split* $Q$ if it can eliminate at least a fraction $\rho$ of edges in the edge-set $Q$, that is, if:

$$\max\left\{\left|Q \cap \binom{\mathcal{H}_x^+}{2}\right|, \left|Q \cap \binom{\mathcal{H}_x^-}{2}\right|\right\} \leq (1-\rho)|Q|.$$

Reducing the diameter of $\mathcal{S}$ to at most $\epsilon$ is equivalent to eliminating all the edges of length $> \epsilon$. Therefore, we only care about edges of length more than $\epsilon$:

$$Q_\epsilon = \{\{h, h'\} \in Q : d(h, h') > \epsilon\}.$$

A subset of hypotheses $\mathcal{S} \in \mathcal{H}$ is $(\rho, \epsilon, \tau)$-*splittable* if for all finite edge-sets $Q \in \binom{\mathcal{S}}{2}$,

$$\mathbb{P}\{x : x \ \rho\text{-splits } Q_\epsilon\} \geq \tau.$$

Another way to measure the volume of the version space is to use the *covering number.* A set of hypotheses $\mathcal{S}_0 = \{h_1, \ldots, h_n\}$ is an $\epsilon$-cover of $\mathcal{H}$ if any $h \in \mathcal{H}$ is within distance $\epsilon$ of some $h_i \in \mathcal{S}_0$, that is, if

$$\mathcal{H} \subset \bigcup_{i=1}^{n} B(h_i, \epsilon).$$

The $\epsilon$-covering number of $\mathcal{H}$ the minimal size of such set:

$$N(\mathcal{H}, \epsilon) = \min\{n : \exists \ \epsilon\text{-cover over } \mathcal{H} \text{ of size } n\}.$$

Upon defining the $\epsilon$-cover of $\mathcal{H}$, suppose the closest element to $h^*$ in $\mathcal{S}_0$ is $h_0$. Then $h_0$ has an error rate at most $\epsilon$. The $\epsilon$-cover of $\mathcal{H}$ serves as a surrogate for the hypothesis class, and our algorithm only need to choose the best hypothesis in $\mathcal{S}_0$.

# 3 Lower bound

The splitting index gives a natural lower bound of label complexity.The lower bound shows that if the target hypothesis is in a subset of the hypothesis space with low splitting index, active learning makes little improvement in sample complexity over supervised learning.

**Theorem 1.** *Suppose that a set $\mathcal{S} \subset \mathcal{H}$ is not $(\rho, \epsilon, \tau)$-splittable for some $0 < \rho, \epsilon < 1$ and some $0 < \tau < 1/2$. Then any active learning algorithm that outputs a hypothesis of error $\leq \epsilon/2$ with probability $> 3/4$ on all target hypotheses in $\mathcal{S}$ needs either $\geq 1/\tau$ unlabeled samples, or $\geq 1/\rho$ labels.*

*Proof.* Suppose that we draw $m < 1/\tau$ unlabeled samples $x_1, \ldots, x_m$. Let $Q_\epsilon$ be a edge-set such that with probability at least $1 - \tau$, a point $x$ eliminates less than $\rho|Q_\epsilon|$ edges in $Q_\epsilon$. Then, with probability at least $(1-\tau)^{1/\tau} > 1/4$, none of $x_1, \ldots, x_m$ $\rho$-splits $Q_\epsilon$. We need at least $1/\rho$ labels to eliminate all the edges in $Q_\epsilon$. □

**Example.** Recall the example of linear separators in $\mathbb{R}^2$. The distance between $h_0$ and $h_i$ satisfies that $d(h_0, h_i) > \epsilon$ for all $i = 1, \ldots, n$, and the sets of points where $h_0$ and $h_i$ disagree on are disjoint. Therefore, any neighborhood of $h_0$ containing $h_1, \ldots, h_n$ is not $(\rho, \epsilon, \tau)$-splittable for $\tau > 0$ and $\rho > 1/n = \epsilon$. From the lower bound given by the splitting index, the label complexity of learning a linear separators in $\mathbb{R}^2$ is $\Omega(1/\epsilon)$.

# 4 Upper bound

Algorithm 1 gives an upper bound of label complexity. The algorithm chooses a hypothesis in an $(\epsilon/2)$-cover of $\mathcal{H}$. By halving the diameter of the version space in each iteration, the algorithm ensures that any hypothesis $h$ remaining in the version space after $T = \log(2/\epsilon)$ iterations has error

$$d(h, h^*) \leq d(h, h_0) + d(h_0, h^*) < \epsilon,$$

where $h_0$ is a hypothesis in the $(\epsilon/2)$-cover of $\mathcal{H}$.

**Algorithm 1** Active learning algorithm

```
input  Hypothesis class $\mathcal{H}$, $\epsilon > 0$.
Let $\mathcal{S}_0$ be an $(\epsilon/2)$-cover of $\mathcal{H}$.
for $t = 1, 2, \ldots, T = \log(2/\epsilon)$ do
    $\mathcal{S}_t = \text{split}(\mathcal{S}_{t-1}, 1/2^t)$.
end for
return  any $h \in \mathcal{S}_T$.
```

**Algorithm 2** split($\mathcal{S}, \Delta$)

```
input  Version space $\mathcal{S}$, $\Delta > 0$.
Let $Q_0 = \left\{ \{h, h'\} \in \binom{\mathcal{S}}{2} : d(h, h') > \Delta \right\}$.
$t \leftarrow 0$.
repeat
    Draw $m$ unlabeled points $x_{t1}, \ldots, x_{tm}$.
    Find the $x_{ti}$ which maximally splits $Q_t$.
    Ask for its label.
    Let $Q_{t+1}$ be the remaining edges.
    $t \leftarrow t + 1$.
until $Q_{t+1} = \emptyset$
return  remaining hypotheses in $\mathcal{S}$.
```

To estimate the sample complexity of Algorithm 1, first we give a lemma that analyzes the number of queries made by each iteration of the inner loop.

**Lemma 1.** *Suppose that $\mathcal{S}$ is $(\rho, \Delta, \tau)$-splittable. Then there is some setting of $m$ that guarantees with probability at least $1 - \delta$, Algorithm 2 will terminate after making*

$$\tilde{O}\left(\frac{d}{\rho} \log\left(\frac{1}{\epsilon}\right)\right)$$

*queries.*

*Proof.* In each step $t$, we draw $m$ unlabeled points. Since $\mathcal{S}$ is $(\rho, \Delta, \tau)$-splittable,

$$\Pr(x_{ti}\ \rho\text{-splits } Q_t) > \tau.$$

Thus, $\Pr(\text{no } x_{ti}\ \rho\text{-splits } Q_t) \leq (1 - \tau)^m$. Let $M$ denote that for some $t$, there is no $x_{ti}$ $\rho$-splits $Q_t$. Then,

$$\begin{aligned} \Pr(M \text{ happens at least once in the first } k \text{ steps}) &\leq k(1 - \tau)^m \\ &< ke^{-m\tau}. \end{aligned}$$

By choosing

$$m = \frac{1}{\tau} \ln \frac{k}{\delta},$$

this probability is less than $\delta$. If $M$ does not happen, each query of labels reduce at least a proportion $\rho$ of the edges. With $k = \frac{1}{\rho} \ln |Q_0|$, we have

$$|Q_k| \leq |Q_0|(1 - \rho)^k < |Q_0| e^{-\rho k} < 1.$$

Since there is an $(\epsilon/2)$-cover of $\mathcal{H}$ of size $O((1/(\epsilon/2))^{2d})$, the size of edge-set $|Q_0| = O((1/(\epsilon/2))^{4d})$. The number of queries made by *split* is

$$\begin{aligned} k &= \frac{1}{\rho} \ln |Q_0| \\ &= \frac{1}{\rho} \cdot O\left(d \log\left(\frac{1}{\epsilon}\right)\right) \\ &= \tilde{O}\left(\frac{d}{\rho} \log\left(\frac{1}{\epsilon}\right)\right). \end{aligned}$$

The total number of unlabeled samples is

$$\begin{aligned} m \cdot k &= \frac{1}{\tau} \ln \frac{1}{\delta} \ln k \\ &= \tilde{O}\left(\frac{d}{\rho\tau} \log\left(\frac{1}{\epsilon}\right)\right). \end{aligned}$$

□

**Theorem 2.** *Suppose $B(h^*, 4\Delta)$ is $(\rho, \Delta, \tau)$-splittable for all $\Delta > \epsilon/2$. Then, with probability at least $1 - \delta$, Algorithm 1 will return a hypothesis of error at most $\epsilon$, with at most*

$$\tilde{O}\left(\frac{d}{\rho\tau}\left(\log\frac{1}{\epsilon}\right)^2\right) \text{ unlabeled points, and at most } \tilde{O}\left(\frac{d}{\rho}\left(\log\frac{1}{\epsilon}\right)^2\right) \text{ labels.}$$

*Proof.* Let $T = \log(2/\epsilon)$. For $t = 1, \ldots, T$, we have

$$\mathcal{S}_t \subset B(h^*, 1/2^t + \epsilon/2) \subset B(h^*, 1/2^{t-1}).$$

Thus, $\mathcal{S}_t$ is $(\rho, 1/2^{t+1}, \tau)$-splittable according to our assumption. The function *split* is called $O(\log(1/\epsilon))$ times in Algorithm 1. From the results in Lemma 1, we have:

$$\text{sample complexity } = \tilde{O}\left(\frac{d}{\rho\tau} \log\left(\frac{1}{\epsilon}\right)^2\right), \text{and label complexity } = \tilde{O}\left(\frac{d}{\rho} \log\left(\frac{1}{\epsilon}\right)^2\right).$$

□

**Example.** Let us return to the simple example of linear separators in $\mathbb{R}^1$ with threshold hypotheses. In this example, $\mathcal{H}$ is $(\rho = 1/2, \epsilon, \epsilon)$-splittable for any $\epsilon > 0$. To see this, consider any finite edge-set $Q = \{\{h_i, h'_i\} : i = 1, \ldots, n\}$, where $h'_i - h_i > \epsilon$. Then the probability that we choose a data point which cuts the edge $\{h_i, h'_i\}$ is at least $\epsilon$. We sort the edges according to their left endpoints, and pick a point $x \in [a, b]$, where exactly half the edges have left endpoints less than $a$, and $b$ is the right endpoint of $a$. This ensures that at least half of the edges are eliminated. Note that in this example, the splitting index $\rho$ is constant on the whole hypothesis space. By directly calling *split*$(\mathcal{S}_0, \epsilon)$, we can achieve label complexity of $\tilde{O}\left(d/\rho \log 1/\epsilon\right)$.

# Bibliographic notes

The $\Omega(1/\epsilon)$ sample complexity of supervised learning and the covering number are due to Haussler [1]. The definition and results of the splitting index are from Dasgupta [2]. More recent work by Tosh and Dasgupta [3] takes the prior on $\mathcal{H}$ into consideration and improves the bounds of sample complexity given by the splitting index.

# References

[1] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. In *Information and computation*, 100(1), pages 78–150, 1992.

[2] S. Dasgupta. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 235–242, 2006.

[3] C. Tosh and S. Dasgupta. Diameter-Based Active Learning. *arXiv preprint arXiv: 1702.08553*, 2017.