COMS 6998-4 Fall 2017 Presenter: Geelon So
October 9, 2017 Scribe: Yanlin Duan

Splitting Index

1 Introduction

Recall that in the active learning setting, learner is provided with unlabeled samples, and can query teacher for the label. The goal is to learn a concept close enough to the target concept while minimizing the number of labels queried. Ideally the number of labels needed is much smaller than $\Omega(1/\epsilon)$, which is the number of labeled examples in the passive learning setting. To characterize the sample complexity, in this lecture we discussed another quantity to measure the effectiveness of active learning on particular concept classes and distributions: splitting index. We provide motivating examples, definition of splitting index, and a (coarse) lower and upper bound for label complexity based on it.

2 Motivating examples

2.1 Good example: linear separator in $\mathbb{R}^1$

We first examine the typical example of learning a threshold in $\mathbb{R}^1$ (Figure 1). If the underlying distribution $\mathbb{P}$ is separable, then as we learned in previous lectures/paper, VC theory shows we can achieve an error rate less than ϵ by drawing $m = O(1/\epsilon)$ labeled examples. Can we do better with active learning?

The answer is yes. If we draw m **unlabeled** samples, we can perform a binary search to find the threshold, which only requires $\log m = O(\log 1/\epsilon)$ labels. Going from $O(1/\epsilon)$ to $O(\log 1/\epsilon)$, we see an exponential improvement. Great!

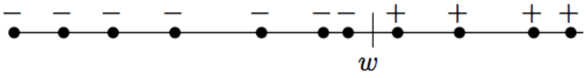

Figure 1: Example of a threshold function in $\mathbb{R}^1$ [1]

2.2 Bad example: linear separator in $\mathbb{R}^2$

We then consider the case of $\mathbb{R}^2$, where the hypothesis class is all linear separators in $\mathbb{R}^2$ and the input distribution $\mathbb{P}$ is supported on the perimeter of the unit circle. Unfortunately, we can construct a setting where $\Omega(1/\epsilon)$ labels have to be obtained.

Consider the following setting, shown in Figure 2. h_0 is an all-positive separator, and for each $h_i (1 \leq i \leq 1/\epsilon)$, they classify all points as positive, except for a small slice B_i with probability mass ϵ. Note that all B_i are disjoint on purpose.

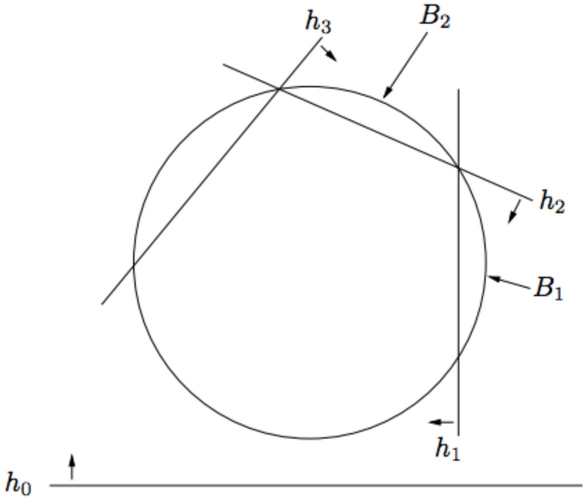

Figure 2: Example of linear separators on $\mathbb{R}^2$ [1]

Why this is bad? each time when we query the label of $x \in B_i$, we can only get rid of h_i one at a time. Therefore, in order to distinguish between $h_0, \ldots, h_i$, the label complexity can be as bad as $\Omega(1/\epsilon)$.

2.3 Why there are good and bad examples?

It is worth thinking why there are good and bad examples. For good examples, somehow, each query to label helps to reduce the 'size/volume' of 'search space' by a constant factor (say 1/2), and therefore, we get exponential improvement from binary-search-like algorithm. However, it is not always the case. In Section 2.2, we observe two **indicators** of bad examples:

1. There aren't good examples that can eliminate hypotheses in an effective way.
2. Even if there are good examples, in some bad scenario, your chance of drawing them from $\mathbb{P}$ can be low.

As we will see soon, splitting index is essentially defined based on those two observations.

3 Problem setting

We consider a binary classification setting. Denote the instance space as $\mathcal{X}$, with underlying probability distribution $\mathbb{P}$. Let $\mathcal{H}$ be the hypothesis class, where each hypothesis $h \in \mathcal{H}$ maps from $\mathcal{X}$ to $\{0, 1\}$.

We make three assumptions:

1. The scenario is **realizable**, meaning there exists a hypothesis $h^* \in \mathcal{H}$ that perfectly classifies all instances.

2. The VC dimension of $\mathcal{H}$, d is **finite**.

3. We consider **non-Bayesian setting**, so no prior on the space $\mathcal{H}$ is assumed.

Now we can define a pseudometric on $\mathcal{H}$, which is naturally suggested by the distribution $\mathbb{P}$:

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

Essentially, this pseudometric provides us with a measure of 'distance' between two hypotheses based on the distribution $\mathbb{P}$. Equipped with the pseudometric, we can further define:

- **Error of a hypothesis**: $\mathrm{err}(h) = d(h, h')$
- **Ball (neighborhood)**: $B(h, r) = \{h' \in \mathcal{H} : d(h, h') \leq r\}$
- **Diameter of a set** S: $\mathrm{diam}(S) = \sup_{h,h' \in S} d(h, h')$

As mentioned in the lecture, it can be helpful to think of the idea of "packing number". Consider Figure 3. The triangle represents the version space, and each ball inside represents a hypothesis and its neighbors with ϵ radius. Then the 'volume' of this version space is related to how many balls it can pack (i.e. the packing number), and as we will see in the next chapter, we are interested in those 'edges' connecting from the center of one ball to another ball, and how each sample x interact with them. More information on packing numbers can be found at [2].

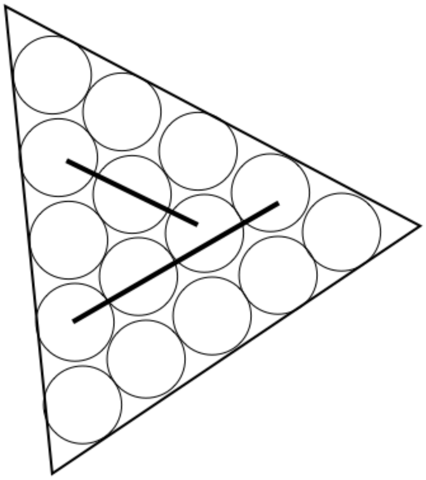

Figure 3: An illustration of packing number

4 Definition of splitting index

Consider any finite edge space $Q \subset \binom{\mathcal{H}}{2}$. Let $x \in \mathcal{X}$, we define the splitting index ρ as follows: we say x ρ-split Q if

$$\max\left\{\left|Q \cap \binom{\mathcal{H}_x^0}{2}\right|, \left|Q \cap \binom{\mathcal{H}_x^1}{2}\right|\right\} \leq (1-\rho)|Q|\,,$$

where $\mathcal{H}_x^0 = \{h \in \mathcal{H} : h(x) = 0\}$, and $\mathcal{H}_x^1 = \{h \in \mathcal{H} : h(x) = 1\}$.

Since our target accuracy is ϵ, for each Q we can construct Q_ϵ to only contain edges of length more than ϵ:

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

Then, a subset of hypotheses $S \subset \mathcal{H}$ is (ρ, ϵ, τ)-*splittable* if $\forall\, Q \subset \binom{S}{2}$:

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

Echoing with Section 2.3, ρ essentially models observation 1 (how good a good example can be), and τ models observation 2 (how easy is it to find a good example). Apparently, the bigger ρ and τ is, the better.

5 Analysis

We first prove the lower bound. The high level idea is, if splitting index in some region of $\mathcal{H}$ is too low, then it must contain hypotheses which active learning cannot improve.

Theorem 1. *Suppose $S \subset \mathcal{H}$ is not (ρ, ϵ, τ)-splittable for some $0 < \rho, \epsilon < 1$, and some $0 < \tau < 1/2$. Then with probability $> 3/4$, for any active learning strategy that achieves an accuracy of $\epsilon/2$ must either draw $\geq 1/\tau$ unlabeled samples, or request $\geq 1/\rho$ labels.*

Proof. The proof is straightforward. Let's draw less than $1/\tau$ unlabeled samples. Then with probability at least $(1-\tau)^{1/\tau} \geq 1/4$, none of these points ρ-splits Q_ϵ. Then there must exist a hypothesis in V for which at least $1/\rho$ labels are needed. □

The upper bound is shown via a simple algorithm. We first present the algorithm.

Algorithm 1 Main algorithm $\mathcal{A}$

```
input  Hypothesis class $\mathcal{H}$, $\epsilon > 0$.
Let $S_0$ be an $(\epsilon/2)$-cover of $\mathcal{H}$.
for t $= 1, 2, \ldots, T = \log 2/\epsilon$ do
    $S_t$ = split($S_{t-1}$,$1/2^t$)
end for
return  any $h \in S_T$.
```

Algorithm 2 split(S,Δ)

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

Lemma 1. *Suppose that $S \subset \mathcal{H}$ is (ρ, ϵ, τ)-splittable. Then with probability at least $1 - (1/\rho)(\ln|Q_0|)e^{-m\tau}$, Algorithm 2 will terminate after making at most $(1/\rho)\ln|Q_0|$ queries.*

Proof. Denote the number of rounds for Algorithm 2 to terminate as k. We 'hope' at each time $t < k$, among the m unlabeled examples we always have at least one x_{ti} that can ρ-splits Q_t. Then each query we can reduce the size of Q_t by (at least) a factor of ρ. Since when we terminate, the number of edges remaining should be less than one (i.e. $Q_{t+1} = \emptyset$), we see:

$$(1-\rho)^k \cdot |Q_0| \leq 1. \tag{1}$$

Solve for k, we get:

$$\begin{aligned} e^{-\rho k} \cdot |Q_0| &\leq 1 \qquad (2) \\ k &\leq (1/\rho)\ln|Q_0|\,. \qquad (3) \end{aligned}$$

From (1) to (2) we use the analytic lemma that $(1-\alpha)^m \leq e^{-m\alpha}, 0 < \alpha \leq 1$.

Now we show that our hope is actually quite likely to happen. Formally, we bound:

$$\begin{aligned} \mathbb{P}(\text{for some } t < k, \text{ no } x_{ti} \text{ can } \rho\text{-splits } Q_t) &\leq \sum_{t=0}^{k-1} \mathbb{P}(\text{no } x_{ti} \text{ can } \rho\text{-splits } Q_t) \qquad (4) \\ &\leq k(1-\tau)^m \qquad (5) \\ &\leq k e^{-m\tau} \qquad (6) \\ &\leq (1/\rho)(\ln|Q_0|)e^{-m\tau}, \qquad (7) \end{aligned}$$

where (4) uses union bound, and (6) uses the analytic lemma again.

Now we examine how big $|Q_0|$ is. Note that Q_0 is initially built on an $(\epsilon/2)$-cover of $\mathcal{H}$, which is of size $O(1/(\epsilon/2)^{2d})$ [2], then $|Q_0| \leq \left|\binom{S}{2}\right| = O(1/(\epsilon/2)^{4d})$. Plugging it in (3), we get:

$$k \leq (1/\rho)\ln|Q_0| = O\left(\frac{d}{\rho}\log\frac{1}{\epsilon}\right). \tag{8}$$

If we further set R.H.S of (6) as δ, and solve for m, we can get:

$$m = \frac{1}{\tau} \log \frac{k}{\delta}. \tag{9}$$

Then the total number of unlabelled examples in k iterations is:

$$\begin{aligned} mk &= \frac{k}{\tau} \log \frac{k}{\delta} \qquad (10) \\ &= \tilde{O}\left(\frac{d}{\rho\tau} \log \frac{1}{\epsilon} \right). \qquad (11) \end{aligned}$$

□

Now we prove Theorem 2 based on the above lemma:

Theorem 2. *Suppose $\mathcal{H}$ is (ρ, Δ, τ)-splittable for all $\Delta \geq \epsilon/2$. To get a version space V such that $h^* \in V$ and $\mathrm{diam}(V) < \epsilon$, the following sample and label complexity apply:*

- *Sample complexity: $\tilde{O}(\frac{d}{\rho\tau}(\log \frac{1}{\epsilon})^2)$*
- *Label complexity: $\tilde{O}(\frac{d}{\rho}(\log \frac{1}{\epsilon})^2)$,*

where d is VC-dimension of $\mathcal{H}$.

Proof. In Algorithm 1, for all $t \leq T - 1$, we have:

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

Therefore, S_t is (ρ, ϵ, τ)-splittable by definition. Then we need to call Algorithm 2 $O(\log 1/\epsilon)$ times. By Lemma 1, the total sample and label complexity would be:

- Sample complexity $= mk \cdot O(\log 1/\epsilon) = \tilde{O}(\frac{d}{\rho\tau}(\log \frac{1}{\epsilon})^2)$
- Label complexity $= k \cdot O(\log 1/\epsilon) = \tilde{O}(\frac{d}{\rho}(\log \frac{1}{\epsilon})^2)$

□

6 Application of splitting index

6.1 Linear separator in $\mathbb{R}^1$

Claim 1. *Let $\mathcal{H}$ be the hypothesis class of linear separators in $\mathbb{R}^1$. Then for any $\epsilon > 0$, $\mathcal{H}$ is $(1/2, \epsilon, \epsilon)$-splittable.*

Proof. Consider any finite set of edges $Q = \{\{h_{w_i}, h_{w'_i}\} : i = 1, \ldots, n\}$. Without losing generality, let's rank all threshold in non-decreasing order, and make sure $h_{w_i} \leq h_{w'_i}$, and only consider those edges with $h_{w'_i} - h_{w_i} > \epsilon$. Now, let's pick x where half the edges have left end less than x, and half the edges have right end greater than x. To put it in another way, x cuts at least half of the edges. Then we can see x can 1/2-split Q_ϵ and by calling Algorithm 2, we achieve the desired label complexity. □

6.2 Linear separator in $\mathbb{R}^2$

Claim 2. *Let $\mathcal{H}$ be the hypothesis class of linear separators in $\mathbb{R}^2$. The setting introduced in Section 2.2 has a label complexity of $\Omega(1/\epsilon)$.*

Proof. Since $\forall i, d(h_0, h_i) > \epsilon$, because of the disjoint slices B_i. Then the region containing h_0 and any of the h_i is not $(\rho > \epsilon, \epsilon, \tau)$-splittable. By Theorem 1, the label complexity is $\Omega(1/\epsilon)$. □

7 Further discussion

One question we discussed in class is why we need an iterative algorithm to step-by-step reduce the version space, rather than just calling split$(\mathcal{H}, \epsilon)$?

The answer is that if we do so, ρ may not be 'nice' when searching on a really big hypothesis class with a small ϵ. It is more efficient to step by step reduce the size of the version space, and change our threshold from Δ to our target ϵ gradually.

Bibliographic notes

The definition and analysis of the splitting index are due to [1]. Figure 1 and Figure 2 are also referred from [1]. The presenter also mentioned [3], a recent work on splitting index, which provides an efficient algorithm that is able to realize the upper bound.

References

[1] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In Y. Weiss, P. B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 235–242. MIT Press, 2006.

[2] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78 – 150, 1992.

[3] Christopher Tosh and Sanjoy Dasgupta. Diameter-based active learning. *CoRR*, abs/1702.08553, 2017.