CS6781 - Theoretical Foundations of Machine Learning

Lecture 6: PAC Sample Complexity Lower Bound

February 6, 2020

Lecturer: Nika Haghtalab *Readings: None*

1 Overview

In the last few lectures, we have proved the number of samples *sufficient* for PAC learning is

$$m_{\mathcal{C}}(\epsilon, \delta) \in O\left(\frac{d \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}\right). \quad (1)$$

In this lecture, we show that this sample complexity is nearly tight, up to a factor of $\ln(1/\epsilon)$.

Theorem 1.1 (PAC Sample Complexity Lower Bound)**.** *Any algorithm that PAC learns, with parameters ϵ and δ, a concept class $\mathcal{C}$ with VC dimension d, must use*

$$m_{\mathcal{C}}(\epsilon, \delta) \in \Omega\left(\frac{d + \ln(1/\delta)}{\epsilon}\right).$$

We will prove this for a constant value of δ in this lecture and leave the added $\frac{1}{\epsilon}\ln(1/\delta)$ sample complexity as an exercise. More precisely we prove the following theorem.

Theorem 1.2. *Any algorithm for PAC learning, with parameters ϵ and $\delta \leq 1/15$, a concept class of $\mathcal{C}$ of VC dimension d must use more than $(d-1)/(64\epsilon)$ samples (for the worst-case choice of $\mathcal{D}$).*

Proof. To prove that there exists a concept $c^* \in \mathcal{C}$ and a distribution $\mathcal{D}$ that requires a large number of samples, we will construct a fixed distribution $\mathcal{D}$ but label it based on a randomly chosen concept c^*. If we see that the expected probability of error is high over the choice of c^*, then there must be a c^* that would also lead to high error.

Consider a concept class $\mathcal{C}$ with VC dimension d. Let $\mathcal{Z} = \{x_1, \ldots, x_d\}$ be a set that is shattered by $\mathcal{C}$. We will construct a distribution $\mathcal{D}$ on the set $\mathcal{Z}$ that requires many samples. Since our distribution by construction will be supported on just $\mathcal{Z}$, without loss of generality we assume that all concepts in $\mathcal{C}$ are also just defined on $\mathcal{Z}$, i.e, we only consider their restriction to $\mathcal{Z}$. Note that because $\mathcal{C}$ shatters $\mathcal{Z}$, we have that $|\mathcal{C}| = 2^d$. So, every labeling of $x_1, \ldots, x_d$ is possible.

In this proof, we assume that a concept c^* for labeling the distribution is drawn uniformly at random from $\mathcal{C}$. Note that this is equivalent to choosing the label of each x_i from $\{-, +\}$ one at a time uniformly at random to determine the labeling induced by c^* on $\mathcal{Z}$.

Let $m = \frac{d-1}{64\epsilon}$ and $\mathcal{A}$ be any algorithm that uses at most m i.i.d. samples before picking a hypothesis h. We need to show that there exist a distribution $\mathcal{D}$ on $\mathcal{Z}$ and a concept $c^* \in \mathcal{C}$ for labeling this distribution such that $\mathrm{err}_{\mathcal{D}}(h) > \epsilon$ with probability at least $1/15$.

At a high level, we use a distribution $\mathcal{D}$ that puts a relatively large probability mass on one of the d points, say x_1, and splits the rest of the remaining probability on $x_2, \ldots, x_d$ uniformly. Our goal is to show that any algorithm that takes m number of samples only, does not even encounter a large fraction of $x_2, \ldots, x_d$. Since these points could have been labeled any which way possible (by the choice of a random $c^* \in \mathcal{C}$) the algorithm cannot predict their labels with probability more than $1/2$ each. Therefore, the algorithm will incur a large error.

More formally, our distribution has point mass $p(\cdot)$ as follows

$$p(x) = \begin{cases} 1 - 16\epsilon & x = x_1 \\ \frac{16\epsilon}{d-1} & x = x_2, \ldots, x_d \end{cases}$$

Let $\mathcal{Z}' = \{x_2, \ldots, x_d\}$. For ease of presentation, we define $\mathrm{err}'(h) = \Pr[c^*(x) \neq h(x) \text{ and } x \in \mathcal{Z}']$. Note that $\mathrm{err}'(h) \leq \mathrm{err}_{\mathcal{D}}(h)$. So it suffices to show that $\mathrm{err}'(h)$ is large.

Let us now define the event $B(S)$: S contains less than $(d-1)/2$ points in $\mathcal{Z}'$. Note that, in expectation S contains $16\epsilon m = (d-1)/4$ points from $\mathcal{Z}'$. Furthermore, the number of points in S that fall in $\mathcal{Z}'$ is a binomial distribution. Since the median of this distribution is $\lfloor (d-1)/4 \rfloor$ or $\lceil (d-1)/4 \rceil$, we have that

$$\Pr_{S \sim \mathcal{D}^m}[B(S)] \geq 1/2. \tag{2}$$

We next show that for a uniformly chosen $c^* \in \mathcal{C}$,

$$\mathbb{E}_{c^*, S}[\mathrm{err}'(h) | B(S)] > 4\epsilon.$$

Recall that choosing a random c^* is equivalent to choosing the label of each x_i from $\{-, +\}$, one at a time and uniformly at random, to determine the labeling induced by c^* on $\mathcal{Z}$. When $B(S)$ holds, $\mathcal{A}$ has not seen at least $(d-1)/2$ of instances in $\mathcal{Z}'$. For each of these points, no matter what h is, in expectation over the labels assigned by c^*, h makes a mistake on that point with probability exactly $1/2$. Therefore,

$$\mathbb{E}_{c^*, S}[\mathrm{err}'(h) | B(S)] > \frac{d-1}{2} \times \frac{1}{2} \times \frac{16\epsilon}{d-1} = 4\epsilon. \tag{3}$$

Using Inequalities 2 and 3 we have that

$$\mathbb{E}_{c^*, S}\left[\mathrm{err}'(h)\right] > 2\epsilon.$$

This means that there is some $c^* \in \mathcal{C}$ such that $\mathbb{E}_S[\mathrm{err}'(h)] > 2\epsilon$. For the remainder of this proof, consider this c^*.

Note that $\mathrm{err}'(h) \leq 16\epsilon$ by definition, because it is only penalized by mistakes that h can make on $\mathcal{Z}'$. Let $p = \Pr_S[\mathrm{err}'(h) > \epsilon]$. We have

$$2\epsilon < \mathbb{E}_{c^*, S}\left[\mathrm{err}'(h)\right] \leq 16\epsilon \Pr[\mathrm{err}'(h) > \epsilon] + \epsilon \Pr[\mathrm{err}'(h) \leq \epsilon] \leq 16\epsilon p + (1-p)\epsilon,$$

which shows that $p < 1/15$. $\square$

2 The Gap between the Upper and Lower Bounds

Comparing the sample complexity upper bound of Equation 1 and lower bound of Theorem 1.1, we see that they differ by $\ln(1/\epsilon)$. So, which one is tight? Interestingly, you could consider both to be tight but in two different settings!

Auer and Ortner [2007] showed that $m_{\mathcal{C}}(\epsilon, \delta) \in \Omega\left(\frac{d\ln(1/\epsilon)+\ln(1/\delta)}{\epsilon}\right)$ is needed (and of course sufficient based on Equation 1), if we want that with probability $1 - \delta$,

$$\text{For all } h \in \mathcal{C}, \text{ if } \mathrm{err}_S(h) = 0 \text{ then } \mathrm{err}_{\mathcal{D}}(h) \leq \epsilon.$$

On the other hand, Hanneke [2016] shows that there is an algorithm that doesn't return just any arbitrary hypothesis $h \in \mathcal{C}$ that is consistent with the data and succeeds at PAC learning the class of $\mathcal{C}$ using only

$$m_{\mathcal{C}}(\epsilon, \delta) \in O\left(\frac{d + \ln(1/\delta)}{\epsilon}\right),$$

which is also needed by Theorem 1.1. The algorithm that achieves this improved bound takes the majority vote of classifiers, each of which are trained on data subsets specified by a recursive algorithm, with substantial overlaps between them. Note that because the majority vote of classifiers in $\mathcal{C}$ may not itself belong to $\mathcal{C}$, this algorithm is improperly PAC learning $\mathcal{C}$.

References

Peter Auer and Ronald Ortner. A new pac bound for intersection-closed concept classes. *Machine Learning*, 66(2-3):151–163, 2007.

Steve Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(1):1319–1333, 2016.