Lecture 4: Towards Characterizing Learnability

Instructor: Aditya Bhaskara *Scribe: Sierra Allred*

CS 5966/6966: Theory of Machine Learning

January 23rd, 2017

Abstract

In this lecture, we introduce the notion of the growth function corresponding to a hypothesis class. This is the key notion that allows us to prove the learnability of infinite hypothesis classes.

1 Introduction

Last class, we proved the "No Free Lunch" theorem, which shows that

- There is no universal learning algorithm.
- If we wish to learn an arbitrary function over a set of m points, we need at least $\frac{m}{2}$ training samples.

In other words, the hypothesis class $\mathcal{H}$ which consists of all possible functions over the domain cannot be learned (in the PAC model). Today we will see how to reason about more restrictive (and reasonable) function classes. The goal in this lecture and the next is to come up with a characterization of *learnable* hypothesis classes.

Towards this end, let us recall what we showed a couple of lectures ago: every *finite* $\mathcal{H}$ is learnable using $O\left(\frac{\log|\mathcal{H}|+\log(1/\delta)}{\epsilon^2}\right)$ samples (to an accuracy ϵ, w.p. $1-\delta$).

What about infinite classes? In the previous lecture, we discussed why we expect the class of "linear thresholds over the line" is learnable. In some sense, this is feasible for every $\mathcal{D}$, because we wish to find a hypothesis with small error, not necessarily the "right threshold". (We refer to the textbook for a formal treatment of this example. It turns out that linear thresholds are PAC learnable using $\approx \frac{\log(1/\delta)}{\epsilon^2}$ samples.)

2 The Growth Function

Suppose we have a (possibly infinite) hypothesis class $\mathcal{H}$, consisting of binary classifiers over a domain $\mathcal{X}$ (the output of the binary classifier is a sign $+$ or $-$). Now, suppose we fix an $S \subseteq \mathcal{X}$ of size m. Every hypothesis $h \in \mathcal{H}$, restricted to the points of S, produces a length-m "sign pattern" (i.e., the *string* $h(x_1)h(x_2)\ldots h(x_m)$, corresponding to $S = \{x_1, x_2, \ldots, x_m\}$). For example, the sign patterns induced by the three hypothesis classes on a set of five points are illustrated in Figure 2.

Let us denote by $\sigma_{\mathcal{H}}(S)$ the number of distinct sign patterns produced as above, as h varies over $\mathcal{H}$. Clearly, for all S of size m, we have $|\sigma_{\mathcal{H}}(S)| \leq 2^m$.

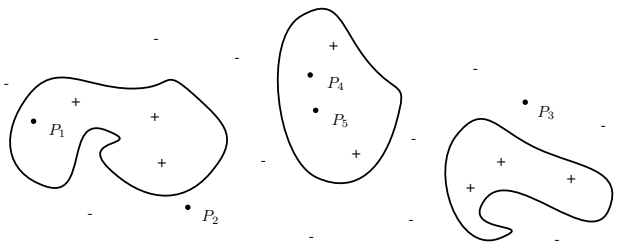

Figure 1: Figure showing five points and three hypotheses (interior of the curved shapes). The sign patterns induced on $p_1p_2p_3p_4p_5$ are $+,-,-,-,-$, $-,-,-,+,+$ and $-,-,-,-,-$.

2.1 definition. The growth function of a class $\mathcal{H}$ is denoted $\tau_{\mathcal{H}} : \mathcal{N} \to \mathcal{N}$, and $\tau_{\mathcal{H}}(m)$ is defined as the max number of sign patterns $\mathcal{H}$ can induce on a set of size m. In other words,

$$\tau_{\mathcal{H}}(m) := \max_{|S|=m} \sigma_{\mathcal{H}}(S).$$

Before proceeding, we observe that the "corollary" of the no-free-lunch theorem we saw in the previous lecture implies that no class for which $\tau_{\mathcal{H}}(m)$ is *exponential* is PAC-learnable.

To illustrate this, suppose $\mathcal{H}$ is a hypothesis class for which $\tau_{\mathcal{H}}(m) \geq 2^{m/10}$ for all m. Now, the corollary of the no-free-lunch theorem implies that as long as we obtain fewer than $m/100$ samples, we cannot learn $\mathcal{H}$ to an accuracy better than $1/20$, with probability $> 1/2$. This proof can be extended to the following theorem:

2.2 theorem. *Let $\mathcal{H}$ be a hypothesis class that satisfies $\tau_{\mathcal{H}}(m) \geq 2^{\Omega(m)}$, for all m. Then $\mathcal{H}$ is not PAC learnable.*

Growth Function for Linear Separators

Before proceeding, let us see the example of a simple hypothesis class, and study its growth function. A natural starting point is the set of linear thresholds over the real line (i.e., $\mathcal{X} = \mathbb{R}$).

Let S be any set of (possibly unordered) points $S = P_1, P_2, P_3, ..., P_m$ on a line, where each P_i is labelled with a $+$ or $-$. How may sign patterns are possible? If the points were ordered, the only patterns possible are

$$\begin{array}{cc} 1 & +,+,+,...,+ \\ 2 & -,+,+,...,+ \\ 3 & -,-,+,...,+ \\ \vdots & \\ m+1 & -,-,-,...,- \end{array}$$

If the points were not ordered, the patterns would simply be a permutation of the above sign patterns. In any case, we have that

$$\tau_{\mathcal{H}}(S) \leq m + 1.$$

Note that this is polynomial in m. The same is true for many natural hypoth-

esis classes, including axis parallel rectangles, half spaces in $\mathbb{R}^2$, etc.

3 Growth Function and Learning

The main theorem for today is that as long as $\tau_{\mathcal{H}}(m)$ is *sub-exponential*, the class $\mathcal{H}$ is PAC learnable. This is a *converse* to Theorem 2.2, and together they provide a characterization of learnability (although we will see a much more 'crisp' characterization in the next class).

3.1 theorem. *Let $\mathcal{H}$ be a class and $\tau_{\mathcal{H}}$ be its growth function. Suppose that*

$$\frac{\log(\tau_{\mathcal{H}}(m))}{m} \to 0 \text{ as } m \to \infty.$$

Then $\mathcal{H}$ is PAC learnable.

This follows from the following, more quantitative result:

3.2 theorem. *Let $\mathcal{H}$ be a class, then for every $\mathcal{D}$ and every $\delta \in (0,1)$ we have that with probability at least $1-\delta$ over the choice of S,*

$$(1) \quad \sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

Suppose there exists an m such that the RHS in Theorem 3.2 is $< \epsilon/2$. Then, the theorem states that w.p. at least $1-\delta$, the sample S provides a good approximation to the loss for every $h \in \mathcal{H}$, and thus empirical risk minimization yields a hypothesis whose loss is within ϵ of the optimal loss (which is what we require for PAC learning).

Now, as long as $\frac{\log(\tau_{\mathcal{H}}(m))}{m} \to 0$, there is bound to exist an m such that the RHS in Eq. (1) is $\leq \epsilon/2$. Thus, using the reasoning above, we conclude that Theorem 3.2 implies Theorem 3.1.

On proving theorem 3.2 The proof of the Theorem will be given after we discuss Rademacher complexity (couple of lectures from now). For now, we simply point out the difficulty in proving such a claim: if we had a single h for which we wish to show that $|L_{\mathcal{D}}(h) - L_S(h)|$ is small, then we could simply apply a Chernoff bound (as shown below). The issue is that we need to conclude this for *all* $h \in \mathcal{H}$. Since there are potentially infinitely many h, we cannot simply take a union bound over all $h \in \mathcal{H}$. (In the case of finite $\mathcal{H}$ we saw earlier, taking a union bound is exactly what gave us our bound.)

Thus, we defer the proof to Lecture 6.

For now, let us illustrate in detail how to apply a Chernoff bound (as we will do it often in the course), by showing that for any single h, $|L_{\mathcal{D}}(h) - L_S(h)|$ is small, with high probability.

One form of the Chernoff bound is the following::

There are a bunch of very good surveys on Chernoff bounds. One is by Chung and Lu, and another is by Boucheron, Lugosi and Bousquet.

3.3 theorem (Chernoff). *Suppose $X_1, X_2, ..., X_m$ are i.i.d. 0/1 random variables, with $\mathbb{E}[X_i] = p$. Then,*

$$\Pr(|\sum_i X_i - mp| \geq t) \leq 2e^{\frac{-t^2}{4m}}, \text{ or equivalently,}$$

$$\Pr\left(\left|\frac{1}{m}\sum_i X_i - p\right| \geq \frac{t}{m}\right) \leq 2e^{\frac{-t^2}{4m}}.$$

Thus for $0/1$ random variables with expectation p, the probability that the "empirical average" deviates from p by an amount t/m is bounded (roughly) by $\exp(-t^2/m)$. Thus, to estimate p up to an additive error of ϵ with probability $\geq 1-\delta$, we need to make sure that t/m is ϵ, and $2e^{-t^2/4m} \leq \delta$. These can be ensured by setting

$$m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right) \qquad \text{and} \qquad t = \sqrt{4m\log(2/\delta)}.$$

In our application $L_{\mathcal{D}}(h) = \mathbb{E}_{x\sim D}[\ell(h,x)]$ (where $\ell(h,x)$ is the loss for the hypothesis h on sample x). The loss is either 0 or 1. We are estimating this expectation via an empirical average over the m samples in S. Hence, we can apply the Chernoff bound, with t, m as above.