CS 598: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Aditya Potukuchi

Lecture # 4
9/15/2017

1 Overview of the lecture

In the last lecture, we derived bounds on the overfitting error of the ERM algorithm. We proved the following:

Theorem 1 (Restated)**.** *For ERM over a function class H, given m training examples where $m \geq \frac{1}{2\epsilon^2} \ln\left(\frac{2|H|}{\delta}\right)$, with probability $1-\delta$, we have that for every $h \in H$, we have:*

$$|\mathrm{err}_S(h) - \mathrm{err}(h)| \leq \epsilon$$

As a sanity check, the increasing-decreasing relationships between various parameters above makes sense. We have an bound on the *overfitting*:

$$|\mathrm{err}_S(h) - \mathrm{err}(h)| \leq \sqrt{\frac{1}{2m} \ln\left(\frac{|H|}{\delta}\right)}$$

The kind of guarantees of the form '$|\mathrm{err}_S(h) - \mathrm{err}(h)| \leq \epsilon$' are called *uniform convergence*. The reason for this name is that for the sequence random variables $\{\mathrm{err}_S(h)\}_{|S|}$ converges to their mean $\mathrm{err}(h)$ uniformly as $|S|$ grows.

Today, we shall see bounds for infinite classes of functions.

2 Uniform convergence bounds for infinite function classes

Let H (always) be the class of functions that we are trying to fit our data to, and let f be the true function we are trying to learn. The training data S is a (always) a set (or sequence) of m independently chosen samples $(x_1, y_1), \ldots, (x_m, y_m)$ according to some distribution $\mathcal{D}$. Similar to the previous (finite) case, we will study two cases, i.e.,

- When the true function f is in the function class H
- When the true function f is not in the function class H.

2.1 Learning an unknown function in the class H

Reminder: In the case where H was finite, the first part gave an error $\mathrm{err}(h) \leq \frac{1}{\epsilon} \ln\left(\frac{|H|}{\delta}\right)$.

In fact, we did even get a bound on an infinite function class (i.e., the thresholds example). We would like to do something similar here. We will start off with an important notion:

For a class of functions function $H = \{h : D \to \{\pm 1\}\}$ over some domain D, and a nonnegative integer $m \in \mathbb{N}$, we use $C[m] = C_H[m]$ to denote the number of ways any set of m points partitions the function class H by the labels induced on them. Formally,

$$C_H[m] = \max_{\substack{S \subseteq D \\ |S|=m}} |\{(h(s))_{s \in S} \mid h \in H\}|$$

Since we are still talking about binary classification, we trivially have that for any function class, $C[m] \leq 2^m$. For 'simple' classes of functions, we have much better bounds, for example, if H is the class of threshold functions (as discussed previously), then $C_H[m] \leq m + 1$. This quantity gives us a way to give error guarantees on the ERM algorithm. More concretely, we have the following theorem:

Theorem 2. *For ERM over H, and a true function $f \in H$, given a set S of m training examples, where $m \geq \frac{4}{\epsilon} \ln\left(\frac{2C[2m]}{\delta}\right)$, then with probability at least $1 - \delta$,*

$$\mathrm{err}(h) \leq \frac{1}{\epsilon} \ln\left(\frac{2C[2m]}{\delta}\right)$$

Where h is the function returned by the ERM algorithm.

Remark: Trying to redo the previous proof, we see that the main problem is in applying union bound. To overcome this, we use an idea that is commonly known as *symmetrization*.

Proof. Denote the bad event $A := \{\exists h \in H \mid \mathrm{err}_S(h) = 0,\ \mathrm{err}(h) \geq \epsilon\}$. This is what we would like to bound. Let S' be another independently chosen sequence of random sample $(x'_1, y'_1), \ldots, (x'_m, y'_m)$. We plan to use $\mathrm{err}_{S'}(h)$ as a proxy for $\mathrm{err}(h)$. Towards this, denote another event $B := \{\exists h \in H \mid \mathrm{err}_S(h) = 0,\ \mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}\}$.

[*Remark: This set S' is traditionally known as the* ghost sample.]

We will first prove the following claim that shows that $\mathrm{err}(h)$ can successfully be 'replaced' with $\mathrm{err}_{S'}(h)$.

Claim 1. $\Pr(A) \leq 2 \Pr(B)$

Proof. We shall prove $\Pr(B|A) \geq \frac{1}{2}$, and since $\Pr(B \cap A) \leq \Pr(B)$, this suffices. Since A holds, We have that there is a $h \in H$ such that $\mathrm{err}_S(h) = 0$ and $\mathrm{err}(h) \geq \epsilon$. We have that the random variable $\mathrm{err}_{S'}(h)$ is given by:

$$\mathrm{err}'_S(h) = \frac{1}{m} \sum_{i \in [m]} Z_i$$

Where Z_i is the indicator random variable $\mathbb{1}[h(x'_i) = y'_i]$. So, we have that

$$\mathbf{E}[\mathrm{err}_{S'}(h)] = \frac{1}{m} \sum_{i \in [m]} \mathbf{E}[Z_i] = \mathrm{err}(h) \geq \epsilon,$$

and so by Chernoff, we have

$$\Pr\left(\mathrm{err}_{S'}(h) \leq \frac{\epsilon}{2}\right) \leq e^{-\epsilon m/4} \leq \frac{1}{2},$$

which completes the proof. Of course, we assume that $C[2m] \geq 2$, since otherwise, it is trivial. □

Now, we just bound the probability of B. For this, we consider the following two random processes:

Process 1: Take m samples, call them S, take another m samples, call it S'.

Process 2: Take $2m$ samples. partition them randomly into S and S'.

We are working with Process 1. However, we note that both these processes produce the same distribution (in the case of discrete r.v.'s it's a simple exercise, the continuous case is somewhat subtle, won't be done here). So, the event B can be defined on either of them, and will have the same probability. So, it suffices to bound $\Pr(B)$ for Process 2.

Proceeding with this goal in mind, denote $T := S \cup S'$ to be a random set, and σ is a random permutation. We want to get an upper bound on $\Pr(B)(= \Pr_{\sigma,T}(B))$, and we will do it by obtaining a bound on $\Pr(B|T)$. The advantage of this is that if two functions look the same on $S \cup S' = T$, we only need to 'count it' once. In other words, we only need to do union bound over $C[2m]$ many function classes.

Let us formalize this: for a sequence T of the inputs from the domain, and a set of possible labels $L \in \{\pm 1\}^{|T|}$, define $H_L := \{h : H \mid h(T) = L(T)\}$, where $h(T) := (h(x))_{x \in T}$ is the sequence of labels that is given to T by h. For a fixed set T_0 and a sequence of labels L, define the bad event

$$B(L) = \{\exists h \in H_L \mid \operatorname{err}_S(h) = 0, \ \operatorname{err}_{S'}(h) \geq \epsilon/2\}$$

.

Clearly, we have:

$$(B|T = T_0) \subseteq \bigcup_{L \in \{\pm 1\}^{|T_0|}} (B(L)|T = T_0).$$

We will, therefore, bound $\Pr(B(L)|T)$. If the labels L give less than $\frac{\epsilon m}{2}$ errors on T, then the probability $\Pr(B(L)|T)$ is zero. Otherwise, there are at least $\frac{\epsilon m}{2}$ errors or an at least $\frac{\epsilon}{4}$ fraction of errors on T. Probability that all of these occur in the second half of the random permutation is at most

$$\prod_{i \in m} \left(1 - \frac{\epsilon}{4} - \frac{i}{2m}\right) < \left(1 - \frac{\epsilon}{4}\right)^m \leq e^{-\frac{\epsilon m}{4}}$$

By union bound over all the $C[2m]$ possible values of L, and using the law of total probability $\Pr(B) = \mathbf{E}[\Pr(B|T)] \leq \max_{T_0} \Pr(B|T = T_0)$, we have

$$\Pr(B) \leq C[2m] e^{-\frac{\epsilon m}{4}}.$$

Therefore, if $m \geq \frac{4}{\epsilon} \ln\left(\frac{2C[2m]}{\delta}\right)$, this probability is at most $1 - \delta$.

□

Let us try this to apply this bound to the previous threshold example, we get that if we want to get it right with probability at least $1 - \delta$, and error ϵ, we have the guarantee when we have at least

$$m \geq \frac{4}{\epsilon} \ln\left(\frac{4m + 2}{\delta}\right),$$

or

$$m \approx \frac{1}{\epsilon} \ln\left(\frac{1}{\epsilon}\right) \ln\left(\frac{1}{\delta}\right)$$

samples.

2.2 The quantity $C[m]$

We will now spend some time studying how $C[2m]$ behaves. For reasonable function classes, we would expect $C[k] \approx 2^k$, for small values of k. In other words, these function class at least should have all possible labels in a small set of inputs. This is a very combinatorial/geometric property of the function classes and leads to the definition of the *VC dimension* of a function class:

Definition 1 (VC dimension)**.** *The* VC *dimension of a function class H is the minimum x such that $C[x+1] < 2^{x+1}$.*

One reason that VC dimension is a good measure to study learnability is that for many function classes, it is reasonably well behaved, i.e., there are theorems of the form:

Theorem 3 (Informal)**.** *There is a d such that for $x \gg d$, $C[d] = O(x^d)$*

We will study this in more detail in the next lecture. Intuitively, this gives a measure on the expressiveness of a function class, with no concern for the underlying distribution that we are learning. This is the reason that VC dimension based bounds are not extremely reliable in practice, i.e., there are function classes of high VC dimension which can be learnt. However, we shall see that if the VC dimension of a function class is high, there is always a (possibly adversarially constructed) distribution which makes learning hard when there are too few samples.