COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Rohan Rao

Lecture #5
February 18, 2019

1 Review

Last time in class we proved Occam's Razor for a finite hypothesis size. To recap we proved that if we have an Algorithm A that finds a hypothesis $h_A \in \mathcal{H}$ consistent with m examples, where $m \geq \frac{1}{\epsilon}(\ln|H| + \ln(\frac{1}{\delta}))$, then $Pr(err_D(h_a) > \epsilon) \leq \delta$.

This theorem gives us a lot of useful intuition but now we want to prove a bound for any hypothesis space — not necessarily finite.

2 Growth Function

2.1 Motivation

Let us examine the positive half-lines problem that we discussed in lecture.

$$\mathcal{X} = \mathbb{R}, \mathcal{H} = \{ \text{ Positive Half Lines } \}$$

We want to think about how many functionally distinct hypotheses we have on a finite set of points. Consider this arbitrary set of m points:

$$\{0.1, 0.5, 3.7, 6.1, 10.4, 11.3\}$$

Let's consider the hypothesis h_1 which starts at 2.5, and h_2 which starts at 2.7 and both extend to the positive half line on $\mathbb{R}$. Are these two hypotheses functionally different on these six points? The answer is no — these hypotheses seem to be functionally equivalent. We can see that on these six points, there are only 7 different hypotheses that are not equivalent. We denote the hypotheses and the labels that these hypotheses give:

$$h_0 : \{(0.1, +), (0.5, +), (3.7, +), (6.1, +), (10.4, +), (11.3, +)\}$$

$$h_1 : \{(0.1, -), (0.5, +), (3.7, +), (6.1, +), (10.4, +), (11.3, +)\}$$

$$h_2 : \{(0.1, -), (0.5, -), (3.7, +), (6.1, +), (10.4, +), (11.3, +)\}$$

$$h_2 : \{(0.1, -), (0.5, -), (3.7, -), (6.1, +), (10.4, +), (11.3, +)\}$$

$$h_4 : \{(0.1, -), (0.5, -), (3.7, -), (6.1, -), (10.4, +), (11.3, +)\}$$

$$h_5 : \{(0.1, -), (0.5, -), (3.7, -), (6.1, -), (10.4, -), (11.3, +)\}$$

$$h_6 : \{(0.1, -), (0.5, -), (3.7, -), (6.1, -), (10.4, -), (11.3, -)\}$$

In fact, we can easily see that for m points, there are $m + 1$ distinct labelings possible given our hypotheses. An equivalent statement is that for our m points there are $m + 1$ distinct "behaviors" or "dichotomies." We observe that this is far fewer than the 2^m total possible labelings. This concept class/hypothesis space is PAC-learnable and we will see that the fact that the possible labelings allowed using our hypothesis space are much

smaller than the total possible labelings is a contributing factor to the PAC-learnability of this problem. Other examples of how many labelings certain hypothesis spaces give us include:

(1) half lines: $2(m+1) - 2 = 2m$. For every positive half line l^+ we reverse the half line to make it a negative half line l^-. This flips the label for every point labeled under l^+ producing $m+1$ additional labelings (except for the labelings that label all points positive and all points negative which we already counted in the set of all positive half lines). This gives us $m+1-2$ more labelings to the $m+1$ labelings we got from looking only at positive half-lines.

(2) intervals: $\binom{m}{2} + m + 1 = O(m^2)$. We choose two different points to be the ends of our interval giving us $\binom{m}{2}$ possibilities, or we choose our intervals around only one data-point (start and end are the same) which gives us m more possibilities, and we have the possibility of an empty interval which gives 1 more option.

These are all polynomial, a fact we will soon see to be important. Our guiding intuition is that we have a nice polynomial case for the number of labelings and a bad exponential case. Our intuition for now should say that the nice polynomial case suggests the problem is PAC-learnable

2.2 Function Definition

Given a hypothesis space $\mathcal{H}$, and given sample points $S = \langle x_1, \ldots, x_m \rangle$ We define the function on $\mathcal{S}$ to the collection of all labelings induced by S by hypotheses in $\mathcal{H}$:

$$\Pi_{\mathcal{H}}(S) = \{\langle h(x_1), \ldots, h(x_m)\rangle : h \in \mathcal{H}\}$$

We will overload this function to get a more useful function with the same spirit:

$$\Pi_{\mathcal{H}}(m) = \max_{|S|=m} |\Pi_{\mathcal{H}}(S)|$$

We call this the growth function. We observe that this counts the maximum number of distinct labelings on any set of m points which captures the quantities of interest in our motivating examples. Our goal will be to get error bounds that don't depend on $|\mathcal{H}|$ but rather on our growth function of the number of examples.

Our goal is to be able to treat the set of labelings on the training set as our hypothesis space, replacing $|\mathcal{H}|$ with $|\Pi_{\mathcal{H}}(S)| \leq \Pi_{\mathcal{H}}(m)$ in the bounds we proved last lecture. We cannot exactly do this, but we will do something similar. We also note that in the case of the $\Pi_{\mathcal{H}}(m)$ being polynomial, this will give a very good bound on the generalization error, almost as good as for the case of finite $|\mathcal{H}|$.

3 Generalization-error Bounds Based on the Growth Function

We spend the final part of class beginning the proof of the following theorem:

3.1 Theorem

With probability $\geq 1-\delta$, $\forall h \in \mathcal{H}$ if h is consistent on random training sample of size m, then $err_D(h) \leq \epsilon$ where:

$$\epsilon = O\left(\frac{\ln(\Pi_{\mathcal{H}}(2m)) + \ln(\frac{1}{\delta})}{m}\right)$$

Remark: If we compare this to the bound we proved last lecture, we see that it is similar with the main difference being the $\ln(\Pi_{\mathcal{H}}(2m))$ term instead of a $\ln(|\mathcal{H}|)$ term. This means that we can use the growth function of the hypothesis space as a proxy for the size of the hypothesis space. This is especially useful for hypothesis spaces that are infinite but have a small growth function (for example, positive half lines).

3.2 Proof:

3.2.1 Double Sample Trick

Here we describe some notation and a proof technique that will help us in this proof. Imagine drawing two samples S: which is our real sample and S': which is our "ghost sample" which is a proxy for the generalization error. This is useful because dealing with S' lets us work with a finite set of points rather than the generalization error over the entire domain. We will usually notate:

$$S : \langle x_1, ..., x_m \rangle$$
$$S' : \langle x'_1, ..., x'_m \rangle$$

Both of these samples are formed by choosing i.i.d samples from D (our distribution). The idea is to use the mistakes that h makes on $\mathcal{S}'$ as a proxy for the true error rate. We define

$$M(h, S) = \text{number of mistakes } h \text{ makes on } S$$

In the spirit of using S' as a stand-in for measuring the true error, we define two events:

$$B : \exists h \in \mathcal{H} : M(h, S) = 0 \wedge err(h) > \epsilon$$

This event describes a hypothesis that is consistent with S but has a high generalization error (h is ϵ-bad). Our goal is to bound the probability of this event.

$$B' : \exists h \in \mathcal{H} : M(h, S) = 0 \wedge M(h, S') \geq \frac{m\epsilon}{2}$$

This event describes a hypothesis that is consistent with S but makes $\frac{m\epsilon}{2}$ mistakes on our ghost sample. We will use this event to help us bound B. We will introduce some more notation later but we will proceed to the proof. We will begin this proof in steps.

3.2.2 Step 1

Claim: $Pr(B'|B) \geq \frac{1}{2}$ if $m \geq \frac{8}{\epsilon}$
Proof: We choose h for which B holds:

$$err(h) \geq \epsilon$$

We know h is consistent with S because B holds, so we only need to show that h makes a lot of mistakes on S'. We compute (using the fact that our ghost samples are drawn i.i.d, and linearity of expectation):

$$E[M(h, S')] \geq m\epsilon$$

We build intuition for the following inequality from the fact that the probability that h makes less than half its expected mistakes is small. We will learn more tools to formalize this intuition but for now we can take the following as a fact.

$$Pr\left(M(h, S') < \frac{m\epsilon}{2}\right) \leq \frac{1}{2}$$

Thus given B we can determine event B' happens with at least $1/2$ probability.

3.2.3 Step 2

Claim: $Pr(B) \leq 2Pr(B')$
Proof:

$$Pr(B') \geq Pr(B' \wedge B)$$

This is true because the event on the right hand side implies the event on the left hand side. We use definition of conditional probability to conclude that:

$$= Pr(B)Pr(B'|B) \geq \frac{1}{2}Pr(B)$$

The last inequality comes from substituting Step 1 for the $Pr(B'|B)$. We conclude that:

$$Pr(B') \geq \frac{1}{2}Pr(B) \Rightarrow Pr(B) \leq 2Pr(B')$$

3.2.4 Step 3

We consider two experiments, (a), (b). In experiment (a) we choose S, S' of size m as usual, and in (b) we choose S, S' as usual and then perform the following process.

for i in $[m]$:
    flip a coin
    if heads, do nothing
    if tails, swap examples x_i and x'_i

We will call our new sets after performing this process T, T' corresponding to S, S' after we apply the process. We observe that the distributions of T, T' are identical to those of S, S'. This is true because since the samples are chosen i.i.d, all permutations of the data are equally likely, so permuting the data after it has already been sampled does not change its distribution.

We now define:

$$B'' : \exists h \in \mathcal{H} : M(h, T) = 0 \wedge M(h, T') \geq \frac{m\epsilon}{2}$$

This event describes a hypothesis that is consistent with T but makes $\frac{m\epsilon}{2}$ mistakes on our transformed ghost sample T'. From our earlier observation that the distributions of T, T', and S, S' are identical, we conclude that:

$$Pr(B') = Pr(B'')$$

3.2.5 Step 4

For a fixed h, We define $b(h)$ as the event that h satisfies the conditions of B''. This means that h is consistent with T but makes $\frac{m\epsilon}{2}$ mistakes on T'.

Claim:

$$Pr(b(h)|S, S') \leq 2^{-m\epsilon/2}$$

Proof: Since we fix h, and condition on S, S', the only randomness is over the random swapping that we use to construct T, T' . We will do this proof in cases:

(I) There is some index i where h makes a mistake on both x_i and x'_i
In this case whether or not we swap x_i and x'_i, h will make a mistake on T at this index, and so $M(h, T) \neq 0$ and so $Pr(b(h)|S, S') = 0$.

For the next cases we define r as the number of pairs $x_i \in S, x'_i \in S'$ where there is exactly one error between the two. We may assume for the remaining cases that h is correct on x_i or x'_i or both for all i, as we have handled the case where h is wrong on both x_i and x'_i for some i.

(II) $r < \frac{m\epsilon}{2}$
In this case we observe that even if we get lucky and after the random swaps all the examples on which h makes mistakes are in T', we still won't have enough mistakes to satisfy $b(h)$ and so we have $Pr(b(h)|S, S') = 0$

(III) $r \geq \frac{m\epsilon}{2}$
Here we need all of the coin flips to come out the right way so that all the mistakes are in T' and none in T. Since the coin flips are all independent:

$$Pr(b(h)|S, S') = \frac{1}{2^r} \leq 2^{-m\epsilon/2}$$

In any of these three cases our claim is satisfied as desired

The remaining steps of the proof will be covered next lecture.