COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Aditya Bhaskara

Lecture # 5
February 18, 2008

1 Introduction

Suppose $\mathcal{H}$ is a set of hypotheses and A is a learning algorithm that takes m training points sampled i.i.d. from some (unknown) distribution D, and produces a consistent hypothesis h. Then we saw last time that for any $\delta > 0$, with probability at least $1 - \delta$, the generalization error of h is at most $\frac{1}{m}(\log |\mathcal{H}| + \log \frac{1}{\delta})$.[1]

In this lecture we will see how to prove non-trivial bounds on the generalization error when $\mathcal{H}$ is infinite. We will introduce the important concept of VC dimension, and see how this purely combinatorial object plays an important role in learning.

In the last lecture, we defined the notion of 'number of possible behaviors' of a set of hypotheses $\mathcal{H}$ on a set of size m. In binary classification (which we are dealing with), a behavior on a set S is just a function mapping S to $\{\pm 1\}$. Define $\Pi_{\mathcal{H}}(S)$ to be the set of distinct behaviors on S, i.e., $\Pi_{\mathcal{H}}(S) = \{\langle h(x_1), h(x_2), \ldots, h(x_n)\rangle : h \in \mathcal{H}\}$ where $S = \{x_1, \ldots, x_m\}$. Further define

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

We saw, for instance, that when $\mathcal{H}$ is the set of positive half-lines, $|\Pi_{\mathcal{H}}(m)| = m+1$ and if $\mathcal{H}$ is the set of intervals, $|\Pi_{\mathcal{H}}(m)| = \frac{m(m+1)}{2} + 1$. The point here was that even though $\mathcal{H}$ is infinite, the number of behaviors it can have on sets of size m is 'small'. Our aim in this lecture is essentially to show that this is what matters for learning – *not* the size of $\mathcal{H}$.

More formally, we prove

Theorem 1. *Suppose $\mathcal{H}$ is a set of hypotheses and A is a learning algorithm that takes m training points sampled i.i.d. from some distribution D, and produces a consistent hypothesis h. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$err_D(h) \leq O\Big(\frac{\ln |\Pi_{\mathcal{H}}(2m)| + \ln(1/\delta)}{m}\Big).$$

2 The Proof

In this section, we prove Theorem 1. We use the so-called 'double-sampling trick', which will be described shortly.

As in the proof of the finite $\mathcal{H}$ case, it suffices to bound the probability of the following 'bad' event B:

$$B \ : \quad \exists h \in \mathcal{H} \text{ such that } h \text{ is consistent but } \mathrm{err}_D(h) > \epsilon.$$

Let us denote the training sample by $S = \{x_1, x_2, \ldots, x_m\}$. Also, let $M(h, S)$ denote the number of *mistakes* made by h on S. Now suppose $S' = \{x'_1, x'_2, \ldots, x'_m\}$ is another sample

[1] We could also turn the bound around and see it as a question of how many examples do we need to ensure an upper bound of some ϵ on the generalization error.

drawn i.i.d. from the same distribution D (the algorithm does not see this in some sense – it is purely for purposes of the proof). Define the event

$$B' \ : \quad \exists h \in \mathcal{H} \text{ such that } h \text{ is consistent and } M(h, S') > \frac{m\epsilon}{2}.$$

Claim 1. If $m > \frac{8}{\epsilon}$, then $\Pr[B'|B] \geq \frac{1}{2}$.

Proof. Suppose h is a consistent hypothesis but $\text{err}_D(h) > \epsilon$. Since S' is drawn i.i.d. from D, $\mathbb{E}[M(h, S')] \geq \epsilon m$. Further, $M(h, S')$ is the sum of i.i.d. binomial random variables, so it is highly concentrated around it's expectation. In particular it can be shown[2] that $\Pr[M(h, S') \leq \epsilon m/2] < \frac{1}{2}$. This proves the claim. □

Note that the claim immediately implies $\Pr[B] \leq 2\Pr[B']$, because

$$\frac{\Pr[B']}{\Pr[B]} \geq \frac{\Pr[B' \cap B]}{\Pr[B]} = \Pr[B'|B] \geq \frac{1}{2}.$$

Thus it suffices to bound $\Pr[B']$ (for clarity, recall that this probability is over choices of S *and* S'). Given S and S', consider the following random process SwapR.

1. For i from 1 to m, do the following:
2. Toss a coin. If you get heads, swap x_i and x'_i, else do nothing.

Say we denote the new collections by T and T'. Then the following is clear.
Claim 2. Suppose we pick S and S' according to D and then perform SwapR. Then the sets T and T' are identically distributed to S and S'.

Now suppose we define the event

$$B'' \ : \quad \exists h \in \mathcal{H} \text{ such that } h \text{ is consistent with } T \text{ (equiv. } M(h, T) = 0) \text{ and } M(h, T') > \frac{m\epsilon}{2}.$$

Claim 2 implies that $\Pr[B''] = \Pr[B']$. The first probability is over the choice of S, S' and the random bits of SwapR while the second is over choice of S, S'.

Now consider some $h \in \mathcal{H}$. We claim that $\Pr\left[M(h, T) = 0 \wedge M(h, T') > \frac{m\epsilon}{2} \middle| S, S'\right] \leq 2^{-m\epsilon/2}$. Consider

$$\begin{array}{cccc} h(x_1) & h(x_2) & \dots & h(x_n) \\ h(x'_1) & h(x'_2) & \dots & h(x'_n) \end{array}$$

First, note that if there is a column with both predictions wrong then $M(h, T) = 0$ can never happen and so we are done (the desired probability is 0). Similarly, if more than $(1 - \epsilon/2)m$ of the columns have both predictions right, we are done since $M(h, T') > \frac{m\epsilon}{2}$ cannot happen. Thus at least $r \geq \frac{m\epsilon}{2}$ columns have one right and one wrong prediction. If we need $M(h, T) = 0$, it must happen that in *all* such columns, SwapR must ensure that the right prediction goes to the top and the wrong one goes to the bottom row. Thus the probability is $2^{-r} \leq 2^{-m\epsilon/2}$.

So far we have not seen how the 'number of behaviors' $|\Pi_{\mathcal{H}}(m)|$ enters the picture. Our final claim shows precisely this.
Claim 3. $\Pr[B''] \leq |\Pi_{\mathcal{H}}(2m)| 2^{-m\epsilon/2}$.

[2] We will see techniques for proving such 'tail bounds' in the next few lectures.

Proof. Given a set S, define $\mathcal{H}'(S) \subseteq \mathcal{H}$ to be a set of size $|\Pi_{\mathcal{H}}(S)|$ where we choose one (*representative*) hypothesis for each different behavior of $\mathcal{H}$ on S. Then

$$\begin{aligned}
\Pr[B''] &= \mathbb{E}_{S,S'}\Big[\Pr\big[\exists h \in \mathcal{H} \text{ such that } M(h,T) = 0 \wedge M(h,T') > \frac{m\epsilon}{2} \mid S, S'\big]\Big] \\
&= \mathbb{E}\Big[\Pr\big[\exists h \in \mathcal{H}'(S \cup S') \text{ such that } M(h,T) = 0 \wedge M(h,T') > \frac{m\epsilon}{2} \mid S, S'\big]\Big] \\
&\leq \mathbb{E}\Big[\sum_{h \in \mathcal{H}'(S \cup S')} \Pr\big[M(h,T) = 0 \wedge M(h,T') > \frac{m\epsilon}{2} \mid S, S'\big]\Big] \\
&\leq \Pi_{\mathcal{H}}(2m) 2^{-m\epsilon/2}
\end{aligned}$$

This proves the claim. □

We are almost done. Combining claims 1 and 3, we have $\Pr[B] \leq \delta$ whenever

$$2|\Pi_{\mathcal{H}}(2m)|2^{-m\epsilon/2} \leq \delta$$

which is equivalent to saying

$$\epsilon \geq 2\Big(\frac{\log|\Pi_{\mathcal{H}}(2m)| + \log(2/\delta)}{m}\Big).$$

This finishes the proof of Theorem 1. We will now try to investigate the 'growth function' $|\Pi_{\mathcal{H}}(m)|$. In the next lecture, we will prove the remarkable theorem that no matter what the $\mathcal{H}$, the function grows either as 2^m or as m^d for some constant d (there is no other behavior!).

Looking at the bound in Theorem 1, we see that these two cases correspond to being able or unable to learn. We will make this connection precise in the next lecture.

3 Vapnik-Chervonenkis (VC) Dimension

As usual, let $\mathcal{H}$ denote a set of hypotheses over a set X.

Definition 1. *A set $Y \subseteq X$ is said to be* shattered *by $\mathcal{H}$ if for every function $f : Y \to \{\pm 1\}$, there exists $h \in \mathcal{H}$ such that $f(y) = h(y)$ for all $y \in Y$.*

Definition 2. *The* Vapnik-Chervonenkis (VC) dimension *of $\mathcal{H}$ is defined to be the size of the largest $Y \subseteq X$ that is shattered by $\mathcal{H}$.*

Note that we get to pick the Y, so showing that VC dimension is at least d is in NP while showing it is at most d is in some sense a co-NP problem. Now let us look at the VC dimension of some typical hypothesis classes $\mathcal{H}$.

Example 1. *$X = \mathbb{R}$, the real line and $\mathcal{H}$ is the set of positive half-lines. The VC dimension is* 1.

Proof. Clearly $\{x\}$ can be shattered (by choosing half-lines starting before and after x we get different behaviors). Also if $x < y$ are two real numbers, then we can never get the behavior $\langle h(x) = +1, h(y) = -1\rangle$. Thus no set of size > 1 can be shattered. □

Example 2. *$X = \mathbb{R}$, the real line and $\mathcal{H}$ is the set of intervals. The VC dimension is* 2.

Proof. Here $\{x, y\}$ can clearly be shattered. But if $x < y < z$ are three real numbers then $h(x) = h(z) = +1$ and $h(y) = -1$ can never occur, thus a set of size > 2 cannot be shattered. □

Example 3. *X is the set of points in the plane and $\mathcal{H}$ is the set of axis parallel rectangles. The VC dimension is* 4.

Proof. Showing the VC dimension is at least 4 is easy – consider 4 points in the shape of a diamond (say we look at $(1, 0), (-1, 0), (0, 1), (0, -1)$). It is clear this set can be shattered. Now suppose we have a set of 5 points. Consider the bottom-most, top-most, left-most and right-most points (they may coincide). We cannot have the behavior that h is $+1$ on these points and -1 on the remaining. Thus no set of size ≥ 5 can be shattered. □

Example 4. *$X = \mathbb{R}^n$ and $\mathcal{H}$ is the set of half-spaces. The VC dimension is $n + 1$.*

This will be left as an exercise on a homework.

Observe that in all the cases above, the VC dimension is in some sense, the 'number of parameters' needed to describe the hypothesis (in the half-line case it is the starting point, in an interval we need the two end-points, for an axis parallel rectangle we need the x, y coordinates of the principal diagonal, and so on). This 'thumb-rule' is quite often true, but not always. For instance, there are pathological examples in which the number of parameters is one, but the VC-dimension is infinite.