COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire — Lecture #5
Scribe: David Bieber — February 19, 2013

Recall Occam's razor. With probability at least $1-\delta$, a hypothesis $h \in \mathcal{H}$ consistent with m examples sampled independently from distribution D satisfies $\text{err}(h) \leq \frac{\ln|\mathcal{H}|+\ln\frac{1}{\delta}}{m}$.

Sample complexity for infinite hypothesis spaces

We seek to generalize Occam's razor to infinite hypothesis spaces. To do so, we look at the set of behaviors $\Pi_{\mathcal{H}}(S)$ of hypotheses from $\mathcal{H}$ on a sample S.

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

$$\Pi_{\mathcal{H}}(m) = \max_{S:|S|=m} |\Pi_{\mathcal{H}}(S)| \text{ defines the growth function of } \mathcal{H}.$$

Our goal is to modify Occam's razor to get a bound of the form

$$\text{err}(h) \leq O\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln\frac{1}{\delta}}{m}\right).$$

First, some definitions. For our proof of this bound, we fix ϵ and let D denote our target distribution. $S = \langle x_1,\ldots,x_m\rangle$ denotes a sample of $m > 8/\epsilon$ points chosen independently from D and provided to the learning algorithm. $S' = \langle x'_1,\ldots,x'_m\rangle$ denotes a second sample of m points, again chosen independently from D, termed the "ghost sample". S' is not provided to the learning algorithm.

Define $M(h,S)$ to be the number of mistakes h makes on S.

And define the event $B \equiv [\exists h \in \mathcal{H} : h$ consistent with S and h ϵ-bad$]$.

Our goal is to show that $Pr[B] \leq \delta$.

To do this we use the "double sample trick," taking the error of a hypothesis on the ghost sample as a proxy for the generalization error of the hypothesis. This allows us to make claims about a hypothesis being ϵ-bad without examining every hypothesis in an infinite hypothesis space. With the double sample trick, we can focus on what is happening on finite sets of points without worrying about what is happening over the entire domain.

Define the event $B' \equiv \left[\exists h \in \mathcal{H} : h \text{ consistent with } S \text{ and } M(h,S') > \frac{m\epsilon}{2}\right]$.

Step 1: $Pr[B'|B] \geq 1/2$

The same hypothesis h that makes B true makes B' true with probability at least $1/2$. To see this, suppose h is consistent with S and has $\text{err}(h) > \epsilon$. Then we must show with probability at least $1/2$ that $M(h,S') > \frac{m\epsilon}{2}$.

$E[M(h,S')] > m\epsilon$ since $\text{err}(h) > \epsilon$. Therefore by Chernoff bounds we conclude $Pr[M(h,S') \leq \frac{m\epsilon}{2}] \leq 1/2$.

Put another way, if B is true, then the probability that B' is not true is less than $1/2$. $Pr[B'|B] \geq 1/2$

Step 2: $Pr[B] \leq 2Pr[B']$

Trivially we have $B \wedge B' \implies B'$.

$$\begin{aligned}\text{So } Pr[B'] &\geq Pr[B \wedge B'] \\ &= Pr[B]Pr[B'|B] \\ &\geq \frac{1}{2}Pr[B] \text{ by step 1.} \\ \text{So } Pr[B] &\leq 2Pr[B'].\end{aligned}$$

Therefore it's sufficient to find an upper bound for $Pr[B']$, which will immediately give an upper bound for $Pr[B]$.

At this point we look at some intuition for why B' has low probability.
B' entails having no error on S and lots of error on S'. Since the samples were chosen independently, all permutations of the data among the sets S and S' are equally likely, so it is unlikely that all errors are in S'.

Consider two experiments for generating S and S'.

Experiment 1: Choose S and S' as usual, independently from D.

Experiment 2: First choose S and S' as usual, then move examples around as follows. For each $i \in \{1 \dots m\}$, with 50% probability swap element i of S with element i of S'. Call the resulting samples T and T'.

Notice that the distributions for T, T' exactly equals those for S, S'.

Now define $B'' \equiv \left[\exists h \in \mathcal{H} : h \text{ consistent with } T \text{ and } M(h, T') > \frac{m\epsilon}{2}\right]$

Step 3: $Pr[B''] = Pr[B']$

Since the distributions for T, T' exactly equals those for S, S' we conclude $Pr[B''] = Pr[B']$.

Define $b(h) \equiv \left[h \text{ consistent with } T \text{ and } M(h, T') > \frac{m\epsilon}{2}\right]$

Step 4: $Pr[b(h)|S, S'] \leq 2^{-m\epsilon/2}$

In step 4 we select S and S' according to the standard proceedure and then look at $b(h)$ before proceeding to construct T, T' according to experiment 2.

We show through three cases that $Pr[b(h)|S, S'] \leq 2^{-m\epsilon/2}$.

Case 1: h produces errors on two corresponding samples in S and S'

In this case, regardless of the permutations of the data selected by experiment 2, $b(h)$ is false. So $Pr[b(h)|S, S'] = 0$.

To illustrate Case 1, we represent in the following table example sets S and S'. A zero in the ith column indicates that the label of the ith element is correct, and a one in the ith column indicates that the label of the ith element is incorrect. Experiment 2 allows for an element to be swapped with the element below it. In this example there will always be an error in the second row because of the column of ones.

S	1	0	1
S’	1	1	0

Let r denote the number of points in S with label different from the corresponding point in S'.

Case 2: $r < \frac{m\epsilon}{2}$

In this case, $b(h)$ cannot happen because there are not enough mistakes. In order for $b(h)$ to be true, all mistakes must occur in S' and none may occur in S and the total number of errors must exceed $\frac{m\epsilon}{2}$. Since there are insufficient errors for this to occur, we again have $Pr[b(h)|S, S'] = 0$.

Case 3: $r \geq \frac{m\epsilon}{2}$

In this case, for $b(h)$ to be true, all r errors must be placed in T' rather than T. These events happen independently, each with probability 1/2. So $Pr[b(h)|S, S'] = 2^{-r} \leq 2^{-m\epsilon/2}$.

In all three cases $Pr[b(h)|S, S'] \leq 2^{-m\epsilon/2}$.

Step 5: $Pr[B''] \leq \Pi_{\mathcal{H}}(2m)2^{-m\epsilon/2}$

We now go to bound the probability of B''. The number of behaviors a hypothesis can take on over the $2m$ samples in T, T' is finite, given by $\Pi_{\mathcal{H}}(2m)$. For each behavior, we select a single representative hypothesis $h \in \mathcal{H}$ giving that behavior, giving a set $\mathcal{H}'(S, S')$ of $\Pi_{\mathcal{H}}(2m)$ representative hypotheses.

To reach the second line that follows we will use marginalization ($Pr[A] = E_X\left[Pr\left[A|X\right]\right]$). We can go from the second to the third line since each hypothesis $h \in \mathcal{H}$ has the same behavior on S, S' as some other $h \in \mathcal{H}'(S, S')$.

$$\begin{aligned}
Pr[B''] &= Pr[\exists h \in \mathcal{H} : b(h)] \\
&= E_{S,S'} Pr[\exists h \in \mathcal{H} : b(h)|S, S'] \text{ using marginalization} \\
&= E_{S,S'} Pr[\exists h \in \mathcal{H}'(S, S') : b(h)|S, S'] \text{ as explained above} \\
&\leq E_{S,S'} \sum_{h \in \mathcal{H}'(S,S')} Pr[b(h)|S, S'] \text{ by union bound} \\
&\leq E_{S,S'} |\Pi_{\mathcal{H}}(2m)| 2^{-m\epsilon/2}
\end{aligned}$$

Lastly we finally get a bound on $Pr[B]$.

$$\begin{aligned}
Pr[B] &\leq 2Pr[B'] = 2Pr[B''] \\
&\leq 2|\Pi_{\mathcal{H}}(2m)|2^{-m\epsilon/2} \\
&\leq \delta.
\end{aligned}$$

If we solve explicitly for ϵ we see the final inequality holds when $\epsilon \leq 2\frac{\lg \Pi_{\mathcal{H}}(2m) + \lg(2/delta) + 1}{m} = O\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln \frac{1}{\delta}}{m}\right)$.

The nice case here is when $\Pi_{\mathcal{H}}(2m)$ is $O(m^d)$, in which case ϵ goes to zero quickly.

Let's now look at a good technique for bounding the growth function $\Pi_{\mathcal{H}}(2m)$.

VC dimension

A sample S of m points is shattered if $\mathcal{H}$ realizes all possible behaviors on S. There are 2^m such behaviors.

$$|\Pi_{\mathcal{H}}(S)| = 2^m$$

The Vapnik-Chervonenkis (VC) dimension of $\mathcal{H}$, VCdim($\mathcal{H}$), is given by the cardinality of the largest set shattered by $\mathcal{H}$.

As an example, consider $\mathcal{H} = \{\text{intervals on the real line}\}$. If S contains a single point then $\mathcal{H}$ shatters S. If S contains two points, again $\mathcal{H}$ shatters S. In both of these cases $\mathcal{H}$ contains hypotheses that produce every possible labeling of the points in S. Thus VCdim($\mathcal{H}$) ≥ 2. If S is a set containing 3 points, then $\mathcal{H}$ does not contain a hypothesis that labels the outer two points positive and the middle point negative, so $\mathcal{H}$ does not shatter S. Therefore VCdim($\mathcal{H}$) $= 2$.