COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Yang Feng

Lecture #6
February 20, 2008

1 First thought

We have $\Pi_{\mathcal{H}}(S) = \{\langle h(X_1), h(X_2), \cdots, h(X_m)\rangle : h \in \mathcal{H}\}$, where $S = \langle X_1, \cdots, X_m\rangle$. And $\Pi_{\mathcal{H}}(m) = \max_{S:|S|=m} |\Pi_{\mathcal{H}}(S)|$.
We say that $\mathcal{H}$ shatters S if $|\Pi_{\mathcal{H}}(S)| = 2^m (m = |S|)$. VC-dim$(\mathcal{H}) = \max\{|S| : \mathcal{H}$ shatters $S\}$.
If $|\mathcal{H}| < \infty$, then $d =$ VC-dim$(\mathcal{H}) \leq \lg |\mathcal{H}|$. In fact, there are only two cases:

- VC-dim $= \infty \Rightarrow \Pi_{\mathcal{H}}(m) = 2^m, \forall m$
- VC-dim $= d < \infty \Rightarrow \Pi_{\mathcal{H}}(m) = O(m^d)$

This follows from Sauer's Lemma, which we now state and prove.

2 Sauer's Lemma

Lemma: $\forall \mathcal{H}$ with $d =$ VC-dim$(\mathcal{H})$,

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = \Phi_d(m) = O(m^d).$$

In other words, the sum of the binomial is just the number of different ways of choosing at most d items from a set of size m.

2.1 The Interval Example

In our examination of intervals, we found that the equation for the number of dichotomies possible was of the form:

$$\Pi_{\mathcal{H}}(m) = \binom{m}{2} + \binom{m}{1} + \binom{m}{0} = \Phi_2(m).$$

So Sauer's Lemma is tight in this example.

2.2 Proof of Sauer's Lemma

First, a few facts and conventions will be used in the proof:

- $\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$
- $\binom{m}{k} = 0$, if $k < 0$ or $k > m$

We will prove Sauer's Lemma by induction on $m + d$.
Base cases:
Our 2 base cases (for our 2 variables) are:

	$\mathcal{H}$						$\mathcal{H}_1$					$\mathcal{H}_2$			
	x_1	x_2	x_3	x_4	x_5		x_1	x_2	x_3	x_4		x_1	x_2	x_3	x_4
h_1	0	1	1	0	0	$\rightarrow$	0	1	1	0					
h_2	0	1	1	0	1						$\rightarrow$	0	1	1	0
h_3	0	1	1	1	0	$\rightarrow$	0	1	1	1					
h_4	1	0	0	1	0	$\rightarrow$	1	0	0	1					
h_5	1	0	0	1	1						$\rightarrow$	1	0	0	1
h_6	1	1	0	0	1	$\rightarrow$	1	1	0	0					

Table 1: Example Datasets for Proof of Sauers Lemma

- $m = 0$: $\Pi_{\mathcal{H}}(m) = 1 = \sum_{i=0}^{d} \binom{0}{i}$. It is the degenerate labeling of the empty set.
- $d = 0$: $\Pi_{\mathcal{H}}(m) = 1 = \binom{m}{0}$. You can not even shatter one point, so only one behavior possible.

Inductive Step:
Assuming lemma holds for any $m' + d' < m + d$. Given $S = \langle x_1, x_2, \cdots, x_m \rangle$, we want to show $|\Pi_{\mathcal{H}}(S)| \leq \Phi_d(m)$.

The main step of the proof is the construction of two new hypothesis spaces: $\mathcal{H}_1$ and $\mathcal{H}_2$ to which we can apply our induction hypothesis. Here, we have $\mathcal{H}_1$ and $\mathcal{H}_2$ defined on $S' = X' = \{x_1, x_2, \cdots, x_{m-1}\}$, that is, on all the points except x_m. $\mathcal{H}_1$ is constructed by just ignoring behavior on x_m. $\mathcal{H}_2$ is constructed by including only dichotomies that "collapsed" in $\mathcal{H}_1$.

As shown in the example in Table 1, h_1 and h_2, h_4 and h_5 are the same if we ignore x_5, so in each of these pairs, only one of goes to $\mathcal{H}_1$, and the other one goes to $\mathcal{H}_2$.

Notice that if a set is shattered by $\mathcal{H}_1$, then it is also shattered by $\mathcal{H}$. The reason is that we can generate $\mathcal{H}$ by using the same x_is when we generate $\mathcal{H}_1$. Thus we have

$$\text{VC-dim}(\mathcal{H}_1) \leq \text{VC-dim}(\mathcal{H}) = d$$

If a set T is shattered by $\mathcal{H}_2$, then $T \cup \{x_m\}$ is shattered by $\mathcal{H}$ since there will be two corresponding hypotheses in $\mathcal{H}$ with each element of $\mathcal{H}_2$ by adding $x_m = 1$ and $x_m = 0$. Thus, $\text{VC-dim}(\mathcal{H}) \geq \text{VC-dim}(\mathcal{H}_2) + 1$, which implies

$$\text{VC-dim}(\mathcal{H}_2) \leq d - 1.$$

Now, by induction, we have:

$$|\mathcal{H}_1| = |\Pi_{\mathcal{H}_1}(S')| \leq \Phi_d(m-1).$$

$$|\mathcal{H}_2| = |\Pi_{\mathcal{H}_2}(S')| \leq \Phi_{d-1}(m-1).$$

Then, we have

$$\begin{aligned}
|\Pi_{\mathcal{H}}(S)| &= |\mathcal{H}_1| + |\mathcal{H}_2| \\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\
&= \sum_{i=0}^{d} \binom{m}{i} \\
&= \Phi_d(m).
\end{aligned}$$

2.3 Upperbound on $\Phi_d(m)$

Claim: $\Phi_d(m) \leq (\frac{em}{d})^d$ for $m \geq d \geq 1$.
Proof:

$$\begin{aligned}
\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} &\leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} & &\text{-----------Since } \left(\frac{d}{m}\right) \leq 1. \\
&\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i 1^{m-i} & &\text{---------We are now adding nonnegative terms.} \\
&= \left(1 + \frac{d}{m}\right)^m & &\text{-------------By the binomial formula.} \\
&\leq e^d.
\end{aligned}$$

Then we have $\Phi_d(m) \leq (\frac{em}{d})^d$.

Using this bound, we will have the following results:
With probability of at least $1-\delta$, $\forall h \in \mathcal{H}$, if h is consistent with m examples, then

$$\text{err}(h) \leq \frac{2}{m}\left[d \lg\left(\frac{em}{d}\right) + \lg\left(\frac{1}{\delta}\right) + 1\right].$$

If $m = O(\frac{1}{\epsilon}[\ln(\frac{1}{\delta}) + d\ln(\frac{1}{\epsilon})])$, we have $\text{err}(h) \leq \epsilon$.

3 About the Lower Bound

Now, let's try to give a lower bound.

3.1 (Bogus) Argument on Lower Bound

Let D be uniform on $z_1, z_2, \cdots, z_d$. We run A with $m = d/2$ examples labeled arbitrarily, say A outputs h_A. Now let $c \in \mathcal{C}$ be any concept that is consistent with labels in S such that $c(x) \neq h_A(x)$ for $x \notin S$. Then we have $\text{err}(h_A) \geq 1/2$.

But, this is not a valid argument because we cannot choose target concept c after we choose h_A. The PAC model requires that we choose c before we choose S. So, in this argument, we are making c a function of h_A, which is in turn a function of S, which is obviously wrong.

3.2 A Theorem on the Lower Bound

We will instead prove the following:
Theorem: $\forall A$, $\exists c \in C, \exists D$, such that if A gets $m = d/2$ examples, where $d = \text{VC-dim}(C)$, then

$$\Pr\left[\text{err}(h_A) > \frac{1}{8}\right] \geq \frac{1}{8}$$

This means that if given only $d/2$ examples, then PAC learning is impossible for $\epsilon \leq 1/8$ and $\delta \leq 1/8$.