COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Brendan Miller

Lecture #6
February 23, 2006

1 Vapnik-Chervonenkis Dimension

1.1 Occam's Razor with the VC Dimension

Last time, we proved: with probability $\geq 1-\delta$, $\forall h \in \mathcal{H}$, if h is consistent with a sample of size m, then

$$err_D(h) \leq \frac{2}{m}\left(\log\left(\Pi_{\mathcal{H}}(2m)\right) + \log\left(\frac{1}{\delta}\right) + 1\right).$$

The size of $\Pi_{\mathcal{H}}(2m)$ is a property of the class of functions $\mathcal{H}$, thereby reducing the probabilistic problem to just a combinatorial problem.

1.2 Today's Goals

Today, we will look at how big $\Pi_{\mathcal{H}}(2m)$. There are only two possible cases:

$$\Pi_{\mathcal{H}}(2m) = \begin{cases} 2^m & \text{if VC-dim } d = \infty \\ \mathbf{O}(m^d) & \text{if VC-dim } d < \infty \end{cases}$$

$\mathcal{S}$ is shattered by $\mathcal{H}$ if

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

VC-dim($\mathcal{H}$) is the cardinality of the largest shattered set. A VC-dim of infinity means that an arbitrarily large set can be shattered by the class. For a finite class, the VC-dim is no greater than the log of the cardinality of the hypothesis class.

The VC-dim could be much smaller than this limit, though. For example, the VC-dim of positive half-lines is 1 (a set of two points cannot be shattered in the case of +/− labeling of the points). If the half-lines are defined by a large, but finite, number of points, then VC-dim($\mathcal{H}$) $\ll \log|\mathcal{H}|$.

1.3 Sauer's Lemma

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

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

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.

$$\binom{m}{i} = \frac{m\cdot(m-1)\cdot\ldots\cdot(m-i+1)}{i!}$$

So, the sum $\sum_{i=0}^{d}\binom{m}{i}$ when multiplied out becomes $\mathbf{O}(m^d)$. This has implications back with the use of the VC-dim in the PAC learning error limits: $\log\big(\Pi_{\mathcal{H}}(2m)\big) = \log\big(\mathbf{O}(m^d)\big) = \mathbf{O}\big(d\cdot\log(m)\big)$.

1.3.1 Example - Intervals

In our examination of intervals, we found that the equation for the number of dichotomies possible was of the form: $\Pi_{\mathcal{H}}(m) = 1 + m + \binom{m}{2}$. Or, now with Sauer's Lemma, we see that this is the exact same form as $\Phi_2(m)$.

1.3.2 Proof of Sauer's Lemma

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

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \quad \text{This comes from Pascal's Triangle}$$

$$\binom{m}{k} = 0 \quad \text{if} \begin{cases} k < 0 \\ k > m \end{cases} \quad \text{This convention is consistent with Pascal's Triangle}$$

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

$m = 0$	$\Pi_{\mathcal{H}}(m) = 1$	degenerate labeling of the empty set
$d = 0$	$\Pi_{\mathcal{H}}(m) = 1$	you cannot shatter 1 point even, so it's a single function

Induction step, $m \geq 1$ $d \geq 1$: assumes lemma holds for all m' d' for which $m' + d' < m + d$.

We are given or already know $\mathcal{H}$, $|S| = m$, $S = \langle x_1, x_2, \ldots, x_m \rangle$, and $d = \text{VC-dim}(\mathcal{H})$. We would like to show that $|\Pi_{\mathcal{H}}(S)| \leq \Phi_d(m)$.

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.

$\mathcal{H}$							$\mathcal{H}_1$						$\mathcal{H}_2$				
	$\mathbf{x_1, \ldots, x_m}$							$\mathbf{x_1, \ldots, x_{m-1}}$						$\mathbf{x_1, \ldots, x_{m-1}}$			
h1	0	1	1	0	0	$\rightarrow$	h1	0	1	1	0	$\rightarrow$	h1	0	1	1	0
h2	0	1	1	0	1	$\nearrow$											
h3	0	1	1	1	0	$\rightarrow$	h3	0	1	1	1						
h4	1	0	0	1	0	$\rightarrow$	h4	1	0	0	1	$\rightarrow$	h4	1	0	0	1
h5	1	0	0	1	1	$\nearrow$											
h6	1	1	0	0	1	$\rightarrow$	h6	1	1	0	0						

Figure 1: Example Datasets for Proof of Sauer's Lemma

$\mathcal{H}_1$ as shown in Figure 1 is defined to be $\mathcal{H}$ restricted to the domain of the first $m-1$ points in the set $\mathbf{S}$. There are as many different functions as there are possible behaviors. In other words:

$$\mathbf{X_1} = \{x_1, \ldots, x_{m-1}\} = \mathbf{S_1}$$
$$|\Pi_{\mathcal{H}_1}(\mathbf{S_1})| = |\mathcal{H}_1|$$

The claim is then that the VC-dim of $\mathcal{H}_1$ is no greater than the VC-dim of the original $\mathcal{H}$ ($\text{VC-dim}(\mathcal{H}_1) \leq d$). This is because all sets shattered by $\mathcal{H}_1$ will also be shattered by $\mathcal{H}$. By induction, then, $|\Pi_{\mathcal{H}_1}(\mathbf{S_1})| \leq \Phi_d(m-1)$.

Hypotheses where the dichotomies of $\mathcal{H}$ collapse into $\mathcal{H}_1$ are placed in $\mathcal{H}_2$ as shown in Figure 1. In the example, we see that both $x_m = 0$ and $x_m = 1$ are possible for $x_1, \ldots, x_{m-1}$ given in h1 and h4, but not for h3 and h6 in $\mathcal{H}_1$, so we only repeat h1 and h4. As for $\mathcal{H}_1$, the hypotheses in $\mathcal{H}_2$ are restricted to the domain $\{x_1, \ldots, x_{m-1}\}$. So:

$$\mathbf{X_1} = \mathbf{X_2} = \mathbf{S_1} = \mathbf{S_2}$$
$$|\Pi_{\mathcal{H}_2}(\mathbf{S_2})| = |\mathcal{H}_2|$$

The claim here is that the VC-dim of $\mathcal{H}_2$ is no greater than one less than the VC-dim of the original $\mathcal{H}$ (VC-dim($\mathcal{H}_2$) $\leq d-1$). This is because when we add x_m back, we will get a set that $\mathcal{H}$ can still shatter. In other words, if $\mathbf{T}$ is shattered by $\mathcal{H}_2$, then $\mathbf{T} \cup \{x_m\}$ will be shattered by $\mathcal{H}$. By induction, then, $|\Pi_{\mathcal{H}_2}(\mathbf{S_2})| \leq \Phi_{d-1}(m-1)$.

$$\begin{aligned}
|\Pi_{\mathcal{H}}(\mathbf{S})| &= |\mathcal{H}_1| + |\mathcal{H}_2| \\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \underbrace{\sum_{i=0}^{d-1} \binom{m-1}{i}}_{=\sum_{i=0}^{d}\binom{m-1}{i-1} \text{ because } \binom{x}{-1}=0} \\
&\leq \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1}\right] \\
&= \sum_{i=0}^{d} \binom{m}{i} \\
&= \Phi_d(m)
\end{aligned}$$

1.4 Upper Bound on Sample Complexity

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

Proof:

$$\begin{aligned}
\Phi_d(m) &= \sum_{i=0}^{d} \binom{m}{i} \\
\Phi_d(m) \cdot \left(\frac{d}{m}\right)^d &= \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^d \\
\left(\frac{d}{m}\right)^d \binom{m}{i} &\leq \left(\frac{d}{m}\right)^i \binom{m}{i} \\
\Phi_d(m) \cdot \left(\frac{d}{m}\right)^d &\leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \\
&\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i 1^{m-i} \qquad (x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \\
&= \left(1 + \frac{d}{m}\right)^m \qquad \forall x\ (1+x) \leq e^x \\
&\leq e^{\frac{d}{m} \cdot m} = e^d \\
\Phi_d(m) &\leq e^d \cdot \left(\frac{m}{d}\right)^d \\
&= \left(\frac{em}{d}\right)^d
\end{aligned}$$

So, now from our earlier limit, $\log\big(\Pi_{\mathcal{H}}(2m)\big)$ becomes roughly $d \cdot \log(\frac{2em}{d})$.

1.5 Lower Bound on Sample Complexity

Now, to get $err_D(h) \leq \epsilon$, we need $m = \mathbf{O}\big(\frac{1}{\epsilon} \cdot (\log \frac{1}{\delta} + d \cdot \log \frac{1}{\epsilon})\big)$ number of examples, which grows linearly with the VC-dim d. This also provides the sufficient conditions for learning. We can also now give a minimum number of examples to describe a class of hypotheses, which is not true when the bound used $\log |\mathcal{H}|$, where no lower bound would be possible.

So, now we will prove the lower bound in terms of the VC-dim to be able to PAC learn. The lower bounds must be in terms of the target concept class $\mathcal{C}$, not the hypothesis class $\mathcal{H}$ (so the limit will be in terms of VC-dim($\mathcal{C}$)).

To gain some intuition on this, we can look at if $\exists\, \overline{x_1}, \ldots, \overline{x_d}$ shattered by $\mathcal{C}$, and if we have $d-1$ points, then we cannot say what the next point d will be because both outcomes are possible.

Theorem Let $d =$ VC-dim(C). Then $\forall$ algorithms A, $\exists\, c \in \mathcal{C}$ and $\exists\, D$ such that if A gets $m \leq \frac{d}{2}$ examples from D labeled by c, then

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

In other words, this theorem says that you can't make ϵ and δ arbitrarily small. If $\epsilon < \frac{1}{8}$ and $\delta < \frac{1}{8}$, then you need at least $\frac{d}{2}$ examples to PAC learn. The textbook expands on this to say you need more than $\mathbf{\Omega}(\frac{d}{\epsilon})$ examples.

1.5.1 (Bad) Argument on Lower Bound

We let D be uniform over a shattered set $T = \langle \overline{x_1}, \ldots, \overline{x_d} \rangle$, and then run the algorithm A on $\frac{d}{2}$ of the examples from D to form S, then we will label them arbitrarily so that the algorithm will then output h_A. Now, we let $c \in \mathcal{C}$ be any concept consistent with the labels in S and such that $c_S(x) \neq h_A(x)\ \forall x \notin S$. Then, by this argument, $err_D(h_A) \geq \frac{1}{2}$.

But, this is not a valid argument because we cannot choose target concept c *after* we choose h_A. We need to 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, so that c is a function of S. This is wrong because we need to choose c *before* S. We want to be able to argue that we can choose c ahead of time and still give a lower bound on the error.

Next class, we will look at having D again be random over all T, but then choose c at random uniformly over the space of all possible dichotomies. Then, we'll finish the valid form of this argument to prove the above theorem.