COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Brian Tsang

Lecture #5
February 18, 2003

---

1 Theorem 3.2 (continued from lecture #4)

In general, we are trying to show that, with probability $\geq 1 - \delta$ for all h in our hypothesis space, that h being consistent implies $err_{\mathcal{D}}(h) \leq \epsilon$. To do this we are bounding the probability that there exists an h such that h is consistent yet has error $\geq \epsilon$.

1.1 Review of Previous Results

We were in the middle of proving that, with probability $\geq 1 - \delta$, $\forall h \in \mathcal{H}$:

$$h \text{ consistent } \Rightarrow err_{\mathcal{D}}(h) \leq O\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln \frac{1}{\delta}}{m}\right) \tag{1}$$

The following has been established (or asserted and deferred):

$$\begin{aligned} \Pr[B] &\leq 2\Pr[B'] && (2)\\ \Pr[e(h)|S,S'] &\leq 2^{\frac{-m\epsilon}{2}} && (3) \end{aligned}$$

Where:

$$\begin{aligned} S &\equiv \text{our training sample of } m \text{ random points according to } \mathcal{D} && (4)\\ S' &\equiv \text{our other sample of } m \text{ random points according to } \mathcal{D} && (5)\\ M(h) &\equiv \text{the number of mistakes } h \text{ makes on } S' && (6)\\ e(h) &\equiv h \text{ consistent with } S \wedge M(h) \geq \frac{m\epsilon}{2} && (7)\\ B &\equiv \exists h \in \mathcal{H} : h \text{ consistent with } S \wedge err_{\mathcal{D}}(h) > \epsilon && (8)\\ B' &\equiv \exists h \in \mathcal{H} : e(h) && (9) \end{aligned}$$

1.2 Working with Fixed S, S'

Let $\mathcal{H}' \equiv \{$one representative from $\mathcal{H}$ for every dichotomy of $S; S'\}$. Clearly, we have another interpretation of B':

$$B' \equiv \exists h \in \mathcal{H}' : e(h) \tag{10}$$

If we call the elements of $\mathcal{H}'$ h_1, h_2, ..., and h_N, we can then use the union bound:

$$\begin{aligned} \Pr[B'|S,S'] &= \Pr[\exists h \in \mathcal{H}' : e(h)|S,S'] && (11)\\ &= \Pr[e(h_1) \vee e(h_2) \vee \ldots \vee e(h_N)|S,S'] && (12)\\ &\leq \sum_{i=1}^{N} \Pr[e(h_i)|S,S'] && (13)\\ &\leq |\mathcal{H}'| \cdot 2^{\frac{-m\epsilon}{2}} && (14)\\ &= |\Pi_{\mathcal{H}}(S;S')| \cdot 2^{\frac{-m\epsilon}{2}} && (15) \end{aligned}$$

1.3 Unfixing Variables in General

We now take a break from the proof to explore a method for eliminating our dependance on a fixed S and S'. Let A be an arbitrary event, and X a random variable (it is irrelevant whether or not X and A are independent). Well, by the definitions of probability (see the notes from lecture #2),

$$\begin{aligned}
\Pr[A] &= \sum_x \Pr[A \wedge X = x] && (16)\\
&= \sum_x \Pr[X = x] \cdot \Pr[A|X = x] && (17)\\
&= \mathrm{E}_X\left[\Pr[A|X]\right] && (18)
\end{aligned}$$

1.4 Unfixing S and S' and Completing the Proof

Now we can use this result to bound $\Pr[B']$ with our bound for $\Pr[B'|S, S']$:

$$\begin{aligned}
\Pr[B'] &= \mathrm{E}_{S,S'}\left[\Pr[B'|S, S']\right] && (19)\\
&\leq \mathrm{E}_{S,S'}\left[\left|\Pi_{\mathcal{H}}(S; S')\right| \cdot 2^{\frac{-m\epsilon}{2}}\right] && (20)\\
&\leq \mathrm{E}_{S,S'}\left[\Pi_{\mathcal{H}}(2m) \cdot 2^{\frac{-m\epsilon}{2}}\right] && (21)\\
&= \Pi_{\mathcal{H}}(2m) \cdot 2^{\frac{-m\epsilon}{2}} && (22)
\end{aligned}$$

Using our other previous result:

$$\begin{aligned}
\Pr[B] &\leq 2Pr[B'] && (23)\\
&\leq 2\Pi_{\mathcal{H}}(2m) \cdot 2^{\frac{-m\epsilon}{2}} && (24)
\end{aligned}$$

Finally, setting this bound $\leq \delta$, we find that, with probability $\geq 1 - \delta, \forall h \in \mathcal{H}$,

$$err_{\mathcal{D}}(h) \leq \epsilon \leq \frac{2 \cdot \left(\lg \Pi_{\mathcal{H}}(2m) + \lg \frac{1}{\delta} + 1\right)}{m} \qquad (25)$$

2 The VC Dimension

The result we just derived is, of course, completely useless if we can't bound $\Pi_{\mathcal{H}}(2m)$ to some sub-exponential order, with respect to m. Sauer's lemma will do just that, but first we need to explore a new concept: the Vapnik-Chervonenkis Dimension.

2.1 Definitions

S is said to be *shattered* by $\mathcal{H}$ if every dichotomy of S has a representative in $\mathcal{H}$ (i.e. $|\Pi_{\mathcal{H}}(S)| = 2^{|S|}$).

The *VC dimension* of $\mathcal{H}$ is defined to be the size of the largest S which is shattered by $\mathcal{H}$ (i.e. $VCdim(\mathcal{H}) = \max\left(\{|S| : S \text{ is shattered by } \mathcal{H}\}\right)$)

2.2 Example: Intervals in $\mathbb{R}$

For example, let $\mathcal{H} = \{\text{intervals in } \mathbb{R}\}$. When S is composed of 1 or 2 samples, S is quite obviously shattered. If follows that $VCdim(\mathcal{H}) \geq 2$.

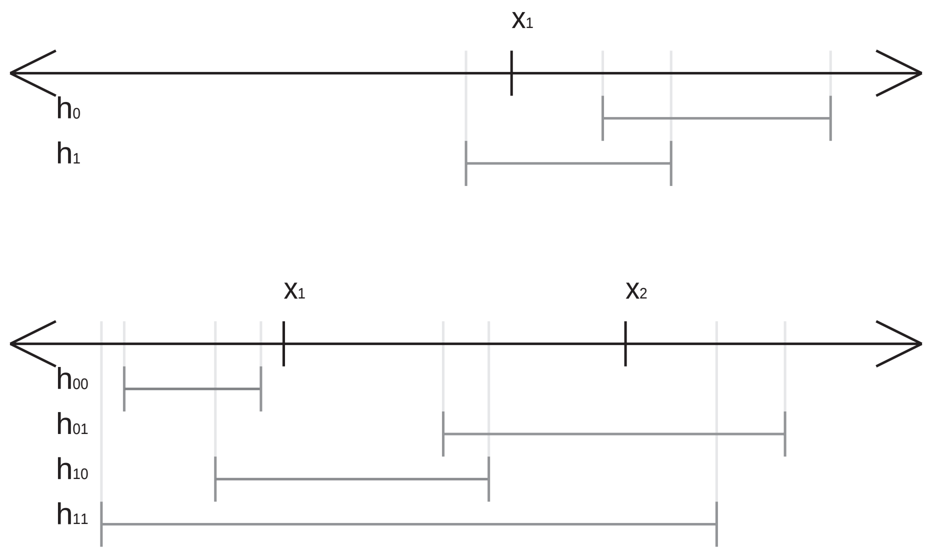

Figure 1: Representatives of $\mathcal{H}$ which shatter S when S is a set of 1 or 2 points.

However, when S is composed of 3 sample points, it is not shattered (if our sample points are x_1, x_2, and x_3 with $x_1 < x_2 < x_3$, there is no hypothesis which can label just x_1 and x_3 positive without also labelling x_2 positive).

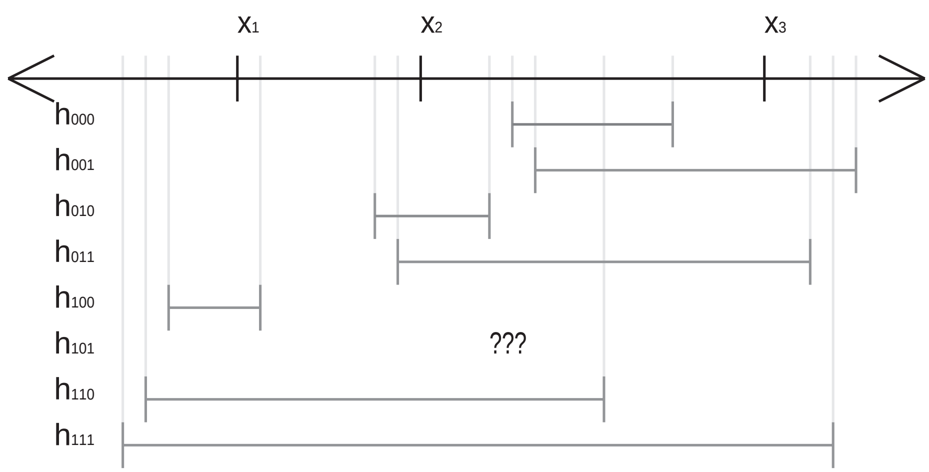

Figure 2: When S is a set of 3 points, we cannot find a hypothesis which marks the two outer points positive without also marking the inner point so.

To show that $VCdim(\mathcal{H}) < 3$, it is not sufficient to show that a single set of size 3 is not shattered. We need to show that *no* set of size 3 is shattered. However, in this case, it is evident that our argument applies to all sets of size 3. Thus, $VCdim(\mathcal{H}) = 2$.

Note that if no set of size d is shattered, then no larger set can be shattered either.

2.3 Example: Rectangles $\mathbb{R}^2$

Let $\mathcal{H} = \{\text{rectangles in } \mathbb{R}^2\}$. We will use a proof by picture to show that there is an S such that $|S| = 4$ and S is shattered by $\mathcal{H}$:

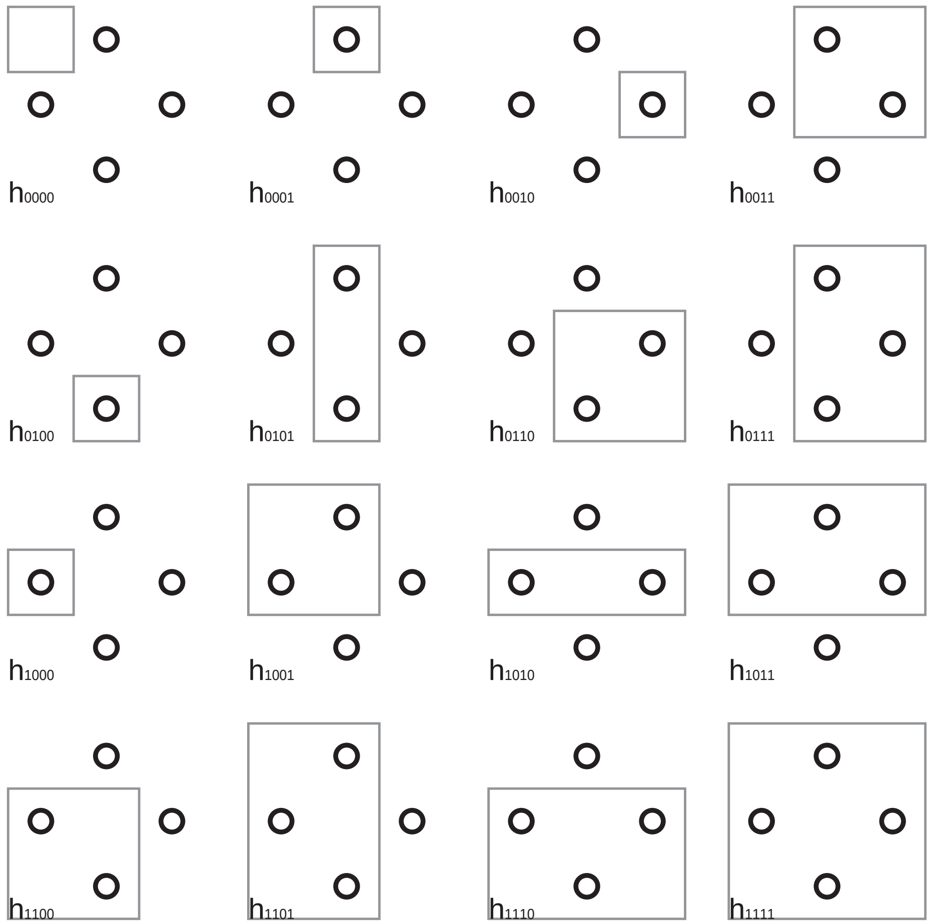

Figure 3: Representatives of $\mathcal{H}$ which shatter S when S is a set of 4 points.

Thus $VCdim(\mathcal{H}) \geq 4$, now we need to show that $VCdim(\mathcal{H}) < 5$.

Suppose $|S| \geq 5$. If you take the leftmost, rightmost, topmost, and bottommost points of S, there is at least one other point, and it must logically be inside. As such, no rectangle can label the leftmost, rightmost, topmost, and bottommost points of S positive without also labeling the interior point positive.

2.4 The VC Dimension of Finite Hypothesis Spaces

Since each hypothesis corresponds to precisely one dichotomy of S, the number of dichotomies of S is less than or equal to $|\mathcal{H}|$. Furthermore, since a shattered S requires $2^{|S|}$ dichotomies,

$$2^{|VCdim(\mathcal{H})|} \leq |\mathcal{H}| \tag{26}$$

So,

$$|VCdim(\mathcal{H})| \leq \lg |\mathcal{H}| \tag{27}$$

2.5 Sauer's Lemma

Sauer's Lemma states that,

$$\Pi_d(m) \leq \Phi_d(m) \tag{28}$$

Where:

$$d \equiv VCdim(\mathcal{H}) \tag{29}$$

$$\Phi_d(m) \equiv \sum_{i=0}^{d} \binom{m}{i} \tag{30}$$

2.6 The Proof of Sauer's Lemma

Note that it is a common convention that, $\binom{n}{k} \equiv 0$ if $k < 0$ or $k > n$. In our proof, we shall also use the following proposition, which turns out to be true even with the aforementioned convention:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \tag{31}$$

The following proof will be done by induction on $(m + d)$:

Base Cases:

Whenever $d = 0$, $\mathcal{H}$ can't even shatter an S of one point. Thus all $h \in \mathcal{H}$ label all points the same way (whether it be positive or negative). Thus, all the h are identical and $|\mathcal{H}| = 1$. So regardless of m, $\Pi_{\mathcal{H}}(m) = 1 = \binom{m}{0} = \Phi_0(m)$.

On the other hand, whenever $m = 0$, there is only one way to label a set of 0 examples. Thus, regardless of $\mathcal{H}$, $\Pi_{\mathcal{H}}(0) = 1 = \binom{0}{0} + \binom{0}{1} + \ldots + \binom{0}{d} = \Phi_d(0)$.

Induction Hypothesis:

Assume the lemma to be true for all m' and d' in which $m' + d' < m + d$.

Induction Step:

Let us work on m sample points, $S = \{x_1, x_2, \ldots, x_m\}$, with a hypothesis space $\mathcal{H}$ of VC dimension d, $VCdim(\mathcal{H}) = d$. For convenience, let $S_{\backslash m} = \{x_1, x_2, \ldots, x_{m-1}\}$.

We define two new (finite) hypotheses spaces, $\mathcal{H}_1$ and $\mathcal{H}_2$, in the following manner:

$$
\begin{aligned}
\mathcal{H}_0 &\equiv \{\text{one representative from } \mathcal{H} \text{ for each dichotomy over } S\} && (32)\\
\mathcal{H}_1 &\equiv \{\text{one representative from } \mathcal{H}_0 \text{ for each dichotomy over } S_{\setminus m}\} && (33)\\
\mathcal{H}_2 &\equiv \mathcal{H}_0 - \mathcal{H}_1 && (34)
\end{aligned}
$$

Take, for example, some $\mathcal{H}$ which contains the dichotomies given by the $\mathcal{H}_0$ column of the table below, where $m = 4$. The following table illustrates the procedure (hypotheses are identified by their dichotomies for the sake of readability):

$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_2$
01100	$\longrightarrow$	0110		
01101	$\longrightarrow$	~~0110~~	$\longrightarrow$	0110
01110	$\longrightarrow$	0111		
10100	$\longrightarrow$	1010		
10101	$\longrightarrow$	~~1010~~	$\longrightarrow$	1010
11001	$\longrightarrow$	1100		

So for $\mathcal{H}_1$ over $S_{\setminus m}$, $m_1 = m - 1$ (because S is one smaller than $S_{\setminus m}$) and $d_1 = VCdim(\mathcal{H}_1) \leq d$ (because reducing the number of hypotheses certainly will not increase the VC dimension of a space).

Similarly, with $\mathcal{H}_2$ over $S_{\setminus m}$, $m_2 = m - 1$ and $d_2 = VCdim(\mathcal{H}_2) \leq d - 1$. Let us explain $d - 1$: By construction, if $S' \subseteq S_{\setminus m}$ is shattered by $\mathcal{H}_2$, then every dichotomy over S' must occur both in $\mathcal{H}_1$ and $\mathcal{H}_2$ but with different labelings of x_m. Thus, $S' \cup \{x_m\}$, which has size $|S'| + 1$, is shattered by $\mathcal{H}$, and so $|S'|$ cannot be more than $d - 1$.
Using induction, $\Pi_{\mathcal{H}_1}(S_{\setminus m}) \leq \Phi_d(m-1)$ and $\Pi_{\mathcal{H}_2}(S_{\setminus m}) \leq \Phi_{d-1}(m-1)$.

Now, by the construction of $\mathcal{H}_1$ and $\mathcal{H}_2$, $\Pi_{\mathcal{H}}(S) = |\mathcal{H}_1| + |\mathcal{H}_2| = \Pi_{\mathcal{H}_1}(S_{\setminus m}) + \Pi_{\mathcal{H}_2}(S_{\setminus m})$. So, using our inequalities along with the convention and proposition put forth at the beginning of this subsection,

$$
\begin{aligned}
\Pi_{\mathcal{H}}(S) &= \Pi_{\mathcal{H}_1}(S_{\setminus m}) + \Pi_{\mathcal{H}_2}(S_{\setminus m}) && (35)\\
&\leq \Phi_d(m-1) + \Phi_{d-1}(m-1) && (36)\\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (37)\\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} && (38)\\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] && (39)\\
&= \sum_{i=0}^{d} \binom{m}{i} && (40)\\
&= \Phi_d(m) && (41)
\end{aligned}
$$

2.7 Sauer's Lemma and Theorem 3.2

We note that:

$$\begin{aligned}
\Phi_d(m) &= \sum_{i=0}^{d} \binom{m}{i} && (42)\\
&= \sum_{i=0}^{d} \frac{m!}{i! \cdot (m-i)!} && (43)\\
&= \sum_{i=0}^{d} \frac{(m-0)(m-1)(m-2)\ldots(m-i+1)}{i!} && (44)\\
&= O(m^d) && (45)
\end{aligned}$$

Thus, $\Pi_d(m) \leq O(m^d)$, and we have just made Theorem 3.2 useful.