COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Aaron Schild

Lecture # 6
February 21, 2013

Last class, we discussed an analogue for Occam's Razor for infinite hypothesis spaces that, in conjunction with VC-dimension, reduced the problem of finding a good PAC-learning algorithm to the problem of computing the VC-dimension of a given hypothesis space. Recall that VC-dimesion is defined using the notion of a *shattered set*, i.e. a subset S of the domain such that $\Pi_{\mathcal{H}}(S) = 2^{|S|}$. In this lecture, we compute the VC-dimension of several hypothesis spaces by computing the maximum size of a shattered set.

1 Example 1: Axis-aligned rectangles

Not all sets of four points are shattered. For example the following arrangement is impossible:

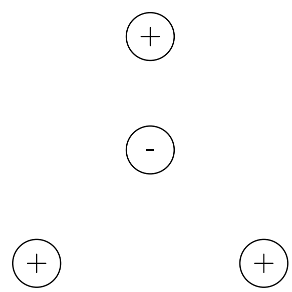

Figure 1: An impossible assignment of +/- to the data, as all rectangles that contain the outer three points (marked +) must also contain the one − point.

However, this is not sufficient to conclude that the VC-dimension is at most three. Note that the following set does shatter:

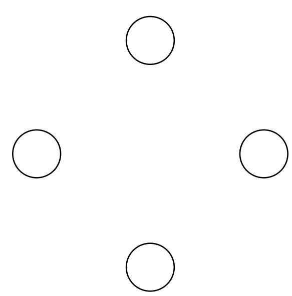

Figure 2: A set of four points that shatters, as there is an axis-aligned rectangle that contains any given subset of the points but contains no others.

Therefore, the VC-dimension is at least four. In fact, it is exactly four. Consider any set of five distinct points $\{v_1, v_2, v_3, v_4, v_5\} \subseteq \mathbb{R}^2$. Consider a rectangle that contains the points with maximum x-coordinate, minimum x-coordinate, maximum y-coordinate, and minimum y-coordinate. These points may not be distinct. However, there are at most four such points. Call this set of points $S \subset \{v_1, v_2, v_3, v_4, v_5\}$. Any axis-aligned rectangle that contains S must also contain all of the points v_1, v_2, v_3, v_4, and v_5. There is at least one v_i that is not in S, but still must be in the rectangle. Therefore, the labeling that labels all vertices in S with $+$ and v_i with $-$ cannot be consistent with any axis-aligned rectangle. This means that there is no shattered set of size 5, since all possible labelings of a shattered set must be realized by some concept.

By a similar argument, we can show that the VC-dimension of axis-aligned rectangles in $\mathbb{R}^n$ is $2n$. By generalizing the approach for proving that the VC-dimension of the positive half interval learning problem is 1, one can show that the VC-dimension of $n-1$ dimensional hyperplanes in $\mathbb{R}^n$ that pass through the origin is n. This concepts are inequalities of the form

$$\mathbf{w} \cdot \mathbf{x} > 0$$

for any fixed $\mathbf{w} \in \mathbb{R}^n$ and variable $\mathbf{x} \in \mathbb{R}^n$. In this case, concepts label points with $+$ if they are one side of a hyperplane and $-$ otherwise.

2 Other remarks on VC-dimension

In the cases mentioned previously, note that the VC-dimension is similar to the number of parameters needed to specify any particular concept. In the case of axis-aligned rectangles, for example, they are equal since rectangles require a left boundary, a right boundary, a top boundary, and a bottom boundary. Unfortunately, this similarity does not always hold, although it often does. There are some hypothesis spaces with infinite VC-dimension that can be specified with one parameter.

Note that if $\mathcal{H}$ is finite, the VC-dimension is at most $\log_2 |\mathcal{H}|$, as at least 2^r distinct hypotheses must exist to shatter a set of size r.

For a hypothesis space with infinite VC-dimension, there is a set of size m that is shattered for any $m > 0$. Therefore, $\Pi_{\mathcal{H}}(m) = 2^m$, which we mentioned last class as an indication of a class that is hard to learn. In the next section, we will show that all classes with bounded VC-dimension d have $\Pi_{\mathcal{H}}(m) = O(m^d)$, completing the description of PAC-learnability by VC-dimension.

3 Sauer's Lemma

Recall that $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ if $0 \leq k \leq n$ and $\binom{n}{k} = 0$ if $k < 0$ or $k > n$. k and n are integers and n is nonnegative for our purposes. Note that $\binom{n}{k} = O(n^k)$ when k is regarded as a positive constant. We will show the following lemma, which immediately implies the desired result:

Lemma 3.1 (Sauer's Lemma)**.** Let $\mathcal{H}$ be a hypothesis with finite VC-dimension d. Then,

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

Proof. We will prove this by induction on $m + d$. There are two base cases:

Case 1 ($m = 0$)**.** There is only one possible assignment of $+$ and $-$ to the empty set, i.e. $\Pi_{\mathcal{H}}(m) = 1$ here. Note that $\Phi_d(0) = \binom{0}{0} + \binom{0}{1} + \ldots + \binom{0}{d} = 1$, as desired.

Case 2 ($d = 0$)**.** Not even a single point can be shattered in this situation. Therefore, on any given point, all hypotheses have the same value. Therefore, there is only one possible hypothesis and $\Pi_{\mathcal{H}}(m) = 1$. This agrees with Φ, as $\Phi_0(m) = \binom{m}{0} = 1$.

Now, we will prove the induction step. For this, we will need Pascal's Identity, which states that

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$$

for all integers n and k with $n \geq 0$. Consider a hypothesis space $\mathcal{H}$ with VC-dimension d and a set of m examples $S := \{x_1, x_2, \ldots, x_m\}$. Let $T := \{x_1, x_2, \ldots, x_{m-1}\}$. Form two hypothesis spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ on T as follows (an example is in Figure 3). Let $\mathcal{H}_1$ be the set of restrictions of hypotheses from $\mathcal{H}$ to T. Let $h|_T$ denote the *restriction* of h to T for $h \in \mathcal{H}$, i.e. the function $h_T : T \to \{-, +\}$ such that $h_T(x_i) = h(x_i)$ for all $x_i \in T$. An element ρ on T is added to $\mathcal{H}_2$ if and only if there are two distinct hypotheses $h_1, h_2 \in \mathcal{H}$ such that $h_1|_T = h_2|_T = \rho$.

Note that $|\Pi_{\mathcal{H}}(S)| = |\Pi_{\mathcal{H}_1}(T)| + |\Pi_{\mathcal{H}_2}(T)|$. What are the VC-dimensions of $\mathcal{H}_1$ and $\mathcal{H}_2$? First, note that the VC-dimension of $\mathcal{H}_1$ is at most d, as any shattering set of size $d+1$ in T is also a subset of S that is shattered by the elements of $\mathcal{H}$, contradicting the fact that the VC-dimension of $\mathcal{H}$ is d.

Suppose that there is a set of size d in T that is shattered by $\mathcal{H}_2$. Since every hypothesis in $\mathcal{H}_2$ is the restriction of two different hypotheses in $\mathcal{H}$, x_m can be added to the shattered set of size d in T to obtain a set shattered by $\mathcal{H}$ of size $d+1$. This is a contradiction, so the VC-dimension of $\mathcal{H}_2$ is at most $d-1$. By the inductive hypothesis, $\Pi_{\mathcal{H}_1}(m-1) \leq \Phi_d(m-1)$. Similarly, $\Pi_{\mathcal{H}_2}(m-1) \leq \Phi_{d-1}(m-1)$. Combining these two inequalities shows that

$$\begin{aligned}
\Pi_{\mathcal{H}}(m) &\leq \Phi_d(m-1) + \Phi_{d-1}(m-1) \\
&= \left(\sum_{i=0}^{d} \binom{m-1}{i}\right) + \left(\sum_{j=0}^{d-1} \binom{m-1}{j}\right) \\
&= \binom{m-1}{0} + \sum_{i=0}^{d-1} \left(\binom{m-1}{i} + \binom{m-1}{i+1}\right) \\
&= \binom{m}{0} + \sum_{i=0}^{d-1} \binom{m}{i+1} \\
&= \Phi_d(m)
\end{aligned}$$

completing the inductive step. □

Often, the polynomial $\Phi_d(m)$ is hard to work with. Instead, we often use the following result:

Lemma 3.2. $\Phi_d(m) \leq (em/d)^d$ when $m \geq d \geq 1$.

Proof. $m \geq d \geq 1$ implies that $\frac{d}{m} \leq 1$. Therefore, since $i \leq d$ in the summand,

$\mathcal{H}$	x_1	x_2	x_3	x_4	x_5
	0	1	1	0	0
	0	1	1	0	1
	0	1	1	1	0
	1	0	0	1	0
	1	0	0	1	1
	1	1	0	0	1

$\mathcal{H}_1$	x_1	x_2	x_3	x_4
	0	1	1	0
	0	1	1	1
	1	0	0	1
	1	1	0	0

$\mathcal{H}_2$	x_1	x_2	x_3	x_4
	0	1	1	0
	1	0	0	1

Figure 3: The construction of $\mathcal{H}_1$ and $\mathcal{H}_2$

$$\begin{aligned}\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} &\leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \\ &= \left(1 + \frac{d}{m}\right)^m \\ &\leq e^d\end{aligned}$$

Multiplying on both sides by $(m/d)^d$ on both sides gives the desired result. $\square$

Plugging this result into the examples bound proven last class shows that

$$err(h) = O\left(\frac{1}{m}\left(d \ln \frac{m}{d} + \ln \frac{1}{\delta}\right)\right)$$

We can also write this in terms of the number of examples required to learn:

$$m = O\left(\frac{1}{\epsilon}(\ln 1/\delta + d \ln 1/\epsilon)\right)$$

Note that the number of examples required to learn scales linearly with the VC-dimension.

4 Lower bounds on learning

The bound proven in the previous section shows that the VC-dimension of a hypothesis space yields an upper bound on the number of examples needed to learn. Lower bounds on the required number of examples also exist. If the VC-dimension of a hypothesis space is d, there is a shattered set of size d. Intuitively, any hypothesis learned from a subset of size at most $d-1$ cannot predict the value of the last element with probability better than $1/2$. This suggests that at least $\Omega(d)$ examples are required to learn.

In future classes, we will prove the following

Theorem 4.1. For all learning algorithms A, there is a concept $c \in \mathcal{C}$ and a distribution D such that if A is given $m \leq d/2$ examples labeled by c and distributed according to D, then

$$\mathbf{Pr}[\mathrm{err}(h_A) > 1/8] \geq \frac{1}{8}$$

One can try to prove this as follows. Choose a uniform distribution D on examples $\{z_1, \ldots, z_d\}$ and run A on $m \le d/2$ examples. Call this set of examples S. Label the elements of S arbitrarily with + and -. Suppose that $c \in \mathcal{C}$ is selected to be consistent with all of the labels on S and $c(x) \ne h_A(x)$ for all $x \notin S$. $\mathrm{err}_D(h_A) \ge \frac{1}{2}$ since c agrees with h_A on at most $(d/2)/2 = 1/2$ of the probability mass of the domain, which means that there is no PAC-learning algorithm on $d/2$ examples.

This proof is flawed, as c needs to be chosen before the examples. We will discuss a correct proof in future classes.