COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Zeyu Jin

Lecture #6
February 20, 2014

1 VC Dimension

Last time we proved the theorem that with high probability $1-\delta$, the generalization error is given by

$$\operatorname{err}(h) \leq \mathrm{O}\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln 1/\delta}{m}\right) \tag{1}$$

where $\Pi_{\mathcal{H}}(m)$ is the growth function. We also defined the concept of shattering where S is shattered by $\mathcal{H}$ if $|\Pi_{\mathcal{H}}(S)| = 2^{|S|}$. Finally, we defined VC-dimension, or Vapnik-Chervonenkis dimension, as the cardinality of the largest shattered set. In this lecture, we are going to derive the bounds of the growth function, which is either $O(m^d)$ or 2^m.

1.1 Examples of VC-dimension

Here are some general results:

- VC-dim(intervals) $= 2$
- VC-dim(Axis-aligned rectangles) $= 4$
- VC-dim(Hyper-rectangles in $\mathbb{R}^n$) $= 2n$
- VC-dim(LTF in $\mathbb{R}^n$) $= n+1$

Note that LTF means linear threshold function (or perceptron), which is defined as a half space with parameters $\mathbf{w}$ where every points in this space is defined as "+". Formally,

$$c_w(x) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x} \geq b \\ 0 & \text{if } \mathbf{w}\cdot\mathbf{x} < b \end{cases} \tag{2}$$

The dot sign means inner product. If b is forced to be 0, the VC-dimension reduces to n. It is often the case that the VC-dimension is equal to the number of free parameters of a concept (for example, a rectangle's parameters are its topmost, bottommost, leftmost and rightmost bounds, and its VC-dimension is 4). However, it is not always true; there exists concepts with 1 parameter but an infinite VC-dimension.

There is also an inequality relationship between VC-dimension and the cardinality of $\mathcal{H}$. If the VC-dimension is d, then there exists a shattered set of size d on which $\mathcal{H}$ realizes all possible labelings. Because for every labeling there must be a corresponding hypothesis, we have $|\mathcal{H}| \geq 2^d$, which gives us:

$$\text{VC-dim}(\mathcal{H}) \leq \lg|\mathcal{H}| \tag{3}$$

1.2 Determining VC-dimension

In the last section, we claimed VC-dim(Axis-aligned rectangles) = 4. Now we show how to prove it. The proof involves two steps: first, we show the VC-dimension is at least 4 by showing that there exists a 4-point set shattered by the concept set (it's worth noting that not every 4-point configuration can be shattered, but we only need one to make the statement). Then, we show that there is no 5-point set that can be shattered.

Proof (1) An example 4-point set is shown in Figure 1 with all typical labelings and the corresponding realization. So we have VC-dim$\geq$ 4.

(2) For any 5-point set, we can construct a data assignment in this way: pick the topmost, bottommost, leftmost and rightmost points and give them the label "+". Because there are 5 points, there must be at least one point left to which we assign "$-$". Any rectangle that contains all the "+" points must contains the "$-$" point, which is a case where shattering is not possible. This proves that VC-dim$<$ 5.

In sum, VC-dim(axis aligned rectangle)$=$ 4.

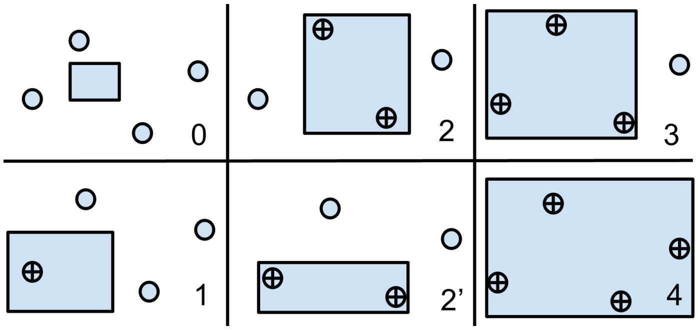


Figure 1: Proving that rectangle concept space shatters at least 4 points

2 Sauer's Lemma

Sauer's Lemma provides an upper bound for $\Pi_{\mathcal{H}}(m)$ parameterized by d, the VC-dimension of $\mathcal{H}$. It also leads to the proof that the growth function is either $O(m^d)$ or 2^m. In this section, we are going to use these definition and facts in binomial coefficients:

$$\binom{m}{k} = 0 \text{ if } k < 0 \text{ or } k > m \tag{4}$$

$$\binom{m}{k} = \binom{m-1}{k-1} + \binom{m-1}{k} \tag{5}$$

$$(a+b)^m = \sum_{k=0}^{m} \binom{m}{k} a^k b^{m-k} \tag{6}$$

Lemma 2.1 (Sauer's Lemma) *Let $\mathcal{H}$ be a hypothesis set with VC-dim($\mathcal{H}$) $= d$. Then, for all $m \in \mathcal{N}$, the following inequality holds*

$$\Pi_{\mathcal{H}}(m) \leq \Phi_d(m) \equiv \sum_{i=0}^{d} \binom{m}{i} \tag{7}$$

Proof The proof is by induction on $m + d$. The base cases are as follows:

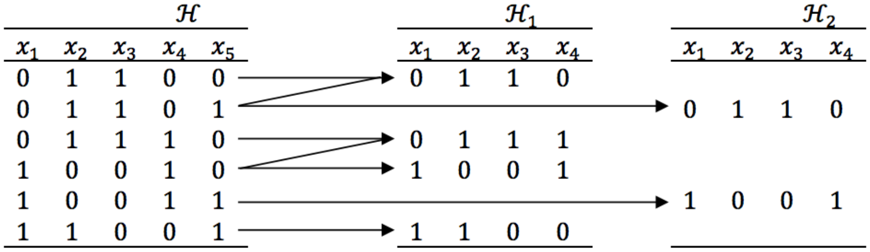
- When $d = 0$, for any m points, there is only a single label possible for every point in the space. So in this case $\Pi_{\mathcal{H}}(m) = 1 = \binom{m}{0} = \Phi_0(m)$.
- When $m = 0$, for any d, there is only one labeling. So $\Pi_{\mathcal{H}}(0) = 1 = \sum_{i=0}^{d} \binom{0}{i} = \Phi_d(0)$

When $m \geq 1$ and $d \geq 1$, assume the lemma holds for any m' and d' if $m' + d' < m + d$. Suppose $S = \{x_1, ..., x_m\}$; we now prove $\Pi_{\mathcal{H}}(S) \leq \Phi_d(m)$. We start by creating two other hypothesis spaces: first, we construct $\mathcal{H}_1$ by restricting the set of concepts in $\mathcal{H}$ to the set $S' = \{x_1, ..., x_{m-1}\}$. Figure 2 shows an example of the construction: suppose there is a concept in $\mathcal{H}$ maps $(x_1, x_2, x_3, x_4, x_5)$ to $(0, 1, 1, 0, 1)$, by restricting this concept on the domain (x_1, x_2, x_3, x_4), we create a new concept in $\mathcal{H}_1$ that maps (x_1, x_2, x_3, x_4) to $(0, 1, 1, 0)$.

In this construction, some pairs of concepts may collapse into single concepts in $\mathcal{H}_1$. The second hypothesis space $\mathcal{H}_2$ is obtained by including all these collapsed concepts in constructing $\mathcal{H}_1$. As illustrated in Figure 2, when concept $(0, 1, 1, 0, 1)$ and $(0, 1, 1, 0, 0)$ both collapse into a concept $(0, 1, 1, 0)$ in $\mathcal{H}_1$, we add another copy of $(0, 1, 1, 0)$ to $\mathcal{H}_2$. Note that both $\mathcal{H}_1$ and $\mathcal{H}_2$ are both defined on the domain $(x_1, ..., x_{m-1})$.

We now derive bounds on the size of these two new hypothesis spaces.

- Any subset $T \subseteq S$ shattered by $\mathcal{H}_1$ is also shattered by $\mathcal{H}$. So VC-dim$(\mathcal{H}_1) \leq$ VC-dim $(\mathcal{H}) = d$. By inductive hypothesis, $|\mathcal{H}| = |\Pi_{\mathcal{H}_1}(S')| \leq \Phi_d(m - 1)$
- Also notice that VC-dim$(\mathcal{H}_2) \leq d - 1$ since for any $T \subseteq S'$ shattered by $\mathcal{H}_2$, $T \cup \{x_m\}$ is shattered by $\mathcal{H}_1$. So $|\mathcal{H}_2| = |\Pi_{\mathcal{H}_2}(S')| \leq \Phi_{d-1}(m - 1)$



$\mathcal{H}$

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 2: Constructing $\mathcal{H}_1$ and $\mathcal{H}_2$ from $\mathcal{H}$: each table represents the content of hypothesis space; each row corresponds to a hypothesis and each row corresponds to a point x_i. The values are the labeling of a point given the row hypothesis. The arrow shows which hypothesis in $\mathcal{H}$ is used to construct a new hypothesis in $\mathcal{H}_1$ and $\mathcal{H}_2$.

In summary,

$$\begin{aligned}
|\Pi_{\mathcal{H}}(S)| &= |\mathcal{H}_1| + |\mathcal{H}_2| \leq \Phi_d(m-1) + \Phi_{d-1}(m-1) \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] && \text{(Equation 4)} \\
&= \sum_{i=0}^{d} \binom{m}{i} && \text{(Equation 5)}
\end{aligned}$$

which completes the proof. □

Now we show an upper bound of $\Phi_d(m)$: if $m \geq d \geq 1$, then

$$\Phi_d(m) \leq (\frac{em}{d})^d = O(m^d)$$

Proof According to the definition, $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$. We multiply $(\frac{d}{m})^d$ on both sides and thus,

$$\begin{aligned}
\left(\frac{d}{m}\right)^d \Phi_d(m) &\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^d \\
&\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i && (d/m \leq 1) \\
&= \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i 1^{m-i} \\
&\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i 1^{m-i} && (d \leq m) \\
&= \left(1 + \frac{d}{m}\right)^m \leq e^d && \text{(Binomial theorem, Equation 6)}
\end{aligned}$$

It follows that $\Phi_d(m) \leq (\frac{em}{d})^d$. □

Corollary 2.2 *Let $\mathcal{H}$ be a hypothesis space with VC-dim($\mathcal{H}$) $= d$. Then for all $m \geq d \geq 1$*

$$\Pi_{\mathcal{H}}(m) \leq (\frac{em}{d})^d = O(m^d)$$

The proof directly follows from Sauer's lemma where $\Pi_{\mathcal{H}}(m) \leq \Phi_d(m)$ and the fact we just proved. With this corollary, we can show that the growth function only exhibits two types of behavior: either VC-dim($\mathcal{H}$) $= d < \infty$, in which case $\Pi_{\mathcal{H}}(m) = O(m^d)$, or VC-dim($\mathcal{H}$) $= \infty$, in which case $\Pi_{\mathcal{H}}(m) = 2^m$ for all $m \geq 1$

Finally, we are able to express the generalization bound using VC-dimension:

Theorem 2.3 *Let $\mathcal{H}$ be a hypothesis space with VC-dim($\mathcal{H}$) $= d$. With probability at least $1-\delta$, for all $h \in \mathcal{H}$, if h is consistent with all m examples ($m \geq d \geq 1$) sampled independently from a distribution D, then the generalization error is*

$$err_D(h) \leq \frac{2}{m}\left(d \lg \frac{2em}{d} + \lg \frac{1}{\delta} + 1\right)$$

It directly follows from Corollary 2.2 and our earlier bound.

3 The lower bound?

VC-dimension also provides necessary conditions for learnability. In this sense, it is also possible to prove lower bounds on the number of examples needed to learn in the PAC model to a given accuracy. The difference is that instead of looking at the hypothesis space $\mathcal{H}$, we evaluate the VC dimension over the concept class $\mathcal{C}$.

Let's suppose we are working with concept class $\mathcal{C}$ and VC-dim($\mathcal{C}$) $= d$, which means there exists a set of size d, $\{z_1, ..., z_d\}$, shattered by concept class $\mathcal{C}$. A natural lower bound is d. The intuition is that even if we are given $z_1, ..., z_{d-1}$, we still lack the information conveyed by the last point; both labelings of the last point are still possible. To prove it rigorously, we need to go back to the definition of PAC learnable.

Claim 3.1 *For any algorithm A that PAC-learns the concept class $\mathcal{C}$, if given $d/2$ examples,then $err(h_A)$ has large generalization error with high probability.*

An incorrect proof is given below:

In the PAC learning setting, there is a target distribution D. To make things as bad as possible, we can choose whatever distribution we want. So we define D as a uniform distribution over the shattered set $z_1, ..., z_d$. Then we run some candidate algorithm A on $d/2$ samples chosen randomly from D. This algorithm will output hypothesis h_A. Pick $c \in \mathcal{C}$ which is consistent with the labels on the training set S. Let the remaining samples be labelled incorrectly, that is choose $c(x)$ so that $c(x) \neq h_A(x)$ for all $x \notin S$. Then err(h_A) is 1/2 since h_A misclassifies at least half the points in the shattered set.

This is wrong because the proof cheats by choosing the concept c after the training samples are selected (we can do that for h but not for c). We will show a correct proof next time.