Recap

Last time, we have the following theorem:

**Theorem.** With probability $\geq 1 - \delta$, $\forall h \in \mathcal{H}$ if $h$ is consistent with sample (of size $m$), then

$$\text{err}_D(h) \leq O\left(\frac{\ln \Pi_H(2m) + \ln \frac{1}{\delta}}{m}\right).$$

For any $\mathcal{H}$ we will see that only the following two cases are possible:

- $\Pi_H(m) = 2^m$, bad case
- $\Pi_H(m) = O(m^d)$, good case. In this case, we will have a generalization bound of the form:

$$\text{err}_D(h) \leq O\left(\frac{d \ln \frac{m}{d} + \ln \frac{1}{\delta}}{m}\right),$$

where PAC-learning is possible if we make $m$ large enough.

Today we will look into the combinatorial property of $\mathcal{H}$ and define VC-dimension. We will derive bounds on the growth function in terms of VC-dimension and show the above is true.

1 VC-dimension

We first introduce the concept of shattering before defining VC-dimension.

**Definition.** (Shattering). A set $S$ of size of $m$ is shattered by $\mathcal{H}$ if $|\Pi_H(S)| = 2^m$, i.e. all possible labelings of the set $S$ are realized by functions in $\mathcal{H}$.

**Definition.** (VC-Dimension). $\text{VC-dim}(\mathcal{H}) = \text{cardinality of the largest set shattered by } \mathcal{H}$.

Note: VC refers to Vapnik and Chervonenkis.

**Example.** (Intervals) For the case when $\mathcal{H} =$ intervals, it is illustrated in Figure 1 that $\mathcal{H}$ can shatter $S$ of 2 points but cannot shatter $S$ of 3 points. Therefore, $\text{VC-dim}(\text{intervals}) = 2$.

Note: we can see that, we need to show VC-dim is at least some number $d$ and then show that VC-dim is at most $d$ to draw the conclusion that VC-dim = $d$. To show VC-dim is at least $d$, we need to find just one set of $d$ points that are shattered (not for every set of $d$ points). To show VC-dim is at most $d$, we need to show every set of $d + 1$ points is not shattered.
Figure 1: Left: Case for 2 points that all labelings are realized. Right: For any three points, when the middle point has “−” label and the other two have “+” labels, this means the interval must contain all three points, which means it can not be shattered.

Figure 2: Left: A set of 4 points that can be shattered by axis-aligned rectangles. Right: For any 5-point set, we can choose the topmost, bottommost, leftmost and rightmost points and assign “+” to them, and the remaining point is assigned to “−”. Any rectangle that contains the “+” points must also contain “−”, which means this case cannot be shattered.

**Example.** (Axis-aligned Rectangles) For the case when $\mathcal{H} =$ axis-aligned rectangles, $VC$-$dim = 4$. (Illustrated in Figure 2)

**Example.** $VC$-$dim$(hyper rectangles in $\mathbb{R}^n$) $= 2n$.

**Example.** $VC$-$dim$(linear threshold functions in $\mathbb{R}^n$) $= n + 1$, where linear threshold function is defined to be

$$f(x) = \begin{cases} 1, & w \cdot x \geq b \\ 0, & \text{else} \end{cases}$$

**Example.** $VC$-$dim$(linear threshold functions through origin in $\mathbb{R}^n$) $= n$ ($b = 0$ here).

Note: in the above cases we see that often $VC$-$dim$ equals the number of parameters, but it is not always the case. For example, for the class of functions mapping real number $x$ to $\text{sign}(\sin(ax))$ with only one parameter $a$, its $VC$-$dim$ is infinite.
Claim. Consider the finite $H$ case, we have $d = \text{VC-dim}(H) \leq \lg |H|$. 

Proof. For VC-dim of size $d$, there must exist a shattered set of size $d$, meaning there are $2^d$ ways of labeling that set. For every labeling, there must be a corresponding hypothesis, therefore we must have $2^d \leq |H|$ for $H$ to shatter it.

1.1 Sauer’s Lemma

After introducing the concept of VC-dimension, we will now prove Sauer’s Lemma, which shows that the growth function $\Pi_H(m)$ is of $O(m^d)$ when VC-dim($H$) = $d$ is finite.

Lemma. (Sauer’s Lemma). Let $H$ be the hypothesis space, and $d = \text{VC-dim}(H)$, then $\Pi_H(m) \leq \Phi_d(m) := \sum_{i=0}^{d} \binom{m}{i}$.

Note: $\sum_{i=0}^{d} \binom{m}{i}$ is the number of ways of choosing at most $d$ items from set of size $m$.

Some facts:

• $\binom{m}{k} = \frac{m(m-1)\cdots(m-k+1)}{k!} = O(m^k)$. This implies that $\Phi_d(m) = O(m^d)$.

• $\binom{m}{k} = \binom{m-1}{k-1} + \binom{m-1}{k}$.

• $\binom{m}{k} = 0$, if $k < 0$ or $k > m$.

Proof. By induction on $m + d$. First, check base case:

• $m = 0$, there is only one labeling possible, $\Pi_H(m) = 1 = \sum_{i=0}^{d} \binom{0}{i} = \Phi_d(0)$.

• $d = 0$, there is only a single label possible for every point, $\Pi_H(m) = 1 = \binom{m}{0} = \Phi_0(m)$.

When $d \geq 1$, $m \geq 1$, assume lemma holds $\forall d', m'$, if $m' + d' < m + d$.

Fix a set $S = \langle x_1, \cdots, x_m \rangle$, we want to show $|\Pi_H(S)| \leq \Phi_d(m)$. Next, we define $H_1$ and $H_2$ on $S' = \langle x_1, \cdots, x_{m-1} \rangle$. Recall that $\Pi_H(S)$ is the set of distinct labellings $H$ induces on $S$. Define $H_1$ to consist of the set of distinct labelings $H$ induces on $S'$. Also, we add the labeling to $H_2$ whenever there is a “collapse” of labelings from $\Pi_H(S)$ to $H_1$, i.e. when there are two labelings in $\Pi_H(S)$ which are only different on $x_m$. A distinct labeling on $S'$ can be regarded as a hypothesis on $S'$.

Illustration of constructing $H_1$ and $H_2$ is given in Figure 3. We can see that by restricting on $S' = \langle x_1, x_2, x_3, x_4 \rangle$, we construct $H_1$ by including all the different labelings on $S'$. In the construction, some pairs of labelings in $\Pi(H(S))$ collapse into a single labeling in $H_1$, for example, from $(0, 1, 1, 0, 0)$ and $(0, 1, 1, 0, 1)$ to $(0, 1, 1, 0)$, causing us to add $(0, 1, 1, 0)$ to $H_2$.

We have the observation that $|\Pi_H(S)| = |H_1| + |H_2|$. And we have the following claims:

Claim: $\text{VC-dim}(H_1) \leq d$.

If $T \subseteq S'$ is shattered by $H_1$, it is also shattered by $H$. We can see from the example in Figure 3, $\{x_1, x_4\}$ are shattered by $H_1$ and therefore also shattered in $\Pi_H(S)$.

Claim: $\text{VC-dim}(H_2) \leq d - 1$.

If $T \subseteq S'$ is shattered by $H_2$, $T \cup \{x_m\}$ is shattered by $H$. In the example in Figure 3, pick $\{x_2\}$ that is shattered by $H_2$, we observe that $\{x_2, x_5\}$ are shattered by $\Pi_H(S)$. 


<table>
<thead>
<tr>
<th>$\Pi_\mathcal{H}(S)$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mathcal{H}_1$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mathcal{H}_2$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 3: An example of how $\mathcal{H}_1$ and $\mathcal{H}_2$ are constructed

From the above two claims, we have $|\mathcal{H}_1| = |\Pi_{\mathcal{H}_1}(S')| \leq \Phi_d(m - 1)$ and $|\mathcal{H}_2| = |\Pi_{\mathcal{H}_2}(S')| \leq \Phi_{d-1}(m - 1)$. Therefore we have,

$$|\Pi_\mathcal{H}(S)| = |\mathcal{H}_1| + |\mathcal{H}_2| \leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \leq \sum_{i=0}^{d} \binom{m}{i} \left( \frac{d}{m} \right)^i \leq \sum_{i=0}^{m} \binom{m}{i} \left( \frac{d}{m} \right)^i 1^{m-i} \leq \left( 1 + \frac{d}{m} \right)^m \leq e^d,$$

where (1) is because $0 < \frac{d}{m} \leq 1, i \leq d$ and (2) comes from binomial expansion. We then have $\Phi_d(m) \leq \left( \frac{em}{d} \right)^d$. □

Next, we will show an upper bound of $\Phi_d(m)$, which can be used to plug into the Theorem mentioned in the beginning and derive generalization bound for $\mathcal{H}$ with finite VC-dim $d$.

Claim. $\Phi_d(m) \leq \left( \frac{em}{d} \right)^d$, if $m \geq d \geq 1$.

Proof.

$$\left( \frac{d}{m} \right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \binom{m}{i} \left( \frac{d}{m} \right)^i \leq \sum_{i=0}^{m} \binom{m}{i} \left( \frac{d}{m} \right)^i 1^{m-i} \leq \left( 1 + \frac{d}{m} \right)^m \leq e^d,$$
From Sauer’s lemma and the above claim, we know there are only two cases for the growth function:

- \( \text{VC-dim}(\mathcal{H}) = d, \Pi_{\mathcal{H}}(m) = O(m^d) \).
- \( \text{VC-dim}(\mathcal{H}) = \infty, \Pi_{\mathcal{H}}(m) = 2^m \).

Plugging the result of Sauer’s Lemma into the Theorem mentioned at the beginning of the class, we have

\[
\text{err}_D(h) \leq O \left( \frac{d \ln \frac{m}{\delta} + \ln \frac{1}{\delta}}{m} \right).
\]

We can further turn it to a sample complexity bound (in other words, a bound on how much data \( m \) is needed to get error \( \epsilon \)) that is linear in \( d \), i.e. \( \text{VC-dim}(\mathcal{H}) \).