1 Overview

We spent the last two lecture learning about the growth function, VC dimension, the relationship between them, and the following theorem. In this lecture, we formally prove these results.\footnote{Several proofs of Theorem 1.1 are known in the literature. The proof approach we cover is similar to that outlined by Robert Schapire’s lecture notes.}

**Theorem 1.1 (PAC Learnability of Infinite Concept Classes).** Let $A$ be an algorithm that learns a concept class $C$ in the consistency model. Then, $A$ learns the concept class $C$ in the PAC learning model using a number of samples that satisfies

$$m \geq \frac{2}{\epsilon} \left( \log_2(\Pi_C(2m)) + \log_2\left(\frac{2}{\delta}\right) \right).$$

2 Proof of Theorem 1.1

In this lecture, we define and work with three “bad” events. First, is the actual failure event, as a function of the training set $S \sim D^m$, we would like to bound:

$$B(S) : \exists h \in C \text{ such that } \text{err}_S(h) = 0 \text{ and } \text{err}_D(h) > \epsilon.$$  

Second, for the sake of analysis we also consider an independently drawn sample set $S' \sim D^m$. We define the following event that is a function of $S$ and $S'$.

$$B'(S, S') : \exists h \in C \text{ such that } \text{err}_S(h) = 0 \text{ and } \text{err}_{S'}(h) > \epsilon/2.$$  

Lastly, given two sample sets $S = \{x_1, \ldots, x_m\}$ and $S' = \{x'_1, \ldots, x'_m\}$, and a vector $\bar{\sigma} \in \{-1, +1\}^m$, we swap the members of $S$ and $S'$ as follows: For each $i \in [m]$, if $\sigma_i = +1$, we let $z_i = x_i$ and $z'_i = x'_i$, otherwise, we let $z_i = x'_i$ and $z'_i = x_i$. Then, let $T = \{z_1, \ldots, z_m\}$ and $T' = \{z'_1, \ldots, z'_m\}$. Given $S$, $S'$, and $\bar{\sigma}$, we define the following bad event:

$$B''(S, S', \bar{\sigma}) : \exists h \in C \text{ s.t., } \text{err}_T(h) = 0 \text{ and } \text{err}_{T'}(h) > \epsilon/2, \text{ where } T \text{ and } T' \text{ correspond to } S, S', \bar{\sigma}.$$
When representing the probability of these events, we typically take $S \sim \mathcal{D}^m$, $S' \sim \mathcal{D}^m$, and $\sigma_i = +1$ or $-1$ with probability $1/2$ for all $i \in [m]$, all independently. When it is clear from the context, we suppress $S$, $S'$, and $\sigma$ in the statement of the probabilities.

To prove Theorem 1.1, it suffices to show that $\Pr_{S \sim \mathcal{D}}[B(S)] \leq \delta$. We do this by first bounding the probability of event $B$ in terms of $B'$ and then in terms of $B''$. We then argue that because $B''$ only depends on the empirical error on $T$ and $T'$ and not the true error, we can union bound only on the number of unique labelings produced on $T$ and $T'$, which is bounded by the growth function.

**Claim 2.1.** If $m > \frac{8}{\epsilon}$, then

$$\Pr_{S, S' \sim \mathcal{D}^m}[B'(S, S') \mid B(S)] \geq \frac{1}{2}$$

**Proof.** Suppose $B(S)$ holds. Then take an $h$ that is consistent with $S$, i.e., $\text{err}_S(h) = 0$, and $\text{err}_\mathcal{D}(h) > \epsilon$. Since $S'$ is drawn i.i.d. from $\mathcal{D}$,

$$\mathbb{E}_{S' \sim \mathcal{D}^m}[\text{err}_{S'}(h)] = \text{err}_\mathcal{D}(h) > \epsilon.$$ 

Furthermore, $\text{err}_{S'}(h)$ is the sample average of $m$ i.i.d. bernoulli variables. Recall that Chernoff bound states that for $X_1, \ldots, X_m$ bernoulli random variables with expectation $\mu$,

$$\Pr\left[\frac{1}{m} \sum_{i \in [m]} X_i \leq \frac{\mu}{2}\right] \leq \exp\left(-\frac{m\mu}{8}\right)$$

Replacing $\mu > \epsilon$, we have that $\Pr[\text{err}_{S'}(h) \leq \epsilon/2] \leq \frac{1}{2}$. This proves the claim. \qed

Note that Claim 2.1 immediately implies that $\Pr_{S \sim \mathcal{D}^m}[B(S)] \leq \Pr_{S, S' \sim \mathcal{D}^m}[B'(S, S')]$, because

$$\frac{\Pr[B'(S, S')]}{\Pr[B(S)]]} \geq \frac{\Pr[B'(S, S') \cap B(S)]}{\Pr[B(S)]} = \Pr[B'(S, S') \mid B(S)].$$

Therefore, it suffices to bound $\Pr_{S, S' \sim \mathcal{D}^m}[B'(S, S')]$.

**Claim 2.2.** For i.i.d. sample sets $S \sim \mathcal{D}^m$ and $S' \sim \mathcal{D}^m$, and a vector $\bar{\sigma}$, where $\sigma_i = +1$ or $-1$ with probability $1/2$ for all $i \in [m]$ independently, we have

$$\Pr_{S, S'}[B'(S, S')] = \Pr_{S, S', \bar{\sigma}}[B''(S, S', \bar{\sigma})].$$

**Proof.** This is true because $(T, T')$ and $(S, S')$ are identically distributed. \qed

**Claim 2.3.** For any $S, S' \in \mathcal{X}^m$ and any $h$ that is fixed (independently of $\bar{\sigma}$), we have

$$\Pr_{\bar{\sigma}}[\text{err}_T(h) = 0 \text{ and } \text{err}_{T'}(h) > \epsilon/2 \mid S, S'] \leq 2^{-m\epsilon/2}$$
Proof. Consider the predictions of \( h \) on \( S \) and \( S' \) as follows.

\[
\begin{align*}
&h(x_1), h(x_2), \ldots, h(x_m) \\
&h(x'_1), h(x'_2), \ldots, h(x'_m)
\end{align*}
\]

First, note that if there is a column with both predictions wrong then \( \text{err}_T(h) = 0 \) can never happen, and the desired probability would be 0. Similarly, if more than \((1 - \frac{\epsilon}{2})m\) of the columns have both predictions right, \( \text{err}_T'(h) \leq \frac{\epsilon}{2} \), so again the desired probability would be 0. Thus, at least \( r \geq \frac{m\epsilon}{2} \) columns have one correct and one incorrect prediction. If \( \text{err}_T(h) = 0 \), it must happen that in all such columns, \( \sigma_i \) must ensure that the right prediction goes to \( T \) and the wrong one goes \( T' \). This happens with probability at most \( 2^{-r} \leq 2^{-m\epsilon/2} \).

\[\square\]

Claim 2.4. For any \( S, S' \in \mathcal{X}^m \),

\[
\Pr_{\vec{\sigma}} \left[ \exists h \in C, \text{err}_T(h) = 0 \text{ and } \text{err}_T'(h) > \frac{\epsilon}{2} \mid S, S' \right] \leq \Pi_C(2m)2^{-m\epsilon/2}
\]

Proof. Given a set \( S \), define \( C'(S) \subseteq C \) to be a set of size \( |C[S]| \) where we choose one (representative) hypothesis for each different labeling of \( C \) on \( S \).

\[
L.H.S = \Pr_{\vec{\sigma}} \left[ \exists h \in C, \text{err}_T(h) = 0 \text{ and } \text{err}_T'(h) > \frac{\epsilon}{2} \mid S, S' \right]
\]

\[
= \Pr_{\vec{\sigma}} \left[ \exists h \in C'(S \cup S'), \text{err}_T(h) = 0 \text{ and } \text{err}_T'(h) > \frac{\epsilon}{2} \mid S, S' \right]
\]

\[
\leq \sum_{h \in C'(S \cup S')} \Pr_{\vec{\sigma}} \left[ \text{err}_T(h) = 0 \text{ and } \text{err}_T'(h) > \frac{\epsilon}{2} \mid S, S' \right]
\]

\[
\leq \Pi_C(2m)2^{-m\epsilon/2} \quad \text{(Claim 2.3)}
\]

Here \( \square \)

Putting together Claims 2.1, 2.2, 2.3, and 2.4, it suffices to find \( m \) such that

\[
2\Pi_C(2m)2^{-m\epsilon/2} \leq \delta,
\]

this gives us \( m \geq \frac{2}{\epsilon} \left( \log_2(\Pi_C(2m)) + \log_2(\frac{2}{\delta}) \right) \).

### 3 Sauer’s Lemma

In the last lecture, we demonstrated the importance of the following lemma.

**Lemma 3.1** (Sauer’s Lemma). Consider any hypothesis class \( C \) and let \( d = \text{VCDim}(C) \). For all \( m \),

\[
\Pi_C(m) \leq \sum_{i=0}^{d} \binom{m}{i}.
\]
In this lecture, we derive the proof of this lemma.

**Proof of Sauer’s Lemma.** The following facts will be used in this proof:

**Fact 3.2.** \( \binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \)

**Fact 3.3.** \( \binom{m}{k} = 0 \), if \( k < 0 \) or \( k > m \).

We will prove Sauer’s Lemma by induction on \( m + d \). Let \( \Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} \).

**Base Cases**

- For \( m = 0 \) and all \( d \). \( \Pi_C(m) = 1 = \sum_{i=0}^{d} \binom{0}{i} = \Phi_d(m) \). This is a degenerate case, where we label the empty set.

- For \( d = 0 \) and all \( m \). \( \Pi_C(m) = 1 = \binom{m}{0} = \Phi_d(m) \). Not even shattering a point, so only one labeling is possible.

**Inductive steps** We assume that the lemma holds for any \( m' + d' < m + d \). We need to show that for any \( S \), \( |C[S]| \leq \Phi_d(m) \). To prove this, we construct two new hypothesis classes that are defined on one fewer instance and apply our induction hypothesis. Take any \( S = \{x_1, \ldots, x_m\} \) and let \( S' = \{x_1, \ldots, x_{m-1}\} \) be the domain of two new hypothesis classes \( C_1 \) and \( C_2 \).

Consider the predictions of \( h \in C \) on \( S \), by consider \( C[S] \). The labeling in \( C[S] \) are all unique and come in one of the following forms:

- **Pairs:** where there are \( h \) and \( h' \) such that, for all \( i \in [m - 1] \), \( h(x_i) = h'(x_i) \) and \( h(x_m) \neq h'(x_m) \). For these pairs, we construct a function \( g : \mathcal{X}' \to \mathcal{Y} \), that is defined similarly as \( h \) and \( h' \), except that it is not defined on \( x_{m} \). We add \( g \) to both \( C_1 \) and \( C_2 \).

- **Singleton:** For \( h \) where there is no \( h' \) that satisfies the pair condition. For these we construct a function \( g : \mathcal{X}' \to \mathcal{Y} \), that is the same as \( h \) except not defined on \( x_m \). We add \( g \) only to \( C_1 \).

Note that, the number of unique labelings in \( C \) is preserved, so \( |C[S]| = |C_1| + |C_2| \). See the following figure for an example of this construction.

<table>
<thead>
<tr>
<th>( C[S] )</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1, x_2, x_3, x_4, x_5 )</td>
<td>( x_1, x_2, x_3, x_4 )</td>
<td>( x_1, x_2, x_3, x_4 )</td>
</tr>
<tr>
<td>( h_1 )</td>
<td>0 1 1 1 0</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>( h_2 )</td>
<td>0 1 1 1 1</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>( h_3 )</td>
<td>1 0 0 1 1</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>( h_4 )</td>
<td>1 0 0 1 0</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>( h_5 )</td>
<td>1 1 1 0 0</td>
<td>1 1 1 0</td>
</tr>
</tbody>
</table>
Moreover, notice that if a set is shattered by \( C_1 \) then it is also shattered by \( C \) because each labeling in \( C[S] \) can be generated using the same labeling (while ignoring \( x_m \)) in \( C_1 \). So,

\[
\text{VCDim}(C_1) \leq \text{VCDim}(C) = d.
\]

Furthermore, if some set \( T \) is shattered by \( C_2 \), then \( T \cup \{x_m\} \) is shattered by \( C \). This is because every labeling in \( C_2 \) refers to two labelings in \( C \), where the labels on \( x_1, \ldots, x_{m-1} \) are the same and \( x_m \) is labeled in two different ways. Hence, \( \text{VCDim}(C) \geq \text{VCDim}(C_2) + 1 \), which implies

\[
\text{VCDim}(C_2) \leq d - 1.
\]

Now, by induction we have that \( |C_1| = |C_1[S']| \leq \Phi_d(m - 1) \) and \( |C_2| = |\Pi_{C_2}(m - 1) \leq \Phi_{d-1}(m - 1) \). We have

\[
|C[S]| = |C_1| + |C_2| \\
\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\
= \sum_{i=0}^{d} \binom{m}{i} \\
= \Phi_d(m).
\]

\( \square \)