Recap

Last lecture, we proved Occam’s Razor, which is that with probability at least $1 - \delta$, $\forall h \in \mathcal{H}$, if $h$ is consistent with all $m$ examples that are sampled independently from distribution $D$, then the generalization error $\text{err}_D(h) \leq \frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{m}$. However, this equation only applies to finite hypothesis spaces since we are using the cardinality of $\mathcal{H}$. This led us to briefly discuss about the generalization of Occam’s Razor to infinite hypothesis spaces at the end of last week’s lecture.

Sample Complexity for Infinite Hypothesis Space

In order to generalize Occam’s Razor to infinite hypothesis spaces, we have to somewhat replace the $|\mathcal{H}|$. Here we first introduce some new concepts and notations which would simplify the later proof and discussion.

$$S = \langle x_1, \ldots, x_m \rangle$$

(sample set)

$$\Pi_{\mathcal{H}}(S) = \{\langle h(x_1), \ldots, h(x_m) \rangle : h \in \mathcal{H}\}$$

(set of all possible dichotomies of $\mathcal{H}$ on $S$)

$$\Pi_{\mathcal{H}}(m) = \max_{S : |S| = m} |\Pi_{\mathcal{H}}(S)|$$

(growth function)

The growth function denotes the maximum number of distinct ways in which $m$ points can be classified using hypotheses in $\mathcal{H}$, which provides another measure of the complexity of the hypothesis set $\mathcal{H}$. We will prove later that $\forall \mathcal{H}$ either:

- $\Pi_{\mathcal{H}}(m) = 2^m$ (impossible for PAC, can’t get enough data)
- or
- $\Pi_{\mathcal{H}}(m) = \mathcal{O}(m^d)$ (possible for PAC)

Recall that our goal is to replace the cardinality of $|\mathcal{H}|$ in Occam’s Razor. It is now clear that a growth function with a form similar to $\Pi_{\mathcal{H}}(m)$ is a good candidate. Therefore, our goal is to modify Occam’s Razor to the following generalized version:

**Theorem:**

with probability at least $1 - \delta$, $\forall h \in \mathcal{H}$, if $h$ is consistent with all $m$ examples that are sampled independently from distribution $D$, then the generalization error $\text{err}_D(h) \leq \mathcal{O}\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln \frac{1}{\delta}}{m}\right)$.

Before the proof, we first introduce some definitions. Let $D$ denote our target distribution, and $S = \langle x_1, \ldots, x_m \rangle$ denote a sample of $m > 8/\epsilon$ points chosen independently from $D$. We also introduce a “ghost sample” $S' = \langle x'_1, \ldots, x'_m \rangle$ that consists of $m$ points drawn i.i.d.
from $D$. By creating this “ghost sample”, we are using the “double-sample trick” to take the mistakes on $S'$ as a proxy for a hypothesis’s generalization error. More importantly, doing so helps us avoid dealing with the potentially infinite space of instances, yet being able to make claims about a hypothesis. $S'$ is called the “ghost sample” because it never actually exists and is not provided to the learning algorithm. We also define:

- $M(h, S) =$ number mistakes $h$ makes on $S$
- $B \equiv \exists h \in \mathcal{H} : (h \text{ is consistent on } S) \land (err_D(h) > \epsilon)$
- $B' \equiv \exists h \in \mathcal{H} : (h \text{ is consistent on } S) \land (M(h, S') \geq \frac{m\epsilon}{2})$

**Proof**

Our goal is to prove that $Pr[B] \leq \delta$

**Step 1: $Pr[B'|B] \geq 1/2$**

In order to show this, suppose $B$ holds, which is that there exists $h$ consistent on $S$ and $err_D(h) > \epsilon$. Since $err_D(h) > \epsilon$, the expectation value of $M(h, S')$, which is simply the number of examples times the probability of making an error would be at least $m\epsilon$. By Chernoff bounds (to be discussed later in the course) we can show that $Pr[M(h, S') < \frac{m\epsilon}{2}] \leq \frac{1}{2}$. Therefore, we can conclude that $Pr[B'|B] \geq 1/2$.

**Step 2: $Pr[B] \leq 2Pr[B']$**

From $A \land B \Rightarrow A$, we can show that:

$$Pr[B'] \geq Pr[B \land B'] = Pr[B]Pr[B'|B] \quad \text{(by product rule)}$$

$$\geq \frac{1}{2}Pr[B] \quad \text{(by step 1)}$$

Now we have reduced the original problem to finding an upper bound for $Pr[B']$.

Now, consider two experiments to generate $S$ and $S'$

**Experiment 1:** Choose $S, S'$ as usual (i.i.d. from $D$)

**Experiment 2:** First choose $S, S'$ as usual (i.i.d. from $D$), but for $i \in \{1, 2, \cdots, m\}$ swap the example $x_i$ in $S$ with $x'_i$ in $S'$ with 0.5 probability and call the resulting samples as $T$ and $T'$.

Notice that $T, T'$ have the exact same distribution as $S, S'$ since they are drawn from i.i.d., so experiment 1 and experiment 2 are actually identical. Also, we define:

- $B'' \equiv \exists h \in \mathcal{H} : (h \text{ is consistent on } T) \land (M(h, T') \geq \frac{m\epsilon}{2})$

$$\equiv \exists h \in \mathcal{H} : (M(h, T) = 0) \land (M(h, T') \geq \frac{m\epsilon}{2})$$
**Step 3:** $Pr[B'''] = Pr[B']$

Because the distributions for $T$, $T'$ are exactly the same as those for $S$, $S'$, $Pr[B'''] = Pr[B']$.

Define $b(h) \equiv [h$ is consistent with $T$ and $M(h, T') \geq \frac{m\epsilon}{2}$]

**Step 4:** $Pr[b(h)|S, S'] \leq 2^{-m\epsilon/2}$

Let us identify each example $x$ in $S$ and $S'$ with a bit which is 0 if $h(x) = c(x)$ and 1 if $h(x) \neq c(x)$. In this step, we want to bound the probability of constructing a set $T$ that is consist of only example 0's and $T'$ that is consist of only example 1's given $S$ and $S'$ that is selected from the standard procedure (drawing i.i.d. from $D$). We denote this as $b(h)$:

$$b(h) \equiv (M(h, T) = 0) \land (M(h, T') \geq \frac{m\epsilon}{2})$$

Let $r$ denote the number of pairs of points from $S$ and $S'$ that has exactly one 1 labeled. $Pr[b(h)|S, S'] \leq 2^{-m\epsilon/2}$ can then be shown by the following three cases:

**Case 1:** $\exists x_i, x'_i$ with both of them labeled as 1

In this case, no matter how the examples in $S$ and $S'$ are swapped by experiment 2, there will always be an error in $T$. Therefore, $Pr[M(h, T) = 0] = 0$ $\Rightarrow$ $Pr[b(h)|S, S'] = 0$

\[
\begin{array}{cccccccc}
| & 1 & 1 & 0 & 0 & 1 & \\
S & | & 1 & 1 & 0 & 0 & 1 & \\
| & 0 & 1 & 0 & 1 & 0 & \\
S' & | & 0 & 1 & 0 & 1 & 0 & \\
\end{array}
\]

We can see from the above example that, no matter how the example in $S$ are swapped with the example in $S'$ below it, the minimum number of 1 labeled in $S$ will be 1 since there are two 1's in column 2.

\[
\begin{array}{cccccccc}
| & 0 & 1 & 0 & 0 & 0 & \\
S & | & 0 & 1 & 0 & 0 & 0 & \\
| & 1 & 1 & 0 & 1 & 1 & \\
S' & | & 1 & 1 & 0 & 1 & 1 & \\
\end{array}
\]

**Case 2:** $r < \frac{m\epsilon}{2}$

In this case $Pr[b(h)|S, S']$ is also 0. This is because in order for $b(h)$ to be true, all $r$ errors have to occur in $T'$ and the total number of errors labeled in $T'$ have to exceed $\frac{m\epsilon}{2}$, which is impossible since there is only one error in each pair in $r$ and $r < \frac{m\epsilon}{2}$.

**Case 3:** $r \geq \frac{m\epsilon}{2}$

Now, the total number of errors exceeds $\frac{m\epsilon}{2}$ so there is a probability that $b(h)$ is true. As mentioned above, experiment 2 would swap examples in $S$ and $S'$ with probability 0.5. Since these events are independent, $Pr[b(h)|S, S'] = (\frac{1}{2})^r \leq 2^{-m\epsilon/2}$.

Now, we can derive the bound of $Pr[b(h)|S, S']$ as follows:

$$Pr[b(h)|S, S'] \leq 2^{-m\epsilon/2}$$
Step 5: \( \Pr[B''|S, S'] \leq \Pi_H(2m)2^{-m\epsilon/2} \)

Let \( \mathcal{H}' \) denote the space of “representative” hypotheses for each labeling of \( S, S' \), which is finite. We can see that \( |\mathcal{H}'| = |\Pi_H(S, S')| \leq \Pi_H(2m) \).

We can now prove \( \Pr[B''|S, S'] \leq \Pi_H(2m)2^{-m\epsilon/2} \) as follows:

\[
\Pr[B''|S, S'] = \Pr[\exists h \in \mathcal{H} : b(h)|S, S'] \\
= \Pr[\exists h \in \mathcal{H} : b(h)|S, S'] \\
\leq \sum_{h \in \mathcal{H}'} \Pr[b(h)|S, S'] \\
\leq |\mathcal{H}'|2^{-m\epsilon/2} \\
\leq \Pi_H(2m)2^{-m\epsilon/2}
\]

Notice that the second step above is true because if \( b(h) \) holds for some \( h \in \mathcal{H} \), it will also hold for some \( h \in \mathcal{H}' \) since they behave the same on \( S \) and \( S' \) (\( b(h) \) only depends on \( S \) and \( S' \)).

Step 6: \( \Pr[B''] \leq \Pi_H(2m)2^{-m\epsilon/2} \)

By marginalization (\( \Pr[a] = \mathbb{E}_X[\Pr[a|X]] \)) we can show that:

\[
\Pr[B''] = \mathbb{E}_{S, S'}[\Pr[B''|S, S']] \\
\leq \Pi_H(2m)2^{-m\epsilon/2}
\]

By the above six steps, we can finally show that:

\[
\Pr[B''] \leq 2\Pr[B'] = 2\Pr[B''] \\
\leq 2\Pi_H(2m)2^{-m\epsilon/2} \\
\leq \delta
\]

Now, by solving the above inequality for \( \epsilon \), we can see that the inequality above holds when

\[
\epsilon \leq \frac{2}{m} (\ln \Pi_H(2m) + \ln \frac{1}{\delta} + 1) = O\left(\frac{\ln \Pi_H(2m) + \ln \frac{1}{\delta}}{m}\right)
\]

which is the error bound we are trying to prove for the generalized Occam’s Razor.

By replacing \( |\mathcal{H}| \) with the growth function \( \Pi_H(2m) \), we have now proved a bound on generalization in terms of the growth function. When the growth function has the form of \( O(m^d) \), we have a useful bound. We will later see when this form of growth function happens.
VC-Dimension

At the end of the class, we also briefly discussed the VC-dimension (Vapnik-Chervonenkis dimension). In order to define the VC-dimension of a hypothesis set $\mathcal{H}$, we first need to introduce the concept of “shattering”. A set $S$ of size $m$ is shattered by $\mathcal{H}$ if all labelings of $S$ can be realized by hypotheses in $\mathcal{H}$, that is when $|\Pi_\mathcal{H}(S)| = \Pi_\mathcal{H}(m) = 2^m$. And the VC-dimension of a hypothesis set $\mathcal{H}$ is the cardinality of the largest set that can be fully shattered by $\mathcal{H}$:

$$VCdim(\mathcal{H}) = \max\{m : \Pi_\mathcal{H}(m) = 2^m\}$$

We now look at an example of $\mathcal{H} = \{\text{intervals}\}$:

Figure 1: $\mathcal{H}$ contains hypotheses that produce every possible labeling of the 1 point in $S$. Therefore, $\mathcal{H}$ shatters $S$, $VCdim \geq 1$

Figure 2: $\mathcal{H}$ contains hypotheses that produce every possible labeling of the 2 points in $S$. Therefore, $\mathcal{H}$ shatters $S$, $VCdim \geq 2$
Figure 3: When $S$ is a set of 3 points, $\mathcal{H}$ does not contain a hypothesis that can label this situation. Therefore, $\mathcal{H}$ does not shatter $S$, $\text{VCdim} < 3$

We can see from the example that $\mathcal{H}$ shatters $S$ when $S$ contains a single point and two points, but not three. Therefore, $\text{VCdim}(\mathcal{H}) = 2$