Today we will cover the following 2 topics:

1. Learning infinite hypothesis class via VC-dimension and Rademacher complexity;

2. Introduction to unsupervised learning and density estimation.

1 Learning Infinite Hypothesis Class

In the first lecture, we showed that a hypothesis class is PAC-learnable if it is finite. What about infinite hypothesis class? First we give a simple example showing the possibility of PAC-learning an infinite hypothesis class.

Consider the family of threshold functions defined on the real line. In particular, let the domain $\mathcal{X} = \mathbb{R}$ and the label set $\mathcal{Y} = \{-1, 1\}$. A threshold function $f_{\theta} : \mathbb{R} \rightarrow \mathcal{Y}$ is defined as,

$$f_{\theta}(x) = \begin{cases} 
1 & \text{if } x \leq \theta \\
-1 & \text{otherwise}
\end{cases} \quad (1)$$

Given $m$ samples in the form $\{(x_i, y_i)\}_{i=1}^{m}$ where $y_i = f_{\theta}(x_i)$, there exists a separator $\hat{\theta} \in \mathbb{R}$ that divides the samples (i.e., for all samples labeled +1 we have $x_i \leq \hat{\theta}$, for those labeled -1 we have $x_i > \hat{\theta}$). We output the following hypothesis $h$ based on $\hat{\theta}$ and this defines our learning algorithm.

$$h(x) = \begin{cases} 
1 & \text{if } x \leq \hat{\theta} \\
-1 & \text{otherwise}
\end{cases} \quad (2)$$

We need to show that $\Pr_{x \sim D}[f(x) \neq h(x)] \leq \epsilon$ with high probability. Let $R$ be the interval between the rightmost +1 data point and the leftmost −1 data point. In other words, $R$ is the set of valid choices for $\hat{\theta}$. Note that $R$ is a random interval that depends on the samples. If $R$ is narrow enough, then $\hat{\theta}$ would be very close to the true $\theta$, implying a small error. In particular, one can see that if $\Pr_{x \sim D}[x \in R] \leq \epsilon$ then our algorithm works.
Choose $\theta_1$ and $\theta_2$ such that $\Pr_{x \sim D}[\theta_1 \leq x \leq \theta] = \epsilon$ and $\Pr_{x \sim D}[\theta \leq x \leq \theta_2] = \epsilon$. If we take $m$ samples, the probability that no sample is inside $[\theta_1, \theta]$ is equal to $(1 - \epsilon)^m$ and likewise for $[\theta, \theta_2]$. Therefore, if we choose $m \geq O(\frac{1}{\epsilon} \log \frac{1}{\delta})$, then with high probability we would have at least one sample inside both $[\theta_1, \theta]$ and $[\theta, \theta_2]$. This would imply $\Pr_{x \sim D}[R] \leq 2\epsilon$ and we are done.

1.1 VC-Dimension

Let $H$ be the hypothesis class over a domain $\mathcal{X}$. Assume $\mathcal{Y} = \{0, 1\}$. In the following, we might represent a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ by its support $\{x \in \mathcal{X} : h(x) = 1\}$.

**Definition 1** (Shattering). A subset $S \subseteq \mathcal{X}$ is shattered by $H$ if for all $T \subseteq S$, there exists $h \in H$ such that $h \cap S = T$ (where $h \cap S := \{x \in X : h(x) = 1\} \cap S$). The VC-dimension of $H$ is the size of the largest subset $S \subseteq \mathcal{X}$ that is shattered by $H$.

To show $H$ has VC-dimension $d$, we need to prove two things:

1. $\exists$ set $S$ with $|S| = d$ that is shattered by $H$;
2. No set $S$ with size $d + 1$ is shattered by $H$.

**Example 2.** Let $\mathcal{X} = \{1, 2, 3, 4, 5\}$. Let $h_1 = \{1, 2, 3\}$, $h_2 = \{2, 4, 5\}$, $h_3 = \{3, 4\}$, $h_4 = \{1, 2, 5\}$, $h_5 = \{1, 3, 5\}$ and $h_6 = \{5\}$.

One can check that $H$ shatters subset $S = \{2, 4\}$, so $VC(H) \geq 2$. In order to shatter a subset of size 3, you need at least $8 = 2^3$ hypotheses, so $VC(H) < 3$. Therefore, $VC(H) = 2$.

Also, we just proved $VC(H) \leq \log_2 |H|$.

**Example 3.** Let $\mathcal{X} = \mathbb{R}$ and $H$ = all closed intervals $[a, b]$. We will show that $VC(H) = 2$. Given any subset $S \subseteq \mathbb{R}$ of size 2, say $\{c, d\}$. We can choose $[c, c]$, $[d, d]$, $[c - 2, c - 1]$ to shatter $\{c, d\}$, $\{c\}$, $\{d\}$ and $\emptyset$, respectively. This proved $VC(H) \geq 2$. However, if one has 3 points $S = \{c, d, e\}$ where $c < d < e$, the subset $T = \{c, e\}$ cannot be shattered by any interval. So $VC(H) < 3$ and $VC(H) = 2$. Note that the family of all intervals is an infinite hypothesis class, and yet it has finite VC-dimension.

1.2 VC-Dimension as a Lower Bound

In this section, we lower bound learnability by VC-dimension.
Theorem 4. Let $H$ be any hypothesis class with $VC(H) = d$. Then any PAC-learner must use at least $\Omega(d^2)$ samples.

Proof. As a warm-up, we would prove this for constant $\epsilon$ and $\delta$. As $VC(H) = d$, let $S = \{x^1, x^2, \cdots, x^d\} \subseteq X$ be shattered by $H$. Let $D$ be the uniform distribution over $S$. Suppose our learner $A$ uses only $d^2$ samples, then $A$ knows at most $d^2$ values of $f(x^i)$ where $f$ is the target function. Let $H_S = \{h_1, h_2, \cdots, h_{2^d}\}$ be the $2^d$ functions that shatter $S$. Let $\mathcal{P}$ be the uniform distribution over $H_S$. Suppose that the target function $f$ is drawn from $\mathcal{P}$, it would be hard for $A$ to learn.

Fix any sample $T$ of size $d/2$, suppose $A$ output $h_T$. As there are at least $d/2$ unseen points from $S$, no matter how the (random) target function labels them $A$ would still output the same hypothesis. So on the unseen half of $S$, any algorithm would make at least $d/4$ mistakes in expectation. Then $E[\text{error}(h)] \geq \frac{1}{4}$, thus by Markov’s inequality $\Pr[\text{error}(h) < \frac{1}{8}] \leq \frac{6}{7}$.

It turns out that VC-dimension exactly characterizes learnability, whether the hypothesis class is infinite or not.

Theorem 5. The following statements are equivalent to binary classification.

1. $VC(H) = d$;
2. $H$ is PAC-learnable with $\frac{1}{\epsilon}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta})$ samples;
3. $H$ is agnostically PAC-learnable with $\frac{1}{\epsilon^2}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta})$ samples;
4. $H$ admit uniform convergence with $\frac{1}{\epsilon}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta})$ samples.

1.3 VC-dimension as an Upper bound

Consider $S \subseteq X$, let $\pi_H(S) = \{h \cap S : h \in H\}$, which is equal to the set of subsets of $S$ induced by $H$.

Example 6. Let $X = \mathbb{R}$, $H = \text{all intervals and } S = \{1, 2, 3\}$. $\pi_H(S) = 2^S - \{\{1, 3\}\}$

We are usually interested in the size of $\pi_H(S)$ rather than the set $\pi_H(S)$ itself.

Definition 7. The growth function $\pi_H(m) := \max_{S \subseteq X : |S| = m} |\pi_H(S)|$.

It is easy to see that $H$ shatters $S \iff |\pi_H(S)| = 2^{|S|}$, so $VC(H) = \text{largest } m \text{ such that } \pi_H(m) = 2^m$. In the worst case, the growth function $\pi_H(m)$ can grow exponentially in $m$, where $\pi_H(S)$ contains all possible subsets of $S$. However, with small VC-dimension, the growth function would grow only polynomially after a certain point. In particular, we have the following lemma.
Lemma 8 (Sauer’s Lemma). If $VC(H) = d$, then

$$
\pi_H(m) = \begin{cases} 
2^m & \text{if } m \leq d \\
O(m^d) & \text{otherwise}
\end{cases}
$$ (3)

In most cases, whenever union bound is applied over a set of hypothesis, one can replace it by a union bound over $\pi_H(m)$ many hypotheses, resulting in smaller sample complexity.

2 Rademacher Complexity

Recall the definition of a representative sample.

Definition 9. A sample $S = \{z_1, z_2, \ldots, z_m\}$ is $\epsilon$-representative (w.r.t domain $Z$, hypothesis class $H$ and loss function $l(h, z)$) if

$$
\sup_{h \in H} |L_D(h) - L_S(h)| \leq \epsilon,
$$ (4)

where $L_D(h) = E_{z \sim D}[l(h, z)]$ and $L_S(h) = E_{z \sim U(S)}[l(h, z)]$ ($U(S)$ is the uniform distribution over $S$).

For each hypothesis $h$, we can rewrite $l(h, z) = f_h(z)$ and $f_h : Z \rightarrow \mathbb{R}$. Let $F = \{f_h : h \in H\}$. Then

$$
Rep_D(F, S) = \sup_{f \in F} |L_D(f) - L_S(f)|.
$$ (5)

The problem is that we don’t know what the true distribution $D$ is, so we can split the training samples into 2 equal-size sets $S_1$ and $S_2$.

$$
Rep_D(F, S) \approx \sup_{f \in F} |L_{S_1}(f) - L_{S_2}(f)| = \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(z_i),
$$ (6)

where $\sigma_i = +1$ if $i \in S_1$ and $\sigma_i = -1$ otherwise.

Inspired by this observation, the Rademacher complexity of $F$ (w.r.t sample $S$) is defined as

$$
R_S(F) = \frac{1}{m} E_{\sigma_1}[\sup_{f \in F} \sum_{i=1}^{m} \sigma_i f(z_i)],
$$ (7)

where each $\sigma_i$ is an independent $\{-1, 1\}$ coin flip. The next lemma shows that the rate of uniform convergence is governed by Rademacher complexity.
Lemma 10.

\[
E_{S \sim D^m}[Rep_D(F, S)] \leq 2E_{S \sim D^m}[R_S(F)]
\]

As uniform convergence guarantees learnability of ERM, this implies an upper bound on the error of ERM learner.

3 Unsupervised Learning

In this section, we introduced an important unsupervised learning problem called ‘density estimation’.

Definition 11. Let \( F \) be a family of probability distribution. Given i.i.d samples from an unknown distribution \( p \in F \), output \( h \in F \) so that \( h \) is ‘close’ to \( p \) with high probability.

We have been vague about what ‘closeness’ means in the above definition and different notions of closeness will lead to different density-estimation problems.

3.1 Most basic setting

Here we consider what might be the most simple density estimation problem: learning discrete distribution under total variation distance.

Let \( F \) be the family of all distribution over \([n]\). The total variation distance is defined as \( d_{TV}(p, q) = \max_{A \subseteq S} |p(A) - q(A)| = \frac{1}{2} ||p - 1||_1 \).

Similar to the Empirical Risk Minimization learner, we can output the empirical histogram. In particular, let \( h_S(i) = \frac{|\{j \in [m] : s_j = i\}|}{m} \). Next we will discuss the performance of this empirical-histogram learner.

Theorem 12. Learning a discrete distribution over \([n]\) requires at least \( O(n) \) samples.

Theorem 13. Let \( h_S \) be the histogram for sample \( S \) and \( m \geq O\left(\frac{n + \log \frac{1}{\epsilon}}{\epsilon^2} \right) \). Then with high probability, \( d_{TV}(h_S, p) \leq \epsilon \).

Proof. To upper bound the total variation distance between \( p \) and \( h_S \), one only needs to upper bound \( |p(A) - h_S(A)| \) simultaneously for all \( A \subseteq [n] \).

Fix an arbitrary \( A \subseteq [n] \), one can use Hoeffding bound to prove \( Pr[|p(A) - h_S(A)| > \epsilon] \leq \frac{\delta}{m} \) when \( m \geq O\left(\frac{n + \log \frac{1}{\epsilon}}{\epsilon^2} \right) \). The proof follows from applying union bound over all \( 2^n \) possible subsets. \( \Box \)