1 Introduction

In the previous lecture, we proved a general sample complexity result under two assumptions:

1. The hypothesis set $\mathcal{H}$ has finite cardinality (i.e., $|\mathcal{H}| < \infty$).
2. The concept to be learned is in $\mathcal{H}$ (i.e., output of the algorithm is consistent).

Specifically, under the above assumptions, we proved that any consistent algorithm is a PAC-learning algorithm. Also, with confidence and accuracy parameters denoted as $\delta$ and $\varepsilon$ respectively, the following bound on the sample size $m$ holds:

$$m \geq \frac{1}{\varepsilon} \left( \log |\mathcal{H}| + \log \frac{1}{\delta} \right).$$

We begin by applying the above result to analyze PAC learning of a specific concept class.

**Example 1.1 (Conjunction of Boolean Literals).** Fix any positive integer $n$. Let $x_i \in \{0, 1\}$, for $i \in [n]$. A Boolean literal is either a variable $x_i$ or its negation $\bar{x}_i$. Consider the concept class $\mathcal{C}_n$ of conjunction of at most $n$ Boolean literals. That is,

$$\mathcal{C}_n = \left\{ (\bigwedge_{i \in F} x_i) \land (\bigwedge_{i \in G} \bar{x}_i) : F \cap G = \emptyset, \ F \cup G \subseteq [n] \right\}.$$

For $n = 6$, a specific training set (with positive examples and negative examples labelled) is given below.
<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>?</td>
<td>?</td>
<td>1</td>
<td>1</td>
<td>A consistent hypothesis: $\bar{x}_1 \land x_2 \land x_5 \land x_6$</td>
</tr>
</tbody>
</table>

Table 1: A training set for $n = 6$. For each example with positive label, $x_i$ is opted out as a candidate literal if the $i^{th}$ location has 0. Similarly, if it is 1, $\bar{x}_i$ is opted out. This strategy yields a consistent hypothesis.

In the light of the above example, we can consider the following algorithm for obtaining a consistent hypothesis. Let $\hat{F}(x_1, x_2, \ldots, x_n) \equiv \hat{F}$ denote the output of the algorithm corresponding to set $F$ in the definition of $C_n$. Similarly, define $\hat{G}$. Also, let $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, for $i \in [n]$. Then, a consistent algorithm is obtained as follows:

$$\hat{F}^c := \bigcap_{i=1}^{m} \{ j \in [n] : x_{ij} = 0 \},$$

$$\hat{G}^c := \bigcap_{i=1}^{m} \{ j \in [n] : x_{ij} = 1 \}.$$

Note that, in any conjunction, a literal could either be inserted positively ($x_i$), with negation ($\bar{x}_i$) or not inserted at all. Thus, the size of the concept class (and hence, the hypothesis set) is $|C_n| = 3^n$. Since we have a consistent algorithm with finite hypothesis set, we can make use of the above mentioned result on sample complexity. For the example discussed in Table 1 if $\delta = \varepsilon = 0.01$, the sample complexity bound is $m \geq 1120$.

### 2 Generalization Bound-Single Hypothesis

First, we will prove certain results which will turn out to be handy in deriving the required bounds. Let $\mathbb{R}, \mathbb{R}_+$ denote the real line and positive real line respectively.

**Lemma 2.1 (Markov’s Inequality).** Let $\Phi : \mathbb{R} \mapsto \mathbb{R}_+$ be a non negative function and let $X$ be a real valued random variable. Then, for all $\lambda > 0$,

$$\mathbb{P}[\Phi(X) \geq \lambda] \leq \frac{\mathbb{E}[\Phi(X)\]}{\lambda}.$$
Proof. Consider the following indicator function

\[ f(x) = \begin{cases} 
1, & \text{if } \Phi(x) \geq \lambda, \\
0, & \text{if } \Phi(x) < \lambda.
\end{cases} \tag{1} \]

Observe that, for all \( x \in \mathbb{R} \), \( f(x) \leq \Phi(x)/\lambda \). The result follows by taking expectation on both sides of the inequality and noting that \( E[f(X)] = \mathbb{P}[\Phi(X) \geq \lambda] \).

Remark 2.2. From Lemma 2.1, we recover the following important special cases:

1. With \( \Phi(x) = |x| \), we obtain the standard form of Markov’s inequality:

\[ \mathbb{P}[|X| \geq \lambda] \leq \frac{E[|X|]}{\lambda}. \]

2. With \( \Phi(x) = x^2 \), we obtain Chebyshev’s inequality:

\[ \mathbb{P}[|X| \geq \lambda] \leq \frac{E[X^2]}{\lambda^2}. \]

3. With \( \Phi(x) = e^{tx}, t > 0 \), we obtain

\[ \mathbb{P}[X \geq \lambda] \leq \frac{E[e^{TX}]}{e^{t\lambda}}. \]

Theorem 2.3 (Hoeffding’s Lemma). Let \( X \) be a random variable with \( E[X] = 0 \), \( X \in [a, b] \) with \( b > a \). Then, for any \( t > 0 \), the following holds:

\[ E[e^{tX}] \leq e^{t^2(b-a)/8}. \]

Proof. For \( x \in [a, b] \), let \( p = \frac{x-a}{b-a} \). Thus, \( 1 - p = \frac{b-x}{b-a} \). Note that \( pa + (1-p)b = x \). From the convexity of \( e^{tx} \), \( x \in [a, b] \),

\[ e^{tx} = e^{pta + (1-p)tb} \leq pe^{ta} + (1-p)e^{tb} = \frac{x-a}{b-a}e^{ta} + \frac{b-x}{b-a}e^{tb}. \]

Using the fact that \( E[X] = 0 \), we obtain

\[ E[e^{tX}] \leq \frac{be^{ta} - ae^{tb}}{b-a} := e^{\phi(t)}. \]
If we denote \( q = \frac{a}{b-a} \) (since \( \mathbb{E}[X] = 0, a < 0 \)), \( \phi(t) = at + \log[qe^{(b-a)t} + (1 - q)] \). It is easy to see that
\[
\phi'(t) = a + (b-a)\left[ 1 - \frac{(1-q)}{qe^{(b-a)t} + (1-q)} \right],
\]
\[
\phi''(t) = (b-a)^2u(t)(1-u(t)), \text{ where, } u(t) = \frac{(1-q)}{qe^{(b-a)t} + (1-q)}.
\]
Since \( u(t) \in [0, 1], u(t)(1-u(t)) \leq \frac{1}{4} \). Thus, \( \phi''(t) \leq \frac{(b-a)^2}{4} \). Also, observe that \( \phi(0) = \phi'(0) = 0 \). From Taylor’s expansion up to the second order, for some \( u \in (0,t) \),
\[
\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(u) \leq \frac{t^2(b-a)^2}{8}.
\]
This concludes the proof.

**Corollary 2.4 (Hoeffding’s Inequality).** Let \( \{X_i, 1 \leq i \leq m\} \) be independent random variables such that \( X_i \in [a_i, b_i] \), for \( a_i < b_i \), \( i \in [m] \). Let \( S_m = \sum_{i=1}^{m} X_i, a = (a_1, \ldots, a_m), b = (b_1, \ldots, b_m) \) and \( ||a-b|| = \sqrt{\sum_{i=1}^{m} (a_i-b_i)^2} \). Then, for any \( \epsilon > 0 \),
\[
\mathbb{P}\left[ S_m - \mathbb{E}[S_m] \geq \epsilon \right] \leq e^{-\frac{2\epsilon^2}{||a-b||^2}}, \tag{2}
\]
\[
\mathbb{P}\left[ S_m - \mathbb{E}[S_m] \leq -\epsilon \right] \leq e^{-\frac{2\epsilon^2}{||a-b||^2}}. \tag{3}
\]
In particular,
\[
\mathbb{P}\left[ \left| S_m - \mathbb{E}[S_m] \right| \geq \epsilon \right] \leq 2e^{-\frac{2\epsilon^2}{||a-b||^2}}. \tag{4}
\]

**Proof.** First we prove the inequality \((2)\). For any \( t > 0 \),
\[
\mathbb{P}\left[ S_m - \mathbb{E}[S_m] \geq \epsilon \right] \overset{(a)}{=} e^{-t\epsilon} \mathbb{E}\left[ e^{t(S_m - \mathbb{E}[S_m])} \right] \overset{(b)}{=} e^{-t\epsilon} \prod_{i=1}^{m} \mathbb{E}\left[ e^{t(X_i - \mathbb{E}[X_i])} \right] \overset{(c)}{=} e^{-t\epsilon} \prod_{i=1}^{m} e^{\frac{(b_i-a_i)^2}{8}} \overset{(d)}{=} e^{-t\epsilon} \sum_{i=1}^{m} e^{\frac{(b_i-a_i)^2}{8}} \leq e^{-\frac{2\epsilon^2}{||a-b||^2}},
\]
where, \((a)\) follows from Lemma 2.1, \((b)\) follows from the fact that the random variables are independent, \((c)\) follows from Theorem 2.3 and \((d)\) follows by choosing \( t = \frac{4\epsilon}{||a-b||^2} \). Thus, \((2)\) follows. The inequality \((3)\) follows in the same way. Finally, \((4)\) follows by invoking the union bound.
Corollary 2.5. Fix $\varepsilon > 0$. Let $S := \{X_i, \; 1 \leq i \leq m\}$ be i.i.d. random variables. For any hypothesis $h : X \mapsto \{0, 1\}$, the following holds:

$$\Pr[\hat{R}(h) - R(h) \geq \varepsilon] \leq e^{-2m\varepsilon^2},$$  \hspace{1cm} (5)

$$\Pr[\hat{R}(h) - R(h) \leq -\varepsilon] \leq e^{-2m\varepsilon^2}. \hspace{1cm} (6)$$

In particular,

$$\Pr[|\hat{R}(h) - R(h)| \geq \varepsilon] \leq 2e^{-2m\varepsilon^2}. \hspace{1cm} (7)$$

Proof. The result follows immediately from 2.4.

Corollary 2.6 (Generalization Bound-Single Hypothesis). Fix a hypothesis $h : \mathcal{X} \mapsto \{0, 1\}$. Then, for any $\delta > 0$, the inequality holds with probability at least $1 - \delta$:

$$|R(h) - \hat{R}(h)| \leq \sqrt{\log \frac{2}{\delta}} \frac{1}{2m}.$$

Example 2.7 (Tossing a Biased Coin). Consider a coin with $\Pr[\text{Head}] = p$. Let the hypothesis $h$ be such that $h(x) = \text{Head}$ for all $x$. Then, the generalization bound $R(h) = \mathbb{E}[1_{h(X_i) \neq Y_i}] = \mathbb{E}[1_{Y_i \neq \text{Head}}] = 1 - p$. From Corollary 2.6

$$\left| p - \frac{1}{m} \sum_{i=1}^{m} 1_{\{h(X_i) \neq Y_i\}} \right| \leq \sqrt{\log \frac{2}{\delta}} \frac{1}{2m}.$$

Remark 2.8. In general, $h_S$ is a random variable. Hence, $R(h_S)$, $\hat{R}(h_S)$ are also random variables.

Theorem 2.9 (Learning bound- finite $|\mathcal{H}| < \infty$, inconsistent case). Let $\mathcal{H}$ be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds:

$$|R(h_S) - \hat{R}(h_S)| \leq \sqrt{\log |\mathcal{H}| + \log \frac{2}{\delta}} \frac{1}{2m}.$$

Proof. The result follows directly from the result for single hypothesis (Corollary 2.6) and applying union bound.

Remark 2.10. Several observations follow from the result in Theorem 2.9:

1. The term $\log |\mathcal{H}|$ appearing in the learning bound has the following interpretation: $\log |\mathcal{H}|$ corresponds to the number of bits required to represent $\mathcal{H}$.

2. Large sample size guarantees tighter generalization bound.
3. The bound increases only logarithmically in $|\mathcal{H}|$.

4. The bound is worse than the bound in consistent case. In fact, quadratically larger number of samples are required for the same guarantee.

5. There is a trade off between empirical error and the size of the hypothesis set.

6. For similar empirical error, use a smaller hypothesis set (Occam’s razor principle).

3 Generalities

3.1 Deterministic versus Stochastic

A generalization of the supervised learning scenario that we discussed till forth is as follows. The training data is a labelled sample $S$ drawn i.i.d. according to distribution $D$ on $\mathcal{X} \times \mathcal{Y}$. The learning problem is to find a hypothesis $h \in \mathcal{H}$ such that the generalization error

$$R(h) = \mathbb{E}[1_{H(X) \neq Y}]$$

is small. This is the stochastic scenario. If the label can be uniquely determined by some measurable function $f : \mathcal{X} \mapsto \mathcal{Y}$ (with probability one), then the scenario is called deterministic. In this case, it suffices to consider distribution $D$ over $\mathcal{X}$ alone.

Definition 3.1 (Agnostic PAC-learning). Let $\mathcal{H}$ be a hypothesis set. $\mathcal{A}$ is an agnostic PAC-learning algorithm if there exists a polynomial function $\text{poly}(\ldots, \ldots)$ such that for any $\delta > 0$, $\varepsilon > 0$ and for all distributions $D$ on $\mathcal{X} \times \mathcal{Y}$, the following holds for any $m \geq \text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c))$:

$$\mathbb{P}[R(h_S) - \min_{h \in \mathcal{H}} R(h) \leq \varepsilon] \geq 1 - \delta.$$ 

Further, if $\mathcal{A}$ runs in $\text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c))$, it is said to be an efficient agnostic PAC learning algorithm.

3.2 Bayes Error and Noise

In the deterministic case, by definition, there exists a hypothesis $h$ such that the generalization error $R(h) = 0$. But for the stochastic case, there is a non-zero error for any hypothesis $h$. This prompts to define Bayes error.

Definition 3.2. Given a distribution $D$ over $\mathcal{X} \times \mathcal{Y}$, the Bayes error $R^*$ is defined as

$$R^* = \inf_{h \text{ measurable}} R(h).$$

A hypothesis $h$ with $R(h) = R^*$ is called a Bayes hypothesis or Bayes classifier.
By definition, $R^* = 0$ in the deterministic case. For the stochastic case, $R^* \neq 0$.

**Theorem 3.3.** The Bayes hypothesis $h_B(x) = \arg \max_y \mathbb{P}[y|x]$, $x \in \mathcal{X}$.

**Proof.** Note that,

$$1 - R^* = \sup_h [1 - R(h)]$$

$$= \sup_h \mathbb{E}[1_{\{h(X) \neq Y\}}]$$

$$= \sup_h \int \sum_h 1_{\{y = h(x)\}} \mathbb{P}[y|x] dF(x).$$

The result follows. \qed