1 Introduction

The PAC bounds we have derived in the previous lecture were restricted to a very simple setting, requiring very strong assumptions. Those do not hold in most real-world learning problems. In this lecture we extend the results to a much more general setting.

2 Agnostic Learning

We will proceed without making any assumptions on the distribution $P_{XY}$. This situation is often termed as Agnostic Learning. The root of the word agnostic literally means not known. The term agnostic learning is used to emphasize the fact that often, perhaps usually, we may have no prior knowledge about $P_{XY}$ and $f^*$. The question then arises about how we can reasonably select an $f \in \mathcal{F}$ in this setting.

2.1 Empirical Risk Minimization - How good is it?

Consider the Empirical Risk Minimization (ERM) selection of a classification rule from a model class $\mathcal{F}$. That is

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f).$$

If we guarantee that with probability at least $1 - \delta$ we have

$$|\hat{R}_n(f) - R(f)| \leq \epsilon, \quad \forall f \in \mathcal{F}, \quad (1)$$

for small $\epsilon > 0$ and $\delta > 0$ then the ERM is quite a reasonable choice. In fact with probability at least $1 - \delta$

$$R(\hat{f}_n) \leq \hat{R}(\hat{f}_n) + \epsilon \leq \hat{R}_n(f) + \epsilon, \quad \text{for any } f \in \mathcal{F} \leq R(f) + 2\epsilon, \quad \text{for any } f \in \mathcal{F}.$$

Therefore with probability at least $1 - \delta$

$$R(\hat{f}_n) \leq \inf_{f \in \mathcal{F}} R(f) + 2\epsilon,$$

and so with high probability the true risk of the selected rule is only a little bit higher than the risk of the best possible rule in the class. This indicates that ERM is quite a reasonable thing to do. Of course we still need to construct a bound like (1).
3 Constructing PAC bounds

To begin, let us recall the definition of empirical risk. Let \( \{X_i, Y_i\}_{i=1}^n \) be a collection of training data. Then the empirical risk is defined as

\[
\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i) .
\]

Note that since the training data \( \{X_i, Y_i\}_{i=1}^n \) are assumed to be i.i.d. pairs, the above is a sum of independent and identically distributed random variables.

Let

\[
L_i = \ell(f(X_i), Y_i) .
\]

The collection of losses \( \{L_i\}_{i=1}^n \) is i.i.d according to some unknown distribution (depending on the unknown joint distribution of \((X,Y)\) and the loss function). The expectation of \(L_i\) is \(E[\ell(f(X_i), Y_i)] = E[\ell(f(X), Y)] = R(f)\), the true risk of \(f\). For now, let’s assume that \(f\) is fixed. Then

\[
E[\hat{R}_n(f)] = \frac{1}{n} \sum_{i=1}^n E[\ell(f(X_i), Y_i)] = \frac{1}{n} \sum_{i=1}^n E[L_i] = R(f) .
\]

We know from the strong law of large numbers that the average (or empirical mean) \( \hat{R}_n(f) \) converges almost surely to the true mean \( R(f) \). That is, \( \hat{R}_n(f) \to R(f) \) almost surely as \( n \to \infty \). The question is how fast.

4 Concentration of Measure

Concentration inequalities are upper bounds on how fast empirical means converge to their ensemble counterparts, in probability. Area of the shaded tail regions in Figure 1 is \( P(|\hat{R}_n(f) - R(f)| > \epsilon) \). We are interested in finding out how fast this probability tends to zero as \( n \to \infty \). In other words, how quickly do the tails shrink when \( n \to \infty \)?

![Figure 1: Distribution of \( \hat{R}_n(f) \)](image.png)
4.1 Markov’s and Chebyshev’s Inequalities

Let’s recall Markov’s inequality. Let $Z$ be a non-negative random variable and $t > 0$. Then

\[
E[Z] = E[Z1\{Z \geq t\}] \\
\geq E[t1\{Z \geq t\}] \\
= P(Z \geq t),
\]

or

\[
P(Z \geq t) \leq \frac{E[Z]}{t}.
\]

This is known as Markov’s inequality. Now we can use this to get a bound on the tails of an arbitrary random variable $X$. Let $t > 0$

\[
P(|X - E[X]| \geq t) = P((X - E[X])^2 \geq t^2) \\
\leq \frac{E[(X - E[X])^2]}{t^2} \\
= \frac{\text{Var}(X)}{t^2},
\]

where $\text{Var}(X)$ denotes the variance of $X$. This inequality is known as Chebyshev’s inequality.

We can use Chebyshev’s inequality to get a simple PAC-style bound. Take $X = \hat{R}_n(f)$ and $t > 0$. Then

\[
P(|\hat{R}_n(f) - R(f)| \geq \epsilon) \leq \frac{\text{Var}(\hat{R}_n(f))}{\epsilon^2} = \frac{\sigma_L^2}{n\epsilon^2},
\]

where $\sigma_L^2 = \text{Var}(L_i)$ (note that the variance of the average of independent random variables is the average of the individual variances).

The tail probability goes to zero at a rate of at least $n^{-1}$, which is the expected behavior, but in light of the Central Limit Theorem (CLT) this seems like an extremely loose bound. According to the CLT

\[
\sqrt{n} \left( \hat{R}_n(f) - R(f) \right) \overset{D}{\to} \mathcal{N}(0, \sigma_L^2),
\]

as $n \to \infty$. This suggests (see footnote\(^1\)) that

\[
P(\sqrt{n} |\hat{R}_n(f) - R(f)| \geq t) \approx O \left( e^{-\frac{t^2}{2\sigma_L^2}} \right).
\]

To get a better looking bound we can take $t = \sqrt{n}\epsilon$, and so

\[
P(|\hat{R}_n(f) - R(f)| \geq \epsilon) \approx O \left( e^{-\frac{n\epsilon^2}{2\sigma_L^2}} \right).
\]

Note that, when $n$ grows the empirical risk converges to the true risk exponentially fast. Clearly we are a bit far of with Chebyshev’s inequality, where the convergence is only polynomial. We need a better concentration inequality.

\(^1\)For a Gaussian random variable $Z \sim \mathcal{N}(0,1)$ and $\gamma > 0$ we have

\[
\frac{1}{\sqrt{2\pi}\gamma^2} \left( 1 - \frac{1}{\gamma^2} \right) e^{-\frac{x^2}{2}} \leq \Pr(Z > \gamma) \leq \frac{1}{\sqrt{2\pi}\gamma^2} e^{-\frac{\gamma^2}{2}}.
\]
5 Hoeffding’s Inequality

Theorem 1 Hoeffding’s Inequality Let $Z_1, Z_2, \ldots, Z_n$ be independent bounded random variables such that $Z_i \in [a_i, b_i]$ with probability 1. Let $S_n = \sum_{i=1}^n Z_i$. Then for any $t > 0$, we have

1. $P(S_n - E[S_n] \geq t) \leq e^{-\frac{t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$

2. $P(S_n - E[S_n] \leq -t) \leq e^{-\frac{t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$

3. $P(|S_n - E[S_n]| \geq t) \leq 2e^{-\frac{t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$

**Proof:** Let $Z$ be any random variable and $s > 0$. Note that

$$P(Z \geq t) = P(e^{sZ} \geq e^{st}) \leq e^{-st} E[e^{sZ}] ,$$

by using Markov’s inequality, and noting that $e^{sx}$ is a non-negative monotone increasing function. For clever choices of $s$ this can be quite a good bound.

Let’s look now at $\sum_{i=1}^n Z_i - E[Z_i]$. Then

$$P(\sum_{i=1}^n Z_i - E[Z_i] \geq t) \leq e^{-st} E\left[e^{s(\sum_{i=1}^n Z_i - E[Z_i])}\right]$$

$$= e^{-st} E\left[\prod_{i=1}^n e^{s(Z_i - E[Z_i])}\right]$$

$$= e^{-st} \prod_{i=1}^n E\left[e^{s(Z_i - E[Z_i])}\right] ,$$

where the last step follows from the independence of the $Z_i$’s. The above procedure is called the Chernoff bounding technique (Chernoff, 1952). To complete the proof we need to find a good bound for $E\left[e^{s(Z_i - E[Z_i])}\right]$.

![Convexity of exponential function](image_url)

Figure 2: Convexity of exponential function.
Lemma 1 Let $X$ be a r.v. such that $E[X] = 0$ and $a \leq X \leq b$ with probability one. Then

$$E[e^{sX}] \leq e^{\frac{2(s(b-a))^2}{8}}.$$ 

Proof: This upper bound is derived as follows. By the convexity of the exponential function,

$$e^{sz} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}, \text{ for } a \leq z \leq b.$$ 

Thus,

$$E[e^{sX}] \leq E \left[ \frac{X-a}{b-a} e^{sb} + \frac{b-X}{b-a} e^{sa} \right]$$

$$= \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}, \text{ since } E[X] = 0$$

$$= (1 - \lambda + \lambda e^{s(b-a)}) e^{-\lambda s(b-a)}, \text{ where } \lambda = \frac{-a}{b-a}.$$ 

Now let $u = s(b - a)$ and define

$$\phi(u) \equiv -\lambda u + \log(1 - \lambda + \lambda e^u),$$

so that

$$E[e^{sX}] \leq (1 - \lambda + \lambda e^{s(b-a)}) e^{-\lambda s(b-a)} = e^{\phi(u)}.$$ 

We want to find a good upper-bound on $e^{\phi(u)}$. Let’s express $\phi(u)$ as its Taylor series with remainder:

$$\phi(u) = \phi(0) + u \phi'(0) + \frac{u^2}{2} \phi''(v) \text{ for some } v \in [0, u].$$

$$\phi'(u) = -\lambda + \frac{\lambda e^u}{1 - \lambda + \lambda e^u} \Rightarrow \phi'(u) = 0$$

$$\phi''(u) = \lambda e^u \left( \frac{1}{1 - \lambda + \lambda e^u} - \frac{\lambda e^u}{(1 - \lambda + \lambda e^u)^2} \right)$$

$$= \lambda e^u \left( \frac{1}{1 - \lambda + \lambda e^u} - \frac{\lambda e^u}{1} \right)$$

$$= \rho(1 - \rho),$$

where $\rho = \frac{\lambda e^u}{1 - \lambda + \lambda e^u}$. Now note that $\rho(1 - \rho) \leq 1/4$, for any value of $\rho$ (the maximum is attained when $\rho = 1/2$, therefore $\phi''(u) \leq 1/4$. So finally we have

$$\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}, \text{ and therefore}$$

$$E[e^{sX}] \leq e^{\frac{2(s(b-a))^2}{8}}.$$ 

Now, we can apply this upper bound to derive Hoeffding’s inequality.
\[ P(S_n - E[S_n] \geq t) \leq e^{-st} \prod_{i=1}^{n} E[e^{s(Z_i - E[Z_i])}] \]
\[ \leq e^{-st} \prod_{i=1}^{n} e^{\frac{s^2(b_i-a_i)^2}{2}} \]
\[ = e^{-st} e^{s^2 \sum_{i=1}^{n} \frac{(b_i-a_i)^2}{2}} \]
\[ = e^{-st} e^{-2t^2 \sum_{i=1}^{n} (b_i-a_i)^2} \]

by choosing \( s = \frac{4t}{\sum_{i=1}^{n} (b_i-a_i)^2} \)

This concludes the proof of (1). To show (2) one just needs to apply (1) to the r.v.'s \((-Z_1), \ldots, (-Z_n)\). Finally (3) follows by using (1) and (2) simultaneously and the union of events bound.

As a final remark notice that Hoeffding’s inequality (Hoeffding 1963) is a generalization of the Chernoff bound, which applies only to Bernoulli random variables.