Instructor: Adarsh Barik
Scribed by: Soumyodeep Dey
Lecture 2, COV878
August 20, 2025

LECTURE 2: PAC LEARNING, HOEFFDING'S INEQUALITY AND APPLICATIONS

In supervised learning, our goal is to learn a mapping from an input space $\mathcal{X}$ to a label space $\mathcal{Y}$. We are given a training set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where each pair $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ is assumed to be drawn independently and identically distributed (i.i.d.) from an unknown joint probability distribution P_{xy}.

Our objective is to find a function, called a **hypothesis** $h : \mathcal{X} \to \mathcal{Y}$, that performs well on new, unseen data from the same distribution. We typically search for this hypothesis within a predefined set of possible functions, known as the hypothesis class $\mathcal{H}$.

To measure the performance of a hypothesis, we use a *loss function* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. Following is a common example for classification tasks:

$$L(h(x), y) = \mathbb{I}[h(x) \neq y] = \begin{cases} 1, & \text{if } h(x) \neq y \\ 0, & \text{if } h(x) = y \end{cases}$$

This loss function simply indicates whether a prediction is correct or not.

True Risk and Empirical Risk

There are two fundamental ways to quantify the error of a hypothesis.

Definition 1 (True Risk)**.** *The true risk (or generalization error) of a hypothesis $h \in \mathcal{H}$ is its expected loss over the true data distribution P_{xy}. It is defined as:*

$$R(h) = E_{(x,y)\sim P_{xy}}[L(h(x), y)]$$

For binary classification with the 0-1 loss, this simplifies to the probability of misclassification:

$$R(h) = \mathbb{P}_{(x,y)\sim P_{xy}}[h(x) \neq y]$$

The true risk is what we ultimately care about, but we cannot compute it directly because the distribution P_{xy} is unknown. Instead, we use the training data to calculate an estimate of the true risk.

Definition 2 (Empirical Risk)**.** *The empirical risk (or training error) of a hypothesis $h \in \mathcal{H}$ is its average loss on the training set S. It is defined as:*

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i), y_i)$$

For binary classification with the 0-1 loss, this is the fraction of misclassified training examples:

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[h(x_i) \neq y_i]$$

1 Probably Approximately Correct (PAC) Learning

The PAC framework provides a formal definition of what it means for a hypothesis class to be "learnable".

Definition 3 (Agnostic PAC-Learnable)**.** *A hypothesis class $\mathcal{H}$ is agnostic PAC-learnable if $\exists$ a function $N : (0,1)^2 \to \mathbb{N}$ and a learning algorithm such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution P_{xy} over $\mathcal{X} \times \mathcal{Y}$: if the algorithm is given $n \geq N(\epsilon, \delta)$ i.i.d. samples from P_{xy}, it returns a hypothesis $\hat{h} \in \mathcal{H}$ such that, with probability at least $1 - \delta$,*

$$R(\hat{h}) \leq \inf_{h^* \in \mathcal{H}} R(h^*) + \epsilon$$

The term "agnostic" means we make no assumptions about the data, not even that the best hypothesis in our class can achieve zero error. A simpler, special case is the "realizable" setting.

Definition 4 (Realizable PAC-Learnable)**.** *A hypothesis class $\mathcal{H}$ is realizable PAC-learnable if it is PAC-learnable under the assumption of realizability. Realizability means there exists at least one hypothesis $h^* \in \mathcal{H}$ with zero true risk, i.e., $R(h^*) = 0$. In this case, the PAC guarantee simplifies to:*

$$R(\hat{h}) \leq \epsilon$$

2 Empirical Risk Minimization (ERM)

A natural and powerful learning strategy is to find the hypothesis that best fits the training data.

Definition 5 (Empirical Risk Minimization (ERM))**.** *The Empirical Risk Minimization (ERM) algorithm is a learning rule that, given a training set S, outputs a hypothesis $\hat{h}$ that minimizes the empirical risk:*

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} [\hat{R}(h)]$$

We will first analyze ERM under the assumptions that we have a finite hypothesis class, that is, $|H| < \infty$, and our problem is realizable.

Theorem 1. *Let $\mathcal{H}$ be a finite hypothesis class. Assume the learning problem is realizable, i.e., there exists an $h^* \in \mathcal{H}$ such that $R(h^*) = 0$. Let $\hat{h}$ be the hypothesis returned by the ERM algorithm on a set of n i.i.d. samples. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left[R(\hat{h}) \leq \epsilon\right] \geq 1 - |\mathcal{H}| \exp(-n\epsilon)$$

Proof. Let $\mathcal{H}_B \subseteq \mathcal{H}$ be the set of "bad" hypotheses, which are those with a true risk greater than ϵ:

$$\mathcal{H}_B = \{h \in \mathcal{H} \mid R(h) > \epsilon\}$$

Our goal is to bound the probability of the "bad event" where the ERM algorithm returns a hypothesis $\hat{h}$ from this set. Let this event be $E_{bad} = \{\hat{h} \in \mathcal{H}_B\}$.

From the realizability assumption, we know there exists an $h^* \in \mathcal{H}$ with $R(h^*) = 0$. This means h^* makes no errors on the true distribution, so its empirical risk on any training set must also be zero, i.e., $\hat{R}(h^*) = 0$.

The ERM algorithm finds a hypothesis $\hat{h}$ that minimizes the empirical risk. Since h^* is a candidate hypothesis in $\mathcal{H}$ with $\hat{R}(h^*) = 0$, the hypothesis $\hat{h}$ returned by ERM must also satisfy $\hat{R}(\hat{h}) = 0$.

The bad event E_{bad} can only happen if ERM selects a hypothesis from $\mathcal{H}_B$. However, for any hypothesis to be selected by ERM in this setting, it must have an empirical risk of 0. Therefore, the bad event E_{bad} implies that at least one hypothesis in $\mathcal{H}_B$ must have had an empirical risk of 0 on the training data. This leads to the following crucial step, which is an application of the union bound:

$$\begin{aligned}\mathbb{P}\left[R(\hat{h}) > \epsilon\right] &= \mathbb{P}\left[\hat{h} \in \mathcal{H}_B\right] \\ &\leq \mathbb{P}\left[\exists h \in \mathcal{H}_B : \hat{R}(h) = 0\right] \\ &\leq \sum_{h \in \mathcal{H}_B} \mathbb{P}\left[\hat{R}(h) = 0\right] \end{aligned} \tag{1}$$

Now we just need to bound the probability $\mathbb{P}\left[\hat{R}(h) = 0\right]$ for any single bad hypothesis $h \in \mathcal{H}_B$. By definition of $\mathcal{H}_B$, we have $R(h) = \mathbb{P}\left[h(x) \neq y\right] > \epsilon$. This means the probability of h correctly classifying a single, randomly drawn example is $\mathbb{P}\left[h(x) = y\right] = 1 - R(h) < 1 - \epsilon$.

The event $\hat{R}(h) = 0$ means that h correctly classifies all n i.i.d. samples in the training set. The probability of this is:

$$\begin{aligned}\mathbb{P}\left[\hat{R}(h) = 0\right] &= \mathbb{P}\left[h(x_1) = y_1, \ldots, h(x_n) = y_n\right] \\ &= \prod_{i=1}^{n} \mathbb{P}\left[h(x_i) = y_i\right] \quad \text{(due to i.i.d. samples)} \\ &\leq (1-\epsilon)^n \end{aligned}$$

Substituting this result back into our sum from (1):

$$\begin{aligned}\mathbb{P}\left[R(\hat{h}) > \epsilon\right] &\leq \sum_{h \in \mathcal{H}_B} (1-\epsilon)^n \\ &= |\mathcal{H}_B|(1-\epsilon)^n \\ &\leq |\mathcal{H}|(1-\epsilon)^n \quad \text{(since } \mathcal{H}_B \subseteq \mathcal{H}\text{)} \end{aligned}$$

Finally, we use the well-known inequality $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$, which gives $(1-\epsilon)^n \leq \exp(-n\epsilon)$. This gives us the final bound:

$$\mathbb{P}\left[R(\hat{h}) > \epsilon\right] \leq |\mathcal{H}| \exp(-n\epsilon)$$

This completes the proof. □

This theorem shows that the probability of ERM selecting a "bad" hypothesis decreases exponentially with the number of training samples n.

2.1 Sample Complexity

From the theorem, we can derive the sample complexity: the number of samples n that is *sufficient* to guarantee a certain level of performance. We want the failure probability to be at most δ:

$$\begin{aligned} |\mathcal{H}| \exp(-n\epsilon) &\leq \delta \\ \exp(-n\epsilon) &\leq \frac{\delta}{|\mathcal{H}|} \\ -n\epsilon &\leq \ln\left(\frac{\delta}{|\mathcal{H}|}\right) \\ n\epsilon &\geq -\ln\left(\frac{\delta}{|\mathcal{H}|}\right) = \ln\left(\frac{|\mathcal{H}|}{\delta}\right) \\ n &\geq \frac{1}{\epsilon}\ln\left(\frac{|\mathcal{H}|}{\delta}\right) \end{aligned}$$

This gives us a concrete number of samples $N(\epsilon, \delta)$ which is sufficient to ensure that with probability $1-\delta$, our ERM-learned hypothesis has true risk at most ϵ. This confirms that for finite hypothesis classes under realizability, ERM is a valid PAC learning algorithm. The analysis for the agnostic case is more involved and requires stronger concentration inequalities like Hoeffding's inequality.

2.2 Hoeffding's Lemma and Inequality

Lemma 1 (Hoeffding's Lemma)**.** *Let X be a random variable with support on $[0,1]$ and mean $\mathbb{E}[X] = \mu$. Then for any $t \in \mathbb{R}$, we have:*

$$\mathbb{E}\left[\exp\left(t(X-\mu)\right)\right] \leq \exp\left(\frac{t^2}{8}\right)$$

Proof Sketch. The proof relies on the convexity of the exponential function.

1. Define $f(x) = \exp(t(x-\mu))$. This function is convex.
2. For any $x \in [0,1]$, we can write $x = (1-x)\cdot 0 + x \cdot 1$. By Jensen's inequality (or the definition of convexity), $f(x) \leq (1-x)f(0) + xf(1)$.
3. Taking the expectation over X, we get $\mathbb{E}[f(X)] \leq (1-\mu)f(0) + \mu f(1)$.
4. Specifically, $\mathbb{E}[\exp(t(X-\mu))] \leq \exp(g(t))$, where $g(t) = -t\mu + \log(1-\mu+\mu e^t)$
5. We can easily verify that $g(0) = 0$, $\frac{dg}{dt}(0) = 0$ and $\frac{d^2g}{dt^2} \leq \frac{1}{4}$.
6. After some algebra and using a Taylor expansion of $g(t)$ around 0 we can show that $g(t) \leq \frac{t^2}{8}$, which proves the lemma.

□

This lemma can be extended to a random variable bounded in any interval $[a,b]$.

Corollary 1. *Let X be a random variable with support on $[a,b]$ and mean $\mathbb{E}[X] = \mu$. Then for any $t \in \mathbb{R}$:*

$$\mathbb{E}\left[\exp\left(t(X-\mu)\right)\right] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

Using Hoeffding's Lemma and the Chernoff bounding technique (which involves applying Markov's inequality to the exponential), we arrive at Hoeffding's inequality.

Theorem 2 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with support on $[0,1]$ and common mean $\mathbb{E}[X_i] = \mu$. For any $\epsilon > 0$, we have:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right] \leq 2\exp(-2n\epsilon^2)$$

A more general version allows for different bounds and means for each random variable.

Theorem 3 (Hoeffding's Inequality, General Version)**.** *Let $X_1, \ldots, X_n$ be independent random variables with X_i having support on $[a_i, b_i]$. For any $\epsilon > 0$, we have:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \epsilon\right] \leq 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

3 ERM in the Agnostic Setting

We can now use Hoeffding's inequality to provide a PAC guarantee for ERM in the agnostic setting.

Theorem 4. *Let $\mathcal{H}$ be a finite hypothesis class. Let $\hat{h}$ be the hypothesis returned by the ERM algorithm on a set of n i.i.d. samples. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left[R(\hat{h}) \leq \inf_{h\in\mathcal{H}} R(h) + 2\epsilon\right] \geq 1 - 2|\mathcal{H}|\exp(-2n\epsilon^2)$$

Proof. The proof consists of two main steps. First, we show that with high probability, the empirical risk is close to the true risk for all hypotheses simultaneously. Second, we use this fact to bound the excess risk of the ERM hypothesis.

Step 1: Uniform Convergence

Fix an arbitrary hypothesis $h \in \mathcal{H}$. Let's define a set of random variables $Z_1, \ldots, Z_n$ where $Z_i = L(h(x_i), y_i) = \mathbb{I}[h(x_i) \neq y_i]$. Since the loss is 0 or 1, each Z_i is a random variable with support on $[0,1]$. The true risk is the mean of this random variable: $R(h) = \mathbb{E}[Z_i]$. The empirical risk is the sample mean: $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Since the Z_i's are i.i.d. (because the data samples are i.i.d.), we can apply Hoeffding's Inequality (Theorem 2) to this specific, fixed h:

$$\mathbb{P}\left[\left|\hat{R}(h) - R(h)\right| \geq \epsilon\right] \leq 2\exp(-2n\epsilon^2)$$

This bound holds for one hypothesis. We need it to hold for *all* hypotheses in $\mathcal{H}$ at the same time. We use the union bound to achieve this:

$$\begin{aligned}\mathbb{P}\left[\exists h \in \mathcal{H} : \left|\hat{R}(h) - R(h)\right| \geq \epsilon\right] &\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left[\left|\hat{R}(h) - R(h)\right| \geq \epsilon\right] \\ &\leq |\mathcal{H}| \cdot 2\exp(-2n\epsilon^2)\end{aligned}$$

This is the probability of a "bad event" where at least one hypothesis has its empirical risk far from its true risk. The complementary "good event" is that for all hypotheses, the risks are close.

$$\mathbb{P}\left[\forall h \in \mathcal{H} : \left|\hat{R}(h) - R(h)\right| \leq \epsilon\right] \geq 1 - 2|\mathcal{H}|\exp(-2n\epsilon^2)$$

This property is known as uniform convergence. For the rest of the proof, we assume we are in this high-probability "good event".

Step 2: Bounding the Excess Risk

Let $h^* = \arg\min_{h\in\mathcal{H}} R(h)$ be the best possible hypothesis in our class. We want to bound the excess risk, $R(\hat{h}) - R(h^*)$. We can decompose this term as follows:

$$R(\hat{h}) - R(h^*) = \left(R(\hat{h}) - \hat{R}(\hat{h})\right) + \left(\hat{R}(\hat{h}) - \hat{R}(h^*)\right) + \left(\hat{R}(h^*) - R(h^*)\right)$$

Now let's bound each of the three terms, assuming the uniform convergence property holds:

1. $\left(R(\hat{h}) - \hat{R}(\hat{h})\right)$: By uniform convergence, we know $\left|R(h) - \hat{R}(h)\right| \leq \epsilon$ for all h, including $\hat{h}$. Thus, $R(\hat{h}) - \hat{R}(\hat{h}) \leq \epsilon$.
2. $\left(\hat{R}(\hat{h}) - \hat{R}(h^*)\right)$: By definition, $\hat{h}$ is the ERM solution, meaning it minimizes the empirical risk. Therefore, $\hat{R}(\hat{h}) \leq \hat{R}(h)$ for all $h \in \mathcal{H}$, including h^*. This implies $\hat{R}(\hat{h}) - \hat{R}(h^*) \leq 0$.
3. $\left(\hat{R}(h^*) - R(h^*)\right)$: Again, by uniform convergence, $\left|\hat{R}(h^*) - R(h^*)\right| \leq \epsilon$. This implies $\hat{R}(h^*) - R(h^*) \leq \epsilon$.

Combining these bounds, we get:

$$\begin{aligned}R(\hat{h}) - R(h^*) &\leq \epsilon + 0 + \epsilon \\ R(\hat{h}) - R(h^*) &\leq 2\epsilon\end{aligned}$$

This inequality holds whenever the uniform convergence event occurs. The probability of this event is at least $1 - 2|\mathcal{H}|\exp(-2n\epsilon^2)$. Therefore, we have shown that with high probability,

$$R(\hat{h}) \leq R(h^*) + 2\epsilon$$

which completes the proof. □