Fondamentaux de l'Apprentissage Automatique

Lecturer: Yann Chevaleyre
Scribe: Jim Wu

Lecture n°4 #
19/10/2023

1 Introduction

Theorem 1. *(Hoeffding Inequality) Let $Z_1, \cdots, Z_N \in [0,1]$ be i.i.d random variables. Then we have the following two bounds :*

$$P\left(\frac{1}{N}\sum_{i=1^N} Z_i - \mathbb{E}[Z] \geq \epsilon\right) \leq \exp\left(-2N\epsilon^2\right)$$

$$P\left(\left|\frac{1}{N}\sum_{i=1^N} Z_i - \mathbb{E}[Z]\right| \geq \epsilon\right) \leq 2\exp\left(-2N\epsilon^2\right)$$

This Jupyter Notebook used in class may be a useful reference as well.

Definition 1. *(Convergence in probability) A sequence of random variables $Z_1, \cdots, Z_N$ converges in probability to Z (denoted $Z_N \xrightarrow[N\to\infty]{prob} Z$ or $Z_N \xrightarrow{p} Z$) iff $\forall \epsilon, \delta \in]0,1[, \exists n$, if $N > n$, $|Z_N - Z| < \epsilon$ with probability (w.p.) $1-\delta$.*

Equivalently, there exists a function $n(\epsilon, \delta)$ such that $\forall \epsilon, \delta \in]0,1[, N > n(\epsilon,\delta) \implies |Z_n - Z| < \epsilon$ w.p. $1-\delta$.

Exercise 1:
Show in the Hoeffding setting that

$$\frac{1}{N}\sum_i Z_i \xrightarrow{p} \mathbb{E}[Z]$$

and give $n(\epsilon, \delta)$.

Solution 1:
We have another variable,

$$Y_i = \frac{1}{N}\sum_{j=1}^{N} Z_j$$

and now we can equivalently show that $Y_N \xrightarrow{p} \mathbb{E}[Z]$. Computing expectation of the variable, we have

$$\begin{aligned}\mathbb{E}[Y_n] &= \frac{1}{N}\sum \mathbb{E}[Z_i] \\ &= \mathbb{E}[Z]\end{aligned}$$

Then using the two-sided Hoeffding inequality, we have that

$$P(|\frac{1}{N}Y_n - \mathbb{E}[Z]| \geq \epsilon) \leq 2\exp(-2N\epsilon^2)$$

Fix ϵ, δ. Then $2\exp(-2n\epsilon^2)$ can be arbitrarily small as N tends towards infinity, so suppose we will have n such that $2\exp(-2n\epsilon) \leq \delta$. Isolating for n, we have the inequality

$$n \geq \frac{\log \frac{2}{\delta}}{2\epsilon^2}$$

Then if $N > n$, we have $P(|Y_n - \mathbb{E}[Z]| \leq \epsilon) \geq 1 - \delta$ and consequently $P(|Y_n - \mathbb{E}[Z]| > \epsilon) \leq \delta$.

Instead of solving the inequality for N, we solve it for ϵ. We have $\epsilon \leq \sqrt{\frac{\log \frac{2}{\delta}}{2N}}$. This finally gives us

$$|\frac{1}{N}Y_n - \mathbb{E}[Z]| < \sqrt{\frac{\log \frac{2}{\delta}}{2N}}$$

w.p. at least $1 - \delta$, and thus Y_i converges in probability to $\mathbb{E}[Z]$ in the Hoeffding setting.

We also studied Bayes Risk and Empirical Risk on two Gaussians through a Jupyter Notebook. The setting is two classes being represented by two normal distributions, $N(1, 1)$ and $N(-1, 1)$; these will be class blue and class red respectively. Suppose f_0 is the classifier that has 0 has a threshold for classification, that is any values greater than 0 will be classified as blue, and any values less than 0 will be classified red (values being 0 could be either, it is negligible).

Then our empeirical risk is $\hat{R}_S f_0 = \frac{1}{N}\sum \mathbb{1}[f_0(x) \neq y_i]$; we essentially have Bernoulli $Z_i = \mathbb{1}[f_0(x) \neq y_i]$, and now this resembles random variables we could use with Hoeffding inequality. Specifically, we have $\left|\hat{R}_S(f_0) - R(f_0)\right| < \sqrt{\frac{\log \frac{2}{\delta}}{2N}}$ w.p. $1 - \delta$.

Another natural question is can we bound this for ERM ? Specifically, can we bound $\left|\hat{R}_S(f_{ERM}) - R(f_{ERM})\right|$? We can't, since f_{ERM} is computed using the data (note that before, we assumed something about the data-generating process, but we didn't create a classifier based off the data we got). As well, in general this difference will be really big : $\hat{R}(f_{ERM}$ can theoretically be 0 through overfitting, and $R(f_{ERM})$ is unbounded, thus the difference is big.

In this lecture, we will consider $S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$ to be a random variable, and the empirical risk $\hat{R}_S(f_S) = \frac{1}{N}\sum \ell(f_s(x_i), y_i)$.

2 Notion of Consistency

Recall the definitions of a few functions :

$$f_S \in \arg\min_{f \in \mathcal{F}} \hat{R}_S(f)$$

$$f^* \in \arg\min_{f \in \text{measurable}} \hat{R}(f)$$

$$f_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \hat{R}(f)$$

We have a few definitions for the learning algorithm f_S :

Definition 2. *(Bayes-consistency). f_S is universally Bayes consistent iff for all distributions P, $R(f_S) \xrightarrow{p} R(f^*)$. In other words, there is a function $n(\epsilon, \delta, P)$ such that for any P, ϵ, δ, if $N > n)\epsilon, \delta, P)$ then for $S \sim P^N$, $|R(f_s) - R(f^*)| < \epsilon$ w.p. $1 - \delta$.*

Note that this is impossible for ERM.

Definition 3. *($\mathcal{F}$-consistency). f_S is universally $\mathcal{F} -$ consistent consistent if for all distributions P, $R(f_S) \xrightarrow{p} R(f_\mathcal{F})$.*

Definition 4. *(PAC-learner). f_S is a PAC-learner (Probably approximately correct) if there is a function $n(\epsilon, \delta)$ such that for all distributions P, $\forall \epsilon, \delta \in]0, 1[$, if $N > n(\epsilon, \delta)$ then for $S \sim P^N$, $|R(f_S) - R(f_\mathcal{F})| < \epsilon$ w.p. $1 - \delta$.*

Note that being a PAC-learner implies $\mathcal{F}$-consistency.

3 PAC learning and Uniform Convergence for ERM

We would like to bound the estimation error, $R(f_S) - R(f_\mathcal{F})$; this can be done with the Hoeffding inequality for a fixed f.

We can manipulate $R(f_S) - R(f_\mathcal{F})$ by adding terms that sum to 0 to get :

$$R(f_S) - R(f_\mathcal{F}) = R(f_S) - \hat{R}(f_S) + \hat{R}(f_S) - \hat{R}(f_\mathcal{F}) + \hat{R}(f_\mathcal{F}) - R(f_\mathcal{F})$$

Since f_S minimizes the empirical risk on function class F, it must be that the empirical risk on f_S is at most that of $f_\mathcal{F}$, or in other words $\hat{R}(f_S) - \hat{R}(f_\mathcal{F}) \leq 0$. These are the third and fourth terms of the right hand side, so we can say that

$$R(f_S) - R(f_\mathcal{F}) \leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \tag{1}$$

Definition 5. *The UnRepresentativeness of S is*

$$UnRep(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|$$

NB : The term more commonly used is Representativeness, but this is a bit unintuitive because it describes the maximal difference between the theoretical error and emperical error. See page 375 of Understanding Machine Learning by Shai Shalev-Shwartz and Shai Ben-David (official free PDF for personal use).

To think about what this measure might mean, consider the data-generating process where two 2D Gaussians have some non-negligible overlap ; say the classes are blue and red. We have two datasets each with two clusters of points in the same area. For the first dataset, we have a cluster of mainly blue points near where the blue Gaussian is, and another cluster of mainly red points near where the red Gaussian is. Then the best linear classifier can separate them with fairly low error, in which case $UnRep(\mathcal{F}, S)$ is low.

On the other hand, the second dataset reverses where the clusters are, but the original blue and red Gaussians stay the same. The best linear classifier will separate the data correctly, but it would be theoretically wrong, i.e. $\hat{R}(f)$ is near 0, but $R(f)$ is much closer to 1. This means that the UnRepresentativeness would be high.

Theorem 2. *If, for class $\mathcal{F}$, there exists $n(\epsilon, \delta)$ such that for any distribution P, any $\epsilon, \delta \in]0,1[$, if $N > n(\epsilon, \delta)$ then $UnRep(\mathcal{F}, S) < \epsilon$ w.p. $1-\delta$ (the uniform convergence property), then ERM is a PAC learner on $\mathcal{F}$.*

Démonstration. If $N > n(\frac{\epsilon}{2}, \delta)$ then $UnRep(\mathcal{F}, S) \leq \frac{\epsilon}{2}$ w.p. $1-\delta$ and $R(f_S) - R(f_{\mathcal{F}}) \leq 2UnRep(\mathcal{F}, S) \leq \epsilon$ w.p. $1-\delta$, so f_S is a PAC learner. The last inequality comes from 1 and the definition of $UnRep$. □

3.1 Application to finite class $\mathcal{F}$

We'd like to show that

$$\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| < \epsilon$$

w.p. $1-\delta$ for $N > n(\epsilon, \delta)$.
We have

$$\begin{aligned} P(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \geq \epsilon) &= P(\exists f \in \mathcal{F}, |R(f) - \hat{R}(f)| \geq \epsilon) \\ &\leq \sum_{f \in \mathcal{F}} P(|R(f) - \hat{R}(f)| \geq \epsilon) \end{aligned}$$

Where the last inequality comes from the Union bound, which states that $P(A \cup B) \leq P(A) + P(B)$ and thus $P(\exists i, A_i) \leq \sum_i P(A_i)$.

Note that this f does not depend on the data, and thus we can apply the Hoeffding inequality. As well, recall that $\hat{R}(f) = \frac{1}{N} \sum \ell(f(x_i), y_i)$ and $R(f) = \mathbb{E}_{S \sim P^N}[\hat{R}(f)]$

With Hoeffding inequality, if we have $N > \frac{\log \frac{2}{\delta}}{2\epsilon^2}$, we also have

$$\begin{aligned} P(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \geq \epsilon) &\leq \sum_{f \in \mathcal{F}} P(|R(f) - \hat{R}(f)| \geq \epsilon) \\ &\leq \delta |\mathcal{F}| \end{aligned}$$

which implies that $UnRep(\mathcal{F}, s) \leq \epsilon$ w.p. $1 - \delta|\mathcal{F}|$ if $N > \frac{\log \frac{2}{\delta}}{2\epsilon^2}$. Importantly, this tells us that ERM on finite classes is PAC-learnable.

Also, this is equivalent to saying $UnRep(\mathcal{F}, S) \leq \sqrt{\frac{\log \frac{2}{\delta'}}{2N}}$ w.p. at least $1-\delta'$ and $|R(f_S) - \hat{R}(f_S)| \leq \sqrt{\frac{\log \frac{2}{\delta'}}{2N}}$ w.p. at least $1-\delta'$

NB : This justification only works for finite classes because the union bound for infinite classes is infinite.

4 The case $|\mathcal{F}| = \infty$, Rademacher Complexity

The goal is to extend our previous result, for $|\mathcal{F}| = \infty$ without using union bound. There are many tools, including Vapnik dimension, Covering numbers, Gaussian Complexity, etc. Here, we visit Rademacher Complexity.

Rademacher applies to arbitrary bounded losses.

Notation : $Z = (x, y)$ is a labelled example. $S = (Z_1, \cdots, Z_N)$.
Given $\mathcal{F}$, define $G = \ell \circ \mathcal{F} = \{(x, y) \mapsto \ell(f(x), y), f \in \mathcal{F}\}$.
Then

$$\begin{aligned} UnRep(\mathcal{F}, S) &= \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \\ &= \sup_{g \in G} |\frac{1}{N} \sum g(Z_i) - \mathbb{E}_{Z \sim P}[g(z)]| \end{aligned}$$

Definition 6. *The Empirical Rademacher Complexity of S on g is*

$$\hat{Rad}(g) = \frac{1}{N} \mathbb{E}_{\sigma_1, \cdots, \sigma_N \sim U\{-1,1\}} [\sup_{g \in G} \sum \sigma_i g_i(z_i)]$$

Intuition 1 : Suppose we have drawn two datasets, S_1 and S_2.

$$\begin{aligned} \sup_{g \in G} |\hat{R}_{S_1}(f) - \hat{R}_{S_2}(f)| &= \sup_{g \in G} \frac{1}{N} [\sum_{Z_i \in S_1} g(Z_i) - \sum_{Z_i \in S_2} g(Z_2)] \\ &= \sup_{g \in G} \frac{1}{N} \sum_{Z_i \in S_1 \cup S_2} \sigma_i g(Z_i) \end{aligned}$$

where $\sigma_i = 1$ if $Z_i \in S_1$, and $\sigma_i = -1$ otherwise.
Assume S is given, and we average $\sup_{g \in G} |\hat{R}_{S_1}(f) - \hat{R}_{S_2}(f)|$ over all partitions of S in $S1, S2$, we get Rademacher complexity.
Intuition 2 : Measures how well $\mathcal{F}$ can fit noisy levels.

Lemma 1. *(Rademacher Lemma).* $\mathbb{E}_{S \sim P^N}[UnRep(\mathcal{F}, S)] \leq 2\mathbb{E}_{S \sim P^N}[\hat{Rad}(g)]$

Theorem 3. *(PAC with Rademacher). Assume $\forall (x, y), |\ell(f(x), y)| \leq c$. For all $f \in \mathcal{F}$, if $S \sim P^N$, then we have w.p. $1 - \delta$ that*

$$R(f) - \hat{R}_S(f) \leq 2\hat{Rad}_S(\ell \circ \mathcal{F}) + 4c\sqrt{\frac{2 \log \frac{4}{\delta}}{N}}$$

so we conclude the PAC result : $R(f) - R(f_{\mathcal{F}}) \leq ?(Exercise)$

Exercise 2:
1. Let $G = \{z \mapsto \alpha : \alpha \in [-1, 1]\}$. What is $\hat{Rad}_S(G)$?
2. Let G be the set of decision trees which can output 1 or -1 at the leaves. What is $\hat{Rad}_S(G)$?

Solution 2:
1. We have

$$\begin{aligned} \sup_{\alpha \in [-1,1]} \sum \sigma_i \alpha &= \sup_{\alpha \in \{-1,1\}} \sum \sigma_i \alpha \\ &= |\sum \sigma_i| \end{aligned}$$

then

$$\begin{aligned}
\hat{Rad}(G) &= \frac{1}{N}\mathbb{E}_{\sigma_1,\cdots,\sigma_N\sim U(\{1,-1\})}\sup_{g\in G}\sum \sigma_i g(z_i) \\
&= \frac{1}{N}\mathbb{E}_{\sigma_1,\cdots,\sigma_N}|\sum \sigma_i| \\
&= \frac{1}{N}\mathbb{E}_{\sigma_1,\cdots,\sigma_N}\sqrt{(\sum \sigma_i)^2} \\
&\leq \frac{1}{N}\sqrt{\mathrm{Var}(\sum \sigma_i)} \\
&= \frac{1}{N}\sqrt{(N\mathrm{Var}(\sigma_\mathrm{i}))} \\
&= \frac{1}{\sqrt{N}}
\end{aligned}$$

2. Let us consider a dataset of 3 points. If we go through all the possibilites of $\sigma_1, \sigma_2, \sigma_3 \in \{-1, 1\}$, we will see that $\sum \sigma_i g(x_i) = 3$ always. In general, $\forall \sigma_1, \cdots, \sigma_N$, $\sup_g \sum \sigma_i g(x_i) = N$, and $\hat{Rad}_S(G) = 1$. In this case, the bound is useless !

Theorem 4. *For any P, for any $\mathcal{F} = \{x \mapsto x^T\theta : ||\theta||_2 \leq w_2\}$, for any 1-Lipschitz loss (hinge, logistic loss, $\cdots$),*

$$UnRep(\mathcal{F}, S) \leq \frac{W_2 X_2}{\sqrt{N}} + 4X_2\sqrt{\frac{2}{N}\log\frac{2}{\delta}}$$

where $X_2 = \sup_{x\in X} ||x||_2$.