Fondamentaux de l'Apprentissage Automatique

Lecturer: Liva Ralaivola | Lecture n°3 #
Scribe: Paul Caucheteux | 12/10/2023

1 Introduction

1.1 What do we want to achieve ?

In this lecture, we want to find a uniform generalization bound for the error of a prediction function $\hat{f}$ which was learnt.

Let us denote $error(\hat{f})_{gen}$ the error of $\hat{f}$ on an other dataset of size n from the one that it was learnt, S the data from which $\hat{f}$ was learnt and $\mathcal{F}$ the class of function of $\hat{f}$. Thus the goal of this lecture is to prove :

$$\forall \hat{f} \in \mathcal{F} \qquad error(\hat{f})_{gen} \leq error_{empirical}(\hat{f}, S) + \epsilon(n, \mathcal{F}, ...) \tag{1}$$

We can already state the thing that we expect :
- If $\mathcal{F}$ is too big, we are going to have over fitting because ϵ depends on $\mathcal{F}$;
- ϵ will decrease as n increases.

1.2 Typical settings of Machine Learning

- $(X, Y) \sim D$, where D is a fixed unknown distribution ;
- $S = (x_i, y_i)_{i=1}^n$, where (x_i, y_i) are independent copies of (X, Y).

In others words, the training sample are *i.i.d* (independent and identically distributed).

1.3 Ultimate goal/criterion

We want to minimize the "l-risk" : $R_l(\hat{f}) := \mathbb{E}_{X,Y\sim D}[l(f(X), Y)]$ with l a loss function.

Example : For binary classification : $\mathcal{F} \subseteq \{-1, 1\}^X \quad l(f(x), y) = \mathbb{1}_{\{f(x)\neq y\}}$
or $\quad \mathcal{F} \subseteq \mathbb{R}^X \quad l(f(x), y) = \mathbb{1}_{\{f(x)y\leq 0\}}$

As we don't know D, we use S to gather information. The bound we are looking for is then :

$$\text{with proba } (1-\delta), \forall f \in \mathcal{F}, \quad R_l(f) \leq \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i) + \epsilon(n, \delta, \mathcal{F}) \tag{2}$$

There are a few things that we have to notice here :

- What appears here is the connection between an empirical quantity and its expectation, it's called a "*concentration* inequality".
- $\frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i)$ is the empirical risk ;
- We are looking to bound the **true risk** ;
- The order of the quantifiers, here "with proba $(1-\delta), \forall f \in \mathcal{F}$" and not vice-versa, is very important.

2 Hoeffding inequality (1963)

Theorem 1. *$X_1, X_2, ..., X_n$ are independent random variable (not necessarily i.i.d) such that : $\forall i \in \{1, .., n\}, \exists a_i, b_i$ such that $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$, let $S_n = \sum_{i=1}^n S_i$ then :*

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^n (a_i - b_i)^2})$$

and

$$\mathbb{P}(\mathbb{E}(S_n) - S_n \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^n (a_i - b_i)^2})$$

combining both inequalities we obtain : $\mathbb{P}(|\mathbb{E}(S_n) - S_n| \geq \epsilon) \leq 2\exp(-\frac{2\epsilon^2}{\sum_{i=1}^n (a_i - b_i)^2})$

Ab important particular case of the previous theorem is :

Proposition 1. *$X_1, X_2, ..., X_n$ are **i.i.d** such that : $\forall i \in \{1, .., n\}$, $\mathbb{P}(0 \leq X_i \leq 1) = 1$, let $\mu = \mathbb{E}(X_1)(= \mathbb{E}(X_2) = ... = \mathbb{E}(X_n))$ then :*

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

and

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

combining both inequalities we obtain : $\mathbb{P}(|\frac{1}{n}\sum_{i=1}^n X_i - \mu| \geq \epsilon) \leq 2\exp(-2n\epsilon^2)$

Example : Coin toss biased : Head(=0) and Tail(=1), we want to estimate the bias μ (wich is also the expectation of X_i) with probability $(1 - \delta)$:

1. $\mathbb{P}(|\frac{1}{n}\sum_{i=1}^n X_i - \mu| \geq \epsilon) \leq \exp(-2n\epsilon^2)$
2. To get an estimation on μ it suffices that : $2\exp(-2n\epsilon^2) \leq \delta \Leftrightarrow \epsilon \geq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$

   ⇒ with probability $(1 - \delta)$, $\mu \in [\frac{1}{n}\sum_{i=1}^n X_i \pm \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}]$

Remark. *There is another concentration inequality : the McDiarmid's inequality. It is a generalization of Hoeffding to the case where the function we look at is more complex.*
It provides an inequality of the following form :

$$\mathbb{P}(\phi(X_1, ..., X_n - \mathbb{E}[\phi(X_1, ..., X_n)] \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^n cste_i})$$

3 Proof of Hoeffding inequality

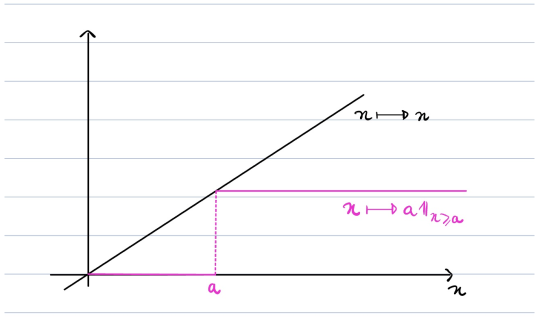
To prove Hoeffding we are going to prove 3 preliminaries results : the Markov inequality, the Laplace transform and the Hoeffding lemma. After this it will be easy to conclude.

3.1 Preliminaries results

Lemma 1 (Markov inequality)**.** *Let X be a random variable taking non-negatives values ($\mathbb{P}(X \geq 0) = 1$) and such that $\mathbb{E}[X] < +\infty$ then :*

$$\forall a > 0, \quad \mathbb{P}(X \geq a)) \leq \frac{\mathbb{E}[X]}{a}$$

Preuve du Lemme 1. $\forall x, x \geq a\mathbb{1}_{x\geq a}$,

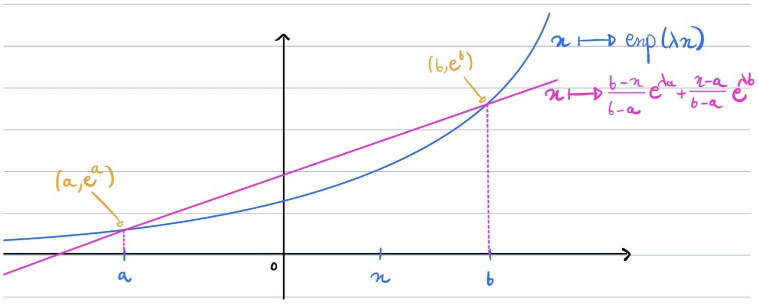


$$\Rightarrow \mathbb{E}[X] \geq \mathbb{E}[a\mathbb{1}_{X\geq a}] = a\mathbb{E}[\mathbb{1}_{X\geq a}] = \mathbb{P}[X \geq a]$$ □

Lemma 2 (Hoeffding lemma)**.** *Let X be a random variable such that $\mathbb{E}[X] = 0$, and $\exists a < 0, b > 0$, such that $\mathbb{P}(a \leq X \leq b) = 1$ then :*

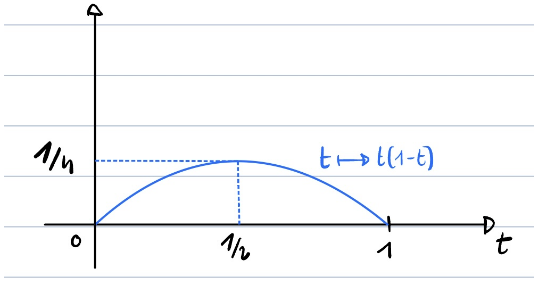
$$\forall \lambda > 0, \quad \mathbb{E}(e^{\lambda X}) \leq \exp(\frac{(\lambda(b-a))^2}{8})$$

Proof of Lemma 2. Using the convexity of $x \mapsto e^{\lambda x}$:
$$\forall x \in [a,b], \quad e^{\lambda x} \leq \tfrac{b-x}{b-a}e^{\lambda a} + \tfrac{x-a}{b-a}e^{\lambda b}$$


$x \mapsto \exp(\lambda x)$
(b, e^b)
$x \mapsto \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$
(a, e^a)
a
0
x
b

Since $\mathbb{P}(a \leq X \leq b) = 1$ we have :
$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[\tfrac{b-X}{b-a}e^{\lambda a} + \tfrac{X-a}{b-a}e^{\lambda b}]$$
Using the linearity of the expectation and $\mathbb{E}[X] = 0$:
$$\mathbb{E}[e^{\lambda X}] \leq \tfrac{b}{b-a}e^{\lambda a} - \tfrac{a}{b-a}e^{\lambda b}$$
By noting $h := \lambda(b-a)$, $p := \frac{a}{b-a} > 0$ and $L_p(h) := -ph + \log(1 - p + pe^h)$, we get :
$$\mathbb{E}[e^{\lambda X}] \leq e^{L_p(h)}$$

By using the Taylor expansion we have : $\exists v \in [0, h], L_p(h) = L_p(0) + hL_p'(0) + \frac{1}{2}h^2 L_p''(v)$ with $L_p(0) = 0$, $L_p(0) = 0$ and $L_p''(v) = \frac{(1-p)pe^v}{(1-p+pe^v)^2} = t(1-t) \leq \frac{1}{4}$ with $t = \frac{1-p}{1-p+pe^v}$.



Then :

$$L_p(h) \leq \frac{1}{8}h^2 = \frac{1}{8}\lambda^2(b-a)^2$$

We conclude by the fact that $x \mapsto e^x$ is increasing.

□

Lemma 3. *Let $X_1, ..., X_n$ be independent random variables, then :*

$$\mathbb{E}[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

3.2 The proof

$\forall \lambda \geq 0$,

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \epsilon) &= \mathbb{P}(\exp(\lambda(S_n - \mathbb{E}[S_n])) \geq \exp(\lambda\epsilon)) \\
&= \mathbb{P}(\exp(\lambda(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]))) \geq \exp(\lambda\epsilon)) \\
&\leq \frac{\mathbb{E}[\exp(\lambda \sum_{i=1}^{n} Z_i)]}{\exp(\lambda\epsilon)} \text{ with } \mu_i = \mathbb{E}[X_i] \text{ and } Z_i = X_i - \mu_i \text{ and applying Markov1} \\
&= \frac{\mathbb{E}[\prod_{i=1}^{n} \exp(\lambda Z_i)]}{\exp(\lambda\epsilon)} \\
&= \frac{\prod_{i=1}^{n} \mathbb{E}[\exp(\lambda Z_i)]}{\exp(\lambda\epsilon)} \quad \text{because } Z_i \text{ are independent and applying the Lemma 3} \\
&\leq \frac{\prod_{i=1}^{n} \exp(\frac{\lambda^2}{8}((b_i - \mu_i) - (a_i - \mu_i))^2)}{\exp(\lambda\epsilon)} \quad \text{using the Hoeffding lemma 2} \\
&= \exp(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\epsilon)
\end{aligned}$$

We have to minimize : $g(\lambda) = \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\epsilon$ with respect to λ to find the optimal bound.
By taking the derivative we easily get : $g'(\lambda) = 0 \Leftrightarrow \lambda = \frac{4\epsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}$
And so with this optimal lambda we find :

$$\mathbb{P}(\mathbb{E}(S_n) - S_n \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2})$$

Remark.

- *We could apply the Hoeffding lemma in the 6th line because :* $\mathbb{P}(Z_i \in [a_i - \mu_i, b_i - \mu_i]) = 1$ *and* $a_i - \mu_i \leq 0, b_i - \mu_i \geq 0.$
- *We prove the other side of the inequality by doing pretty much the same thing, and so we have the result with absolutes values.*

4 Conclusion

In this lecture we have proved the Hoeffding inequality, thanks to the Hoeffding Lemma and the Markov Lemma. This inequality permits to bound the difference of the sum of i.i.d variables. We may use it later to try to bound the difference of the empirical risk and the real risk of a classifier.