Fondamentaux de l'Apprentissage Automatique

Lecturer: Yann Chevaleyre
Scribe: Leroy Amélie

Lecture n°1 #
26/10/2023

1 Introduction

In this lesson, we will explore techniques for deriving bounds on the true risk.
Let $S = (X_i, Y_i)_{i=1}^n$ be an IID sample, our goal is to have this result :
With probability $1 - \delta$

$$\forall f \in \mathcal{F}, R(f, D) \leq \hat{\mathcal{R}}_n(f, S) + \varepsilon(\delta, n, \mathcal{C}(\mathcal{F}))$$

or, equivalently ("contraposée")

$$\mathbb{P}_{S \sim D^n} \left[\exists f \in \mathcal{F} : R(f, D) \geq \hat{\mathcal{R}}_n(f, S) + \varepsilon(\delta, n, \mathcal{C}(\mathcal{F}))\right] \leq \delta$$

We will study such bounds in :

- the case of countable and finite $\mathcal{F}$ ($|\mathcal{F}| < +\infty$).
- the case where we don't have $|\mathcal{F}| < +\infty$, and where we are going to use Vapnik-Chervonenkis Dimension.

2 The case $|\mathcal{F}| < +\infty$

2.1 A first bound.

Let $f \in \mathcal{F}$:

$$\begin{aligned}
\hat{\mathcal{R}}_n(f, S) &:= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i) \neq Y_i} \quad \text{We can use other loss functions.} \\
R_D(f) = R(f, D) &:= \mathbb{E}\left[\mathbb{1}_{f(X_1) \neq Y_1}\right] = \mathbb{P}\left[f(X_1) \neq Y_1\right] \\
&= \mathbb{E}_S\left[\hat{\mathcal{R}}_n(f, S)\right] \quad \text{By linearity of expectation and } S \text{ is IID.}
\end{aligned}$$

According to the Hoeffding inequality, since $\mathbb{1}_{f(X_i) \neq Y_i}$ are IID, $\hat{\mathcal{R}}_n(f, S)$ is the sample average and $\mu = R(f, D)$,

$$\forall f \in \mathcal{F} : \mathbb{P}_S\left(|\hat{\mathcal{R}}_n(f, S) - R(f, D)| \geq \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right)$$

or using the one-sided inequality :

$$\mathbb{P}\left(R(f, D) - \hat{\mathcal{R}}_n(f, S) \geq \varepsilon\right) \leq \exp\left(-2n\varepsilon^2\right)$$

Thus, given that previous result, we can state that $\forall f \in \mathcal{F}$, with probability $1 - \delta$:

$$R(f, D) \leq \hat{\mathcal{R}}_n(f, S) + \sqrt{\frac{1}{2n} \ln(\frac{1}{\delta})} \tag{1}$$

Démonstration. Let $\exp(-2n\varepsilon^2) \leq \delta$

$$\begin{aligned}\exp(-2n\varepsilon^2) = \delta &\iff -2n\varepsilon^2 = \ln(\delta)\\ &\iff 2n\varepsilon^2 = \ln(\frac{1}{\delta})\\ &\iff \varepsilon = \sqrt{\frac{1}{2n}\ln(\frac{1}{\delta})}\end{aligned}$$

□

Remarks on inequality (1) :

- the rate of the bound is $\mathcal{O}(\frac{1}{\sqrt{n}})$
- it's not a uniform generalization bound because "$\forall f \in \mathcal{F}$" and "probability $1-\delta$" are inverted.

2.2 A uniform bound.

To get a uniform generalization bound , we would rather look at achieving a result like :

$$\mathbb{P}\left(\exists f \in \mathcal{F} : R(f, D) - \hat{\mathcal{R}}_n(f, S) \geq \varepsilon\right) \leq \delta \tag{2}$$

As a reminder : $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

$$\begin{aligned}&\mathbb{P}\left(\exists f \in \mathcal{F} : R(f, D) - \hat{\mathcal{R}}_n(f, S) \geq \varepsilon\right)\\ &= \mathbb{P}\left(\bigcup_{p=1}^{|\mathcal{F}|}\{R(f_p, D) - \hat{\mathcal{R}}_n(f_p, S) \geq \varepsilon\}\right)\\ &\leq \sum_{p=1}^{|\mathcal{F}|}\mathbb{P}\left(R(f_p, D) - \hat{\mathcal{R}}_n(f_p, S) \geq \varepsilon\right) \quad \text{Union bound}\\ &\leq \sum_{p=1}^{|\mathcal{F}|}\exp(-2n\varepsilon^2) \quad \text{Hoeffding inequality}\\ &= |\mathcal{F}|\exp(-2n\varepsilon^2)\end{aligned}$$

As in the previous proof, we are solving

$$\begin{aligned}\delta &= |\mathcal{F}|\exp(-2n\varepsilon^2)\\ \iff \varepsilon &= \sqrt{\frac{1}{2n}\ln\frac{|\mathcal{F}|}{\delta}}\end{aligned}$$

Given this ε, you then have

$$\mathbb{P}\left(\exists f \in \mathcal{F} : R(f, D) - \hat{\mathcal{R}}_n(f, S) \geq \sqrt{\frac{1}{2n}\ln\frac{|\mathcal{F}|}{\delta}}\right) \leq \delta$$

So that, with probability $1-\delta$:

$$\forall f \in \mathcal{F}, R(f, D) \leq \hat{\mathcal{R}}_n(f, S) + \sqrt{\frac{1}{2n}\ln\frac{|\mathcal{F}|}{\delta}} \tag{3}$$

Remarks

— The result is obtained with the "union bound"
— We used the fact that $|\mathcal{F}| < +\infty$
— Here, $\mathcal{C}(\mathcal{F}) = |\mathcal{F}|$
— In practice, it is very rare to be in the case where $|\mathcal{F}| < +\infty$

To cope with the situation where $|\mathcal{F}| < +\infty$ does not hold, we will use another tool to find a bound, the VC dimension.

3 The Vapnik-Chervonenkis dimension/ VC dimension

3.1 High-level idea

$\mathcal{F} \subseteq \{\mathcal{X} \mapsto \{-1, +1\}\}$ $\left(\text{ex} : \mathcal{F} = \left\{x \mapsto \text{sign}(w \circ x), w \in \mathbb{R}^d\right\}\right)$

Observe that if you have n points $S = \{x_1, ..., x_n\}$, then

$$|\mathcal{F}_S := \{(f(x_1), ..., f(x_n)) : f \in \mathcal{F}\}| \leq 2^n \quad \text{(because binary classification)}$$

In VC dimension, we are going to be looking at hte following situation : $\sup\limits_{S:|S|=n} |\mathcal{F}_S| < 2^n$

3.2 VC dimension

Definition 1. *Restriction of $\mathcal{F}$ to a sample :*

$$\begin{aligned} \mathcal{F} &\subseteq \{-1, +1\}^{\mathcal{X}} \, (\equiv \{\mathcal{X} \mapsto \{-1, +1\}) \\ S &= \{x_1, ..., x_n\}, \, x_i \in \mathcal{X} \; \forall i \\ \mathcal{F}_S &:= \{(f(x_1), ..., f(x_n)) : f \in \mathcal{F}\} \end{aligned}$$

Remark Sometimes in the literature you can see a "functional" way of writing things :

$$\mathcal{F}_S := \{(x_1, ..., x_n) \mapsto (f(x_1), ..., f(x_n)), f \in \mathcal{F}\}$$

Definition 2. *Shattered set*
Let $S = \{x_1 ... x_n\}$. We say that S is shattered by $\mathcal{F}$ if $|\mathcal{F}_S| = 2^n$.

In others words : you can realize all the labellings on S given $\mathcal{F}$.

Definition 3. *Vapnik-Chervonenkis dimension*
The Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$ is the size of the largest set that is shattered by $\mathcal{F}$.

It may happen that VC $\dim(\mathcal{F}) = +\infty$.

Remarks :

— VC dimension appeared in the 70's.
— Connected to "Computational Machine Learning".
— Connected to the Probably Approximately Correct (PAC) framework of learning, that took into consideration complexity (from a computer science point of view)/ NP classes/decidable problems.

3.3 Examples of VC dimension for some classes of functions

3.3.1 VC dimension of axis-aligned rectangles

$\mathcal{F}$ is the set of all axis aligned rectangles in $\mathbb{R}^2$ such that points inside the rectangle are classified as positive instances (+1) by the rectangle and those outside are classified as negative instances (-1).

Claim The VC dimension of Figure 1 $\mathcal{F}$ is 4.

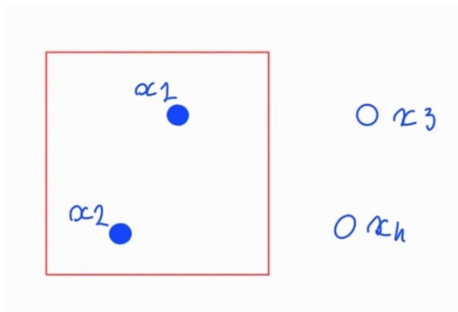


FIGURE 1 – Example of Classification with a rectangle.

We can see on Figure 2 that for n = 4, there are configurations of the dataset S which are not shattered. Indeed in the example given, it is not possible to realize all labellings : no rectangle can classify all positive examples as positive and classify the negative right.

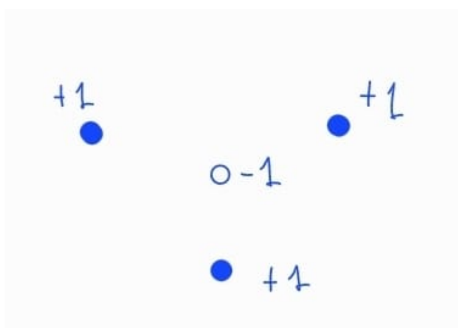


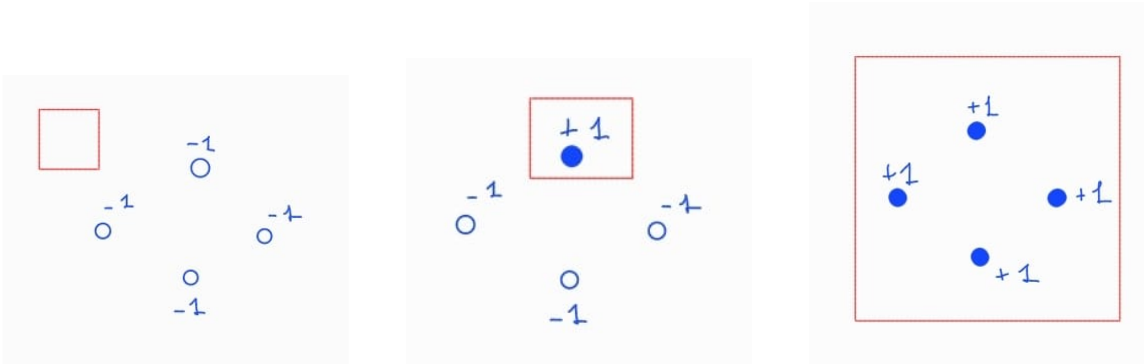
FIGURE 2 – Example of set of 4 points not shattered by $\mathcal{F}$

However, there exists a configuration S of 4 points such that all labellings are possible. This is the case for Figure 3. In that example, S is shattered by the class of rectangles. We can show how it is shattered graphically in the same Figure.

But if you take $|S| = 5$, then the subclass of $\mathcal{F}$ defined as axis-aligned rectangle delimited by the max and min values of x and y values (which does not lose in labelling power), has a configuration which cannot be realized. An illustration of this is given in Figure4. Thus VC dimension is 4.

3.3.2 VC dimension of Hyperplanes

Claim For a Hyperplane of dimension d the VC dimension is $d + 1$ (the result can be formaly proved by induction).



(a) Labelling 1 (b) Labelling 2 (c) Labelling 16

FIGURE 3 – 3 labellings of S among the 16 possible

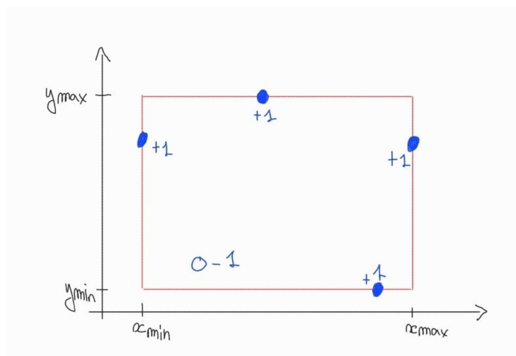


FIGURE 4 – Example of how an S is not shattered

In the case of $d = 2$, we can show that $VCdim(\mathcal{F}) = 3$. We show an example in figure 5.
For $|S| = 4$, we are never able to separate all configurations of labellings, because the XOR situation cannot be handled by the hyperplanes, as shown in Figure 6.

4 VC dimension and generalization error bound.

Definition 4. *Growth function The growth function* $\Pi_{\mathcal{F}} : \mathbb{N} \to \mathbb{N}$ *is*

$$\Pi_{\mathcal{F}}(n) := \max_{S \subseteq \mathcal{X} : |S| = n} |\mathcal{F}_S| \tag{4}$$

Remark if VCdim$(\mathcal{F}) = d$ then $\forall n \leq d, \Pi_{\mathcal{F}}(n) = 2^n$.

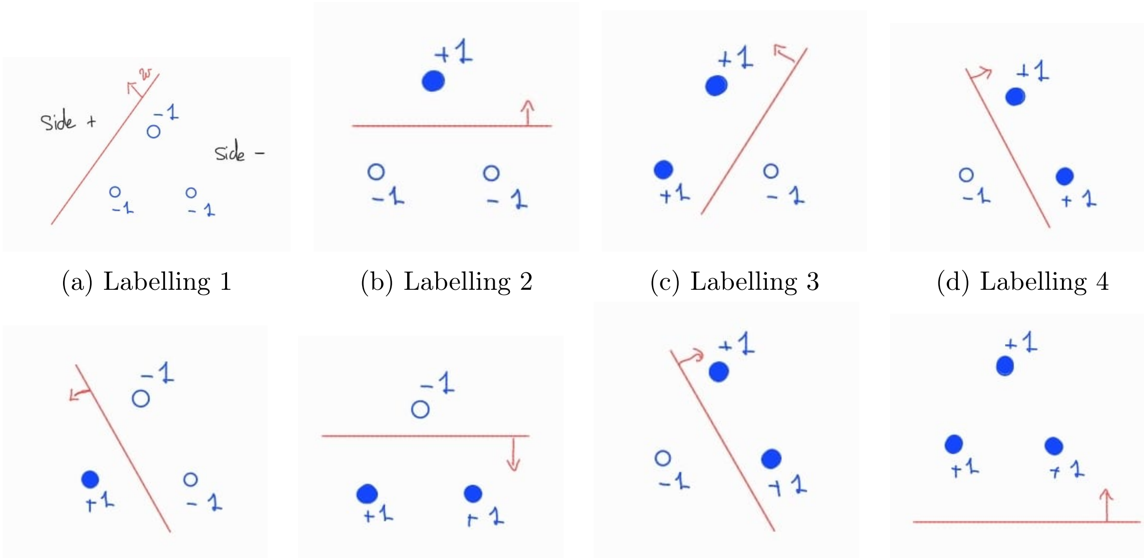
Theorem 1. *Let* $\mathcal{F} \subseteq \{-1, +1\}^{\mathcal{X}}$ *with* $d := VCdim(\mathcal{F}) < +\infty$. *With probability* $1 - \delta$ *:*

$$\forall f \in \mathcal{F}, \mathcal{R}(f, D) \leq \hat{\mathcal{R}_n}(f, D) + \sqrt{\frac{2d \ln\left(\frac{en}{d}\right)}{n}} + \mathcal{O}(\sqrt{\frac{1}{n} \ln\left(\frac{1}{\delta}\right)}) \tag{5}$$

where $\ln e = 1$

To prove the theorem, we have to use other theorem/lemma :

— Massart's Lemma



(a) Labelling 1 (b) Labelling 2 (c) Labelling 3 (d) Labelling 4

(e) Labelling 5 (f) Labelling 6 (g) Labelling 7 (h) Labelling 8

FIGURE 5 – All possible labellings of S

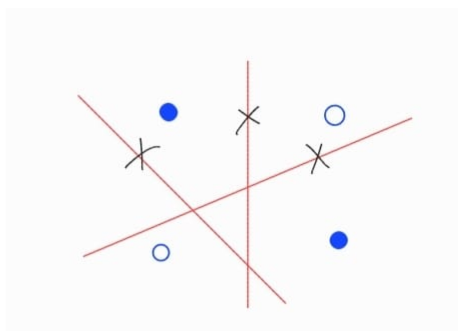


FIGURE 6 – Example of how an S is not shattered

- Bound on the growth function using the Rademacher Complexity
- Sauer's Lemma

Reminder : With the Rademacher complexity :

$$\forall f \in \mathcal{F}, \mathcal{R}\left(f, D\right) \leq \hat{\mathcal{R}_n}\left(f, S\right) + \hat{Rad}\left(\mathcal{F}, S\right) + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right)$$

Lemma 1. *Massart's Lemma*

Let $A \subseteq \mathbb{R}^n$ and $\varepsilon_1 ... \varepsilon_n$ independant Rademacher variables ($P(\varepsilon_i = +1) = P(\varepsilon_i = -1) = \frac{1}{2}$).

Let $r := \sup_{a \in A} \|a\|_2$ then

$$\mathbb{E}_{\varepsilon_1 ... \varepsilon_n}\left[\sup_{a \in A}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i a_i\right)\right] \leq \frac{r\sqrt{2\ln(|A|)}}{n} \tag{6}$$

Démonstration. a_i is the i-th component of a.

$$\begin{aligned}
\exp(\lambda\mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\sup_{a\in A}\left(\sum_{i=1}^{n}\varepsilon_i a_i\right)\right]) &\leq \mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\exp\left(\lambda\sup_{a\in A}\left(\sum_{i=1}^{n}\varepsilon_i a_i\right)\right)\right] \text{ By convex of exp and Jensen inequality.}\\
&= \mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\sup_{a\in A}\left(\exp(\lambda\sum_{i=1}^{n}\varepsilon_i a_i)\right)\right] \text{ Because exp is increasing.}\\
&\leq \mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\sum_{a\in A}\left(\exp(\lambda\sum_{i=1}^{n}\varepsilon_i a_i)\right)\right]\\
&= \sum_{a\in A}\mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\left(\exp(\lambda\sum_{i=1}^{n}\varepsilon_i a_i)\right)\right] \text{ By linearity of expectaion.}\\
&= \sum_{a\in A}\mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\prod_{i=1}^{n}(\exp(\lambda\varepsilon_i a_i))\right]\\
&= \sum_{a\in A}\prod_{i=1}^{n}\mathbb{E}_{\varepsilon_i}\left[\exp(\lambda\varepsilon_i a_i)\right]\\
&= \sum_{a\in A}\prod_{i=1}^{n}\left[\frac{1}{2}\exp(-\lambda a_i)+\frac{1}{2}\exp(\lambda a_i)\right]\\
&\leq \sum_{a\in A}\prod_{i=1}^{n}\exp\left(\frac{\lambda^2 a_i^2}{2}\right)\\
&= \sum_{a\in A}\exp\left(\frac{\lambda^2}{2}\sum_{i=1}^{n}a_i^2\right)\\
&= \sum_{a\in A}\exp\left(\frac{\lambda^2}{2}\|a\|_2^2\right)\\
&\leq \sum_{a\in A}\exp\left(\frac{\lambda^2}{2}r^2\right)\\
&= |A|\exp\left(\frac{\lambda^2}{2}r^2\right)
\end{aligned}$$

We thus have

$$\begin{aligned}
\exp(\lambda\mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\sup_{a\in A}\left(\sum_{i=1}^{n}\varepsilon_i a_i\right)\right]) &\leq |A|\exp\left(\frac{\lambda^2}{2}r^2\right)\\
\Rightarrow \mathbb{E}_{\varepsilon_1\ldots\varepsilon_n}\left[\sup_{a\in A}\left(\sum_{i=1}^{n}\varepsilon_i a_i\right)\right] &\leq \frac{\ln(|A|)}{\lambda}+\lambda r
\end{aligned}$$

The right-bound side is minimal when

$$\lambda = \sqrt{\frac{2\ln(|A|)}{r}}$$

which gives us the result stated in the theorem. □

Lemma 2. *Let $S = \{x_1 ... x_n\}$.*

$$\hat{Rad}(\mathcal{F}, S) \leq \sqrt{\frac{2\ln(\Pi_{\mathcal{F}}(n))}{n}} \tag{7}$$

With $\hat{Rad}(\mathcal{F}, S) := \mathbb{E}_{\sigma_1 ... \sigma_n}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(xi)\right)\right]$.

Démonstration. $\forall a \in \mathcal{F}_S : \|a\|_2 = \sqrt{\sum_{i=1}^{n}(a_i^2)} = \sqrt{n}$, since $\mathcal{F} \subseteq \{-1,+1\}^{\mathcal{X}}$.

$$\begin{aligned}\hat{Rad}(\mathcal{F}, S) = \mathbb{E}_{\sigma_1 ... \sigma_n}\left[\sup_{a\in\mathcal{F}_S}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right)\right] &\leq \sqrt{n}\frac{\sqrt{2\ln(|\mathcal{F}_S|)}}{n} \quad \text{By Massart's Lemma.}\\ &= \sqrt{\frac{2\ln(|\mathcal{F}_S|)}{n}}\end{aligned}$$

$|\mathcal{F}_S|$ is bounded by growth function, it concludes the proof. □

Lemma 3. *Sauer's Lemma*
Let $\mathcal{F} \subseteq \{-1,+1\}^{\mathcal{X}}$ such that VCdim($\mathcal{F}$) $\leq d < +\infty$.

$$\forall n \geq d, \Pi_{\mathcal{F}}(n) \leq \sum_{i=1}^{n}\binom{n}{i} \leq \left(\frac{en}{d}\right)^d \tag{8}$$

With $\binom{n}{i} = \frac{n!}{i!(n-i)!}$ *(i choose n).*

We don't prove this Lemma.

Proof of theorem 1

$$\begin{aligned}\forall f \in \mathcal{F}, \mathcal{R}(f, D) &\leq \hat{\mathcal{R}}_n(f, S) + \hat{Rad}(\mathcal{F}, S) + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right)\\ &\leq \hat{\mathcal{R}}_n(f, S) + \hat{Rad}(\mathcal{F}, S)\sqrt{\frac{2\ln(\Pi_{\mathcal{F}}(n))}{n}} + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right) \quad \text{By Lemma 2}\\ &\leq \hat{\mathcal{R}}_n(f, S) + \sqrt{\frac{2d\ln\left(\frac{en}{d}\right)}{n}} + \mathcal{O}\left(\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right) \quad \text{By Lemma 3}\end{aligned}$$