Fondamentaux de l'Apprentissage Automatique

Lecturer: Liva Ralaivola
Scribe: Gervreau Augustin

Lecture n°6 #
26/10/2023

Table of Contents

1 Introduction

1.1 Reminders

Hoeffding Inequality - corollary

$$\begin{gathered}\text{Let } X_1, ..., X_n \text{ be IID r.v. such that } \mathbb{P}(0 \leq X_1 \leq 1) = 1 \\ \mu := \mathbb{E}[X_1] (= \mathbb{E}[X_2] = ... = \mathbb{E}[X_n]) \\ \forall \epsilon > 0,\ \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right) \leq \exp\left(-2n\epsilon^2\right) \\ \mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \epsilon\right) \leq \exp\left(-2n\epsilon^2\right)\end{gathered}$$

In books or ML litterature, this is in the form

$$\forall \epsilon > 0,\ \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right)$$

The Hoeffding inequality provides an interesting bound since it becomes smaller as n grows, showing the estimation becomes better as n grows.

1.2 This course

This lesson goes over techniques to find bounds on the true risk of classifiers (with a focus on the binary classification case).

formally, let $S = (X_i, Y_i)_{i=1}^n$ be an IID sample, we want for this S with probability $1 - \delta$

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

or equivalently, by contrapositive,

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

The greater $\mathcal{C}(\mathcal{F})$ is, the better the approximation is (the closer f_{opt} can be from f^*, not to be confused with $\hat{f}$). S is a sample dataset, in the sense that it is the observation of a possible dataset of points in $\mathcal{X}$ among all possible datasets.

We will study such bounds in :

— the case of countable and finite $\mathcal{F}$ ($|\mathcal{F}| < +\infty$).

— the case where $|\mathcal{F}| = +\infty$, where we will use Vapnik-Chervonenkis Dimension (VC dimension or VCdim)

2 When $|\mathcal{F}| < +\infty$

2.1 A first bound.

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

$$\begin{aligned} \hat{\mathcal{R}}_n(f, S) :&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i) \neq Y_i} \\ (\mathcal{R}_D(f)) = \mathcal{R}(f, D) :&= \mathbb{E}\left[\mathbb{1}_{f(X_1) \neq Y_1}\right] = \mathbb{P}\left[f(X_1) \neq Y_1\right] \\ &= \mathbb{E}_S\left[\hat{\mathcal{R}}_n(f, S)\right] \end{aligned}$$

Because $\mathbb{E}$ is linear and S is IID.

Note we could use any other loss functions instead of $\mathbb{1}_{f(X_i) \neq Y_i}$.

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 $\mu = \mathcal{R}(f, D)$,

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

or equivalently

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

Thus, $\forall f \in \mathcal{F}$, with probability $1 - \delta$:

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

Démonstration. equation 1

Let $\exp(-2n\epsilon^2) \leq \delta$

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

□

Remarks on inequality 1 :
— the *rate* of the bound on the risk is $\mathcal{O}(\frac{1}{\sqrt{n}})$
— it's not a uniform generalization bound. The bound is different for each f chosen. We would prefer to find a bound where "$\forall f \in \mathcal{F}$" and "with probability $1-\delta$" are inverted.

2.2 A uniform bound.

To get a uniform generalization bound , we look at the union of all bounds.
As a reminder : $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

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

So in the same way as in the previous proof,

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

Thus, for this ϵ,

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

So, with probability $1-\delta$:

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

Remarks

— The result is obtained with the "union bound"

— The result is interesting because $|\mathcal{F}| < +\infty$

— 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 need to use another tool to find a bound : the VC dimension.

3 The Vapnik-Chervonenkis dimension/ VC dimension

3.1 High-level idea

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

If you have n points $S = \{x_1, ..., x_n\}$, then

$$|\mathcal{F}_S := \{(f(x_1), ..., f(x_n)) : f \in \mathcal{F}\}| \leq 2^n, \text{ (number of possible binary vectors of size n)}$$

$|\mathcal{F}|$ could be bigger but two functions in $\mathcal{F}$ with the same output vector are considered to be "the same".

VC dimension looks at the situation where $\sup_{S:|S|=n} |\mathcal{F}_S| < 2^n$

Note we can extend this easily to the multiclass setting. The case of regression is outside of the scope of the course but it can be treated with VC dimension.

3.2 VC dimension

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

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

Note Sometimes in the literature it is written in the "functional" way :

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

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($\mathcal{F}$) $= +\infty$.

Remark

— VC dimension appeared in the 70's.

— Connected to "Computational Machine Learning", thus the formulation and proof are very combinatorial.

— 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 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 such that points ($\in \mathbb{R}^2$) within the rectangle are classified as positive instances (+1) by the rectangle and those outside are classified as negative instances (-1). An example is given in Figure 1.

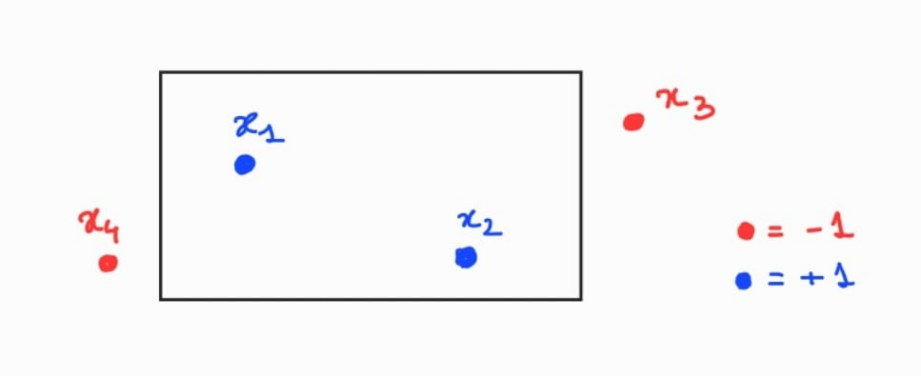


FIGURE 1 – Example of Classification with a rectangle in $\mathcal{F}$

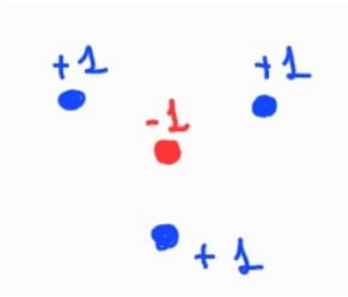
Claim The VC dimension of this $\mathcal{F}$ is 4.

We can see on Figure 2a 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. However, VCdim looks at the supremum over all datasets of the size of a shattered S so it suffices to have one instance of S which is shattered.

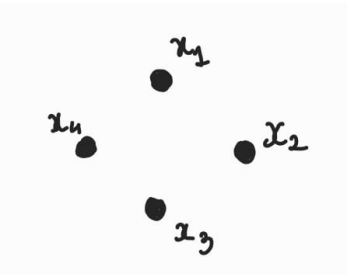
There is indeed a configuration S of 4 points such that all labellings are possible. This is the case for S in Figure 2b. In that example, S is shattered by the class of rectangles. We can see how it is shattered in Figure 3.

In the case of $|S| = 5$, 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, and thus has the same VCdim as the case of any axis aligned rectangle), has a configuration which cannot be realized. An illustration of this is given in Figure 4. Each configuration will always have such a bounded box where the point not imposing the bound is problematic.

Thus VC dimension is 4.

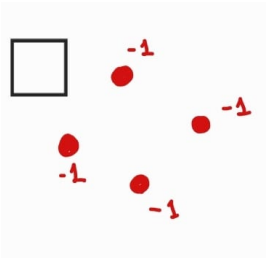


(a) Example of set of 4 points not shattered by $\mathcal{F}$

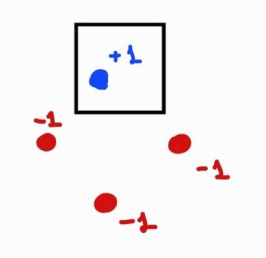


(b) Example of set of 4 points shattered by $\mathcal{F}$

FIGURE 2 – 2 configurations of set S with $|S| = 4$

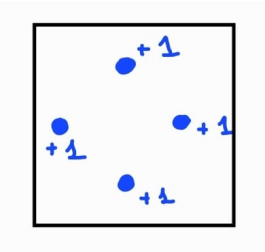


(a) Labelling 1



(b) Labelling 2

(c) ...



(d) Labelling 16

FIGURE 3 – All possible labellings of S from Figure 2b

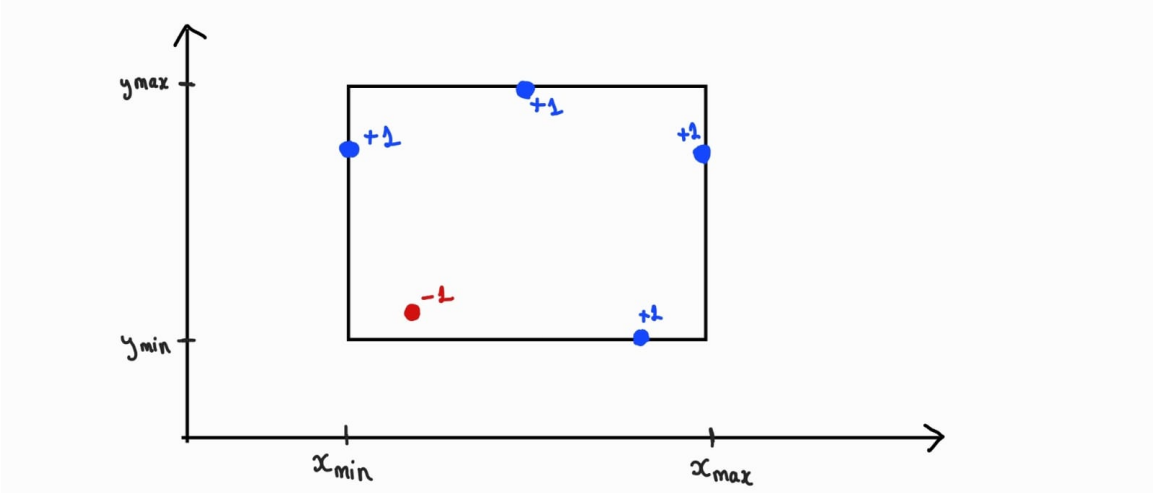
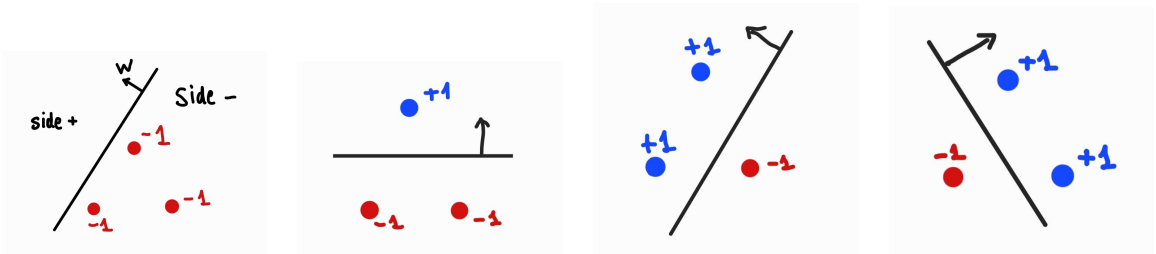


FIGURE 4 – Example of how an S is not shattered

3.3.2 VC dimension of Hyperplanes

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

In the case of $d = 2$, we can show Graphically that $VCdim(\mathcal{F}) = 3$. We show an example in figure 5 of an S which is shattered with $|S| = 3$.



(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

For $|S| = 4$, we are never able to separate all configurations of labellings as shown in Figure 6. This is because it is always possible to make a XOR situation appear which cannot be handled by the hyperplanes.

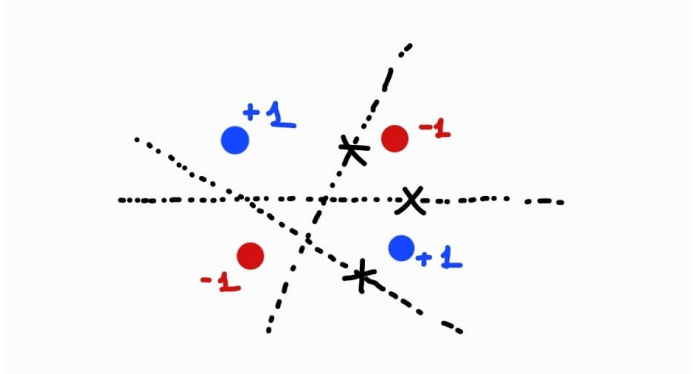


FIGURE 6 – Example of how an S is not shattered

4 VC dimension and generalization error bound.

Definition 4. *Growth function*

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

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

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* $1 - \delta$ *:*

$$\forall f \in \mathcal{F},\ \mathcal{R}(f, D) \leq \hat{\mathcal{R}}_n(f, S) + \sqrt{\frac{2d \log\left(\frac{en}{d}\right)}{n}} + \mathcal{O}\left(\sqrt{\frac{1}{n} \log\left(\frac{1}{\delta}\right)}\right) \tag{3}$$

The end of this lesson is dedicated to proving Theorem 1.
To prove the theorem, we need :
— Massart's Lemma
— Bound on the growth function using the Rademacher Complexity
— Sauer's Lemma (proof for this is outside of the scope of this course)

Reminder with the Rademacher complexity :

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

Lemma 1. *Massart's Lemma*
Let $A \subseteq \mathbb{R}^n$ *and* $\epsilon_1, ..., \epsilon_n$ *independant Rademacher variables (i.e.* $\mathbb{P}(\epsilon_i = +1) = \mathbb{P}(\epsilon_i = -1) = \frac{1}{2}$*).*
Let $r := \sup_{a \in A} \|a\|_2$ *then*

$$\mathbb{E}_{\epsilon_1, ..., \epsilon_n}\left[\sup_{a \in A}\left(\frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i\right)\right] \leq \frac{r\sqrt{2\log(|A|)}}{n} \tag{4}$$

with a_i *the* i*-th component of* a.

Démonstration. Lemma 1 (Massart's Lemma)

$$\begin{aligned}
\exp\left(\lambda\mathbb{E}_\epsilon\left[\sup_{a\in A}\left(\sum_{i=1}^n \epsilon_i a_i\right)\right]\right) &\leq \mathbb{E}_\epsilon\left[\exp\left(\lambda\sup_{a\in A}\left(\sum_{i=1}^n \epsilon_i a_i\right)\right)\right] \text{ (convexity of exp and Jensen inequality)}\\
&= \mathbb{E}_\epsilon\left[\sup_{a\in A}\left(\exp(\lambda\sum_{i=1}^n \epsilon_i a_i)\right)\right] \text{ (exp is increasing)}\\
&\leq \mathbb{E}_\epsilon\left[\sum_{a\in A}\left(\exp(\lambda\sum_{i=1}^n \epsilon_i a_i)\right)\right]\\
&= \sum_{a\in A}\mathbb{E}_\epsilon\left[\exp(\lambda\sum_{i=1}^n \epsilon_i a_i)\right] \text{ (linearity of expectation)}\\
&= \sum_{a\in A}\mathbb{E}_\epsilon\left[\prod_{i=1}^n (\exp(\lambda\epsilon_i a_i))\right]\\
&= \sum_{a\in A}\prod_{i=1}^n \mathbb{E}_{\epsilon_i}\left[\exp(\lambda\epsilon_i a_i)\right]\\
&= \sum_{a\in A}\prod_{i=1}^n \left[\frac{1}{2}\exp(-\lambda a_i)+\frac{1}{2}\exp(\lambda a_i)\right]\\
&\leq \sum_{a\in A}\prod_{i=1}^n \exp\left(\frac{\lambda^2 a_i^2}{2}\right),\ \left(\frac{\exp x+\exp -x}{2}\leq \exp\frac{x^2}{2}\right)\\
&= \sum_{a\in A}\exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n a_i^2\right)\\
&= \sum_{a\in A}\exp\left(\frac{\lambda^2}{2}\|a\|_2^2\right)\\
&\leq \sum_{a\in A}\exp\left(\frac{\lambda^2}{2}r^2\right)\\
&= |A|\exp\left(\frac{\lambda^2}{2}r^2\right)
\end{aligned}$$

We thus have

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

The right-bound side is minimal when

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

which gives us the result stated in the theorem. □

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

$$\hat{Rad}(\mathcal{F}, S) \leq \sqrt{\frac{2\log(\Pi_{\mathcal{F}}(n))}{n}} \tag{5}$$

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. Lemma 2

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

$$\begin{aligned}
\hat{Rad}(\mathcal{F}, S) &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(xi)\right)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{a\in\{(f(x_1),...,f(x_n)|f\in\mathcal{F}\}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{a\in\mathcal{F}_S}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right)\right] \text{ (by definition of } \mathcal{F}_S) \\
&\leq \sqrt{n}\frac{\sqrt{2\log(|\mathcal{F}_S|)}}{n} \quad \text{(by Massart's Lemma)} \\
&= \sqrt{\frac{2\log(|\mathcal{F}_S|)}{n}}
\end{aligned}$$

□

Lemma 3. *Sauer's Lemma*

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

$$\forall n \geq d,\ \Pi_{\mathcal{F}}(n) \leq \sum_{i=1}^{n}\binom{n}{i} \leq \left(\frac{en}{d}\right)^d \tag{6}$$

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

Démonstration. Theorem 1

By combining (5) and (6). We can easily explicit an upper bound for $\hat{Rad}(\mathcal{F}, S)$, and we get the expression in (3), proving theorem 1. □