Fondamentaux de l'Apprentissage Automatique

Lecturer: Liva Ralaivola
Scribe: BISWAS Suparna

Lecture n°5 #
26/10/2023

1 Introduction

1.1 Reminders

Hoeffdding inequality : For $X_1, ...X_n$ which is Independent Identically Distributed (IID) with $\mathbb{P}(0 \leq X_1 \leq 1) = 1$

$$\mu := \mathbb{E}X_1 (= \mathbb{E}X_2 = ... = \mathbb{E}X_n)$$

$\forall \varepsilon > 0$

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

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

As per triangle inequality : $\forall \varepsilon > 0$

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

1.2 Goal

We want to have results like the following one :
$S = \{(X_i, Y_i)\}_{i=1}^{n}$, IID sample (lets imagine binary classification) :
$\forall f \in \mathcal{F}$, with probability at least $1 - \delta$ (over S),

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

which is uniform generalization bound.
or, equivalently,

$$\mathbb{P}_{S \backsim D^n}(\exists f \in \mathcal{F} : (\mathcal{R}(f, D) \geq \hat{\mathcal{R}_n}(f, S) + \mathcal{E}(\delta, n, \mathcal{C}(\mathcal{F})) \leq \delta$$

2 Today

— The case of countable and finite $\mathcal{F}$.

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

— The case where we don't have $|\mathcal{F}| < +\infty$, and where we're going to use the Vapnik–Chervonenkis dimension (VC dimension or VC dim)

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

Let $f \in \mathcal{F}$

$$\hat{\mathcal{R}}_n(f, S) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{f(x_i) \neq y_i}$$

Note : we can use other loss functions

$$\mathcal{R}(f, D) := \mathbb{E}(\mathbf{1}_{f(x_1) \neq y_1}) = \mathbb{P}(f(x_1) \neq y_1)$$

$$\Rightarrow \mathcal{R}_D(f) = \mathbb{E}_S \hat{\mathcal{R}}_n(f, S)$$

which is linearity of $\mathbb{E}$ and IID-ness of S
- According to Hoeffding inequality, $\forall \varepsilon > 0$

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

remember that
- $\mathbf{1}_{f(x_i) \neq y_i}$are IID
- $\mu = \mathbb{E}\mathbf{1}_{f(x_i) \neq y_i}$
- $\frac{1}{n} \sum \mathbf{1}_{f(x_i) \neq y_i}$= sample average

or, using the one-sided inequality :

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

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

$$\mathcal{R}(f, D) \leq \hat{\mathcal{R}}_n(f, S) + \sqrt{\frac{1}{2n} log \frac{1}{\delta}} \text{................(2.1.1)}$$

* **Proof :** Impose $exp(-2n\mathcal{E}^2) \leq \delta$
$exp(-2n\mathcal{E}^2) = \delta$
$\Leftrightarrow -2n\mathcal{E}^2 = log\delta$
$\Leftrightarrow 2n\mathcal{E}^2 = log\frac{1}{\delta}$
$\Leftrightarrow \mathcal{E}' = \sqrt{\frac{1}{2n} log\frac{1}{\delta}}$

Remarks on (2.1.1) :
- The rate of the bound is $O(\frac{1}{\sqrt{n}})$
- It's not a uniform generalization bound because ,
  "$\forall f \in \mathcal{F}$" and "prob$1 - \delta$" are inverted
- To get a uniform generalization bound, we would rather look at achieving result like :

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

! **Remember :**
$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

So,

$$
\begin{aligned}
&= \mathbb{P}(\{\mathcal{R}(f_1, D) - \hat{\mathcal{R}_n}(f_1, S) \geq \mathcal{E}\} \\
&\quad or\{\mathcal{R}(f_2, D) - \hat{\mathcal{R}_n}(f_2, S) \geq \mathcal{E}\} \\
&\quad or \ldots \\
&\quad or\{\mathcal{R}(f_{|\mathcal{F}|}, D) - \hat{\mathcal{R}_n}(f_{|\mathcal{F}|}, S) \geq \mathcal{E}\}) \\
&\leq \sum_{p=1}^{|\mathcal{F}|} \mathbb{P}(\mathcal{R}(f_p, D) - \hat{\mathcal{R}_n}(f_p, S) \geq \mathcal{E}) \quad (Union\, bound) \\
&\leq \sum_{p=1}^{|\mathcal{F}|} exp(-2n\mathcal{E}^2) \quad (Hoeffding\; inequality) \\
&= |\mathcal{F}| exp(-2n\mathcal{E}^2)
\end{aligned}
$$

As before, we're solving :

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

Given this $\mathcal{E}$, we have,

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

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

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

! **Remarks :**

1. We used the "union bound".
2. We used the fact that $|\mathcal{F}| < +\infty$
3. $\mathcal{C}(\mathcal{F}) = |\mathcal{F}|$, $\mathcal{C}(\mathcal{F})$ is the complexity/capacity = The number of functions we have.
4. "In practice", it is very rare to be in the case where $|\mathcal{F}| < +\infty$
5. VC dimension helps us to cope with the situation where $|\mathcal{F}| < +\infty$ does not hold.

2.2 Vapnik–Chervonenkis dimension/VC dimension :

VC (Vapnik-Chervonenkis) dimension is a concept in machine learning and statistical learning theory that measures the capacity or expressiveness of a hypothesis set (a set of functions or classifiers) in its ability to shatter a set of data points.
High level idea :

$$\mathcal{F} \subseteq \{X \rightarrow \{-1, +1\}\}$$

e.g.

$$\mathcal{F} = \{x \mapsto sign(w \bullet x), w \in \mathbb{R}^d\}$$

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

$$|\mathcal{F}_{\mathcal{S}} := \{(f(x_1), ...f(x_n)), f \in \mathcal{F}\}| \leq 2^n$$

The VC dimension is important because it helps us understand the trade-off between the complexity of a hypothesis set and its ability to fit arbitrary data.
(!) In VC dimension we're going to look at the following situation :

$$sup_{\mathcal{S} \backsim |\mathcal{S}|=n} |\mathcal{F}_{\mathcal{S}}| < 2^n$$

Definition 1. *Restriction of $\mathcal{F}$ to a sample :*
$\mathcal{F} \subseteq \{-1, +1\}^X (\equiv \{X \mapsto \{-1, +1\}\})$
$\mathcal{S} = \{x_1, ..., x_n\}, x_i \in X \; \forall i$

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

Note : Sometimes in the literature we can see a "functional" way of writing things

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

Definition 2. *Shattered set :*
Let $\mathcal{S}\{x_1, ...x_n\}$. We say that $\mathcal{S}$ is shattered by $\mathcal{F}$ if $|\mathcal{F}_{\mathcal{S}}| = 2^n$. ***In other words we can realize all the labellings on S given F.***

Definition 3. *VC dim/Vapnik–Chervonenkis dimension :*
The VC dimension of $\mathcal{F}$ is the size of the largest set that is shattered by $\mathcal{F}$
It may happen that, VC dim $(\mathcal{F}) = +\infty$

! Notes :

- 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 dividable problems.

2.2.1 Examples of VC dimension for some classes of functions

- The VC dimension of axis-aligned rectangles is 4. Everything that is inside the rec-

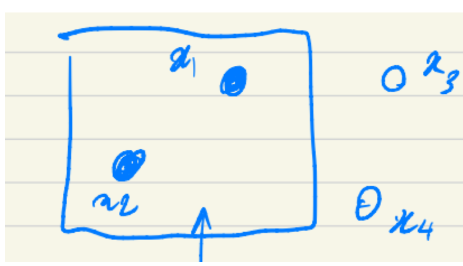
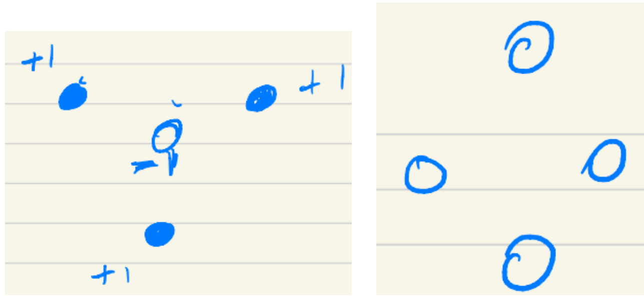


Figure 1 – Classification example with a rectangle ib $\mathcal{F}$



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

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

FIGURE 2 – Example configuration of set, S with $|S| = 4$

tangle above (Fig.1) is classified as a positive instance by the rectangle.

It's fine that for this conjugation of points (Fig.2.a) we can not realize <u>all labellings</u> here BUT there exists a conjugation of 4 points (Fig.2.b) such that all labellings are possible.
This conjugation is shattered by the class of rectangles.

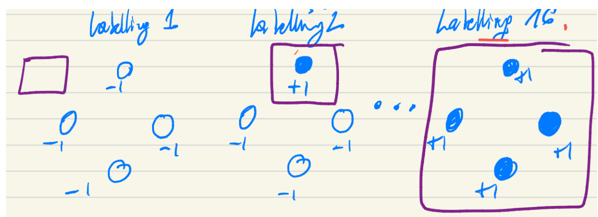


FIGURE 3 – All possible labelling of S from Fig.2.b

But if we take 5 points that the axis aligned rectangle delimited by the max and min X values and the max-min y values has a conjugation that can't be realized.

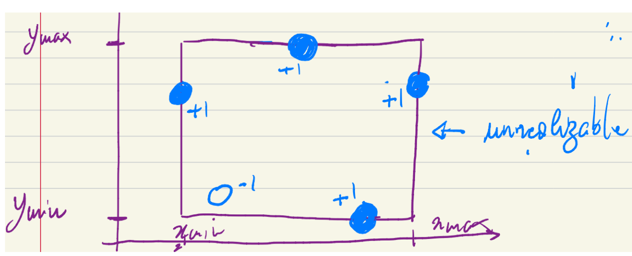
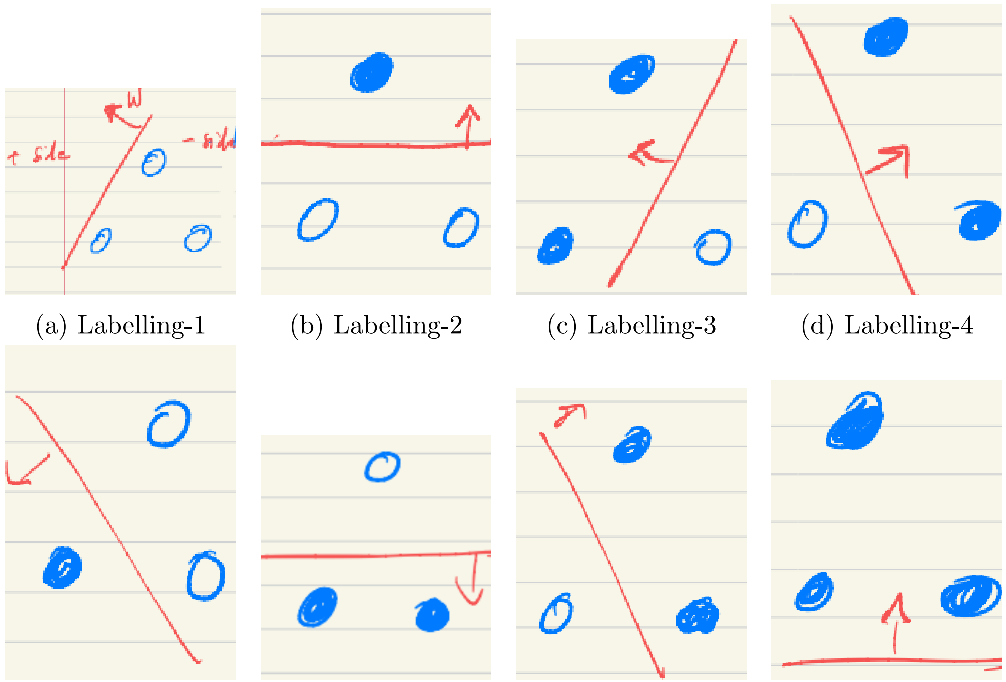


FIGURE 4 – How S can not be shattered example

— The VC dimension of hyperplanes in dimension$d$ is $d+1$. e.g. The VC dim of $d = 2$ is $VCdim(\mathcal{F}) = 3$ (Fig.5)
  If we're looking at 4 points :
  The XOR situation can't be handled by the hyperplanes.

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

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



(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 with $|S| = 3$

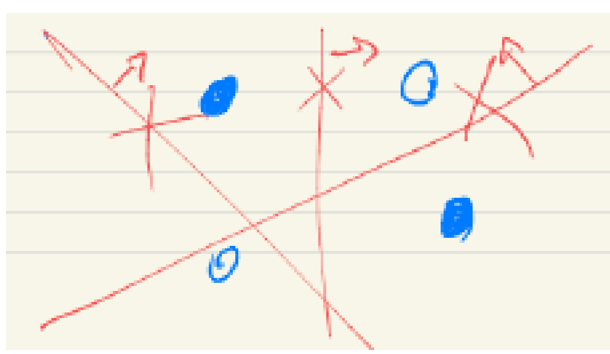


FIGURE 6 – S not being shattered with 4 points

! Remark : If $VCdim(\mathcal{F}) = d$, then

$$\forall n \leq d, \Pi_{\mathcal{F}}(n) = 2^n$$

2.3 VC dimension and generalization error bound :

Theorem 1. *Let* $\mathcal{F} \subseteq \{-1, +1\}_{\mathcal{F}}$ *with* $d := VC - dim(\mathcal{F}) < +\infty$, *with probability* $1 - \delta$,

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

when $\ln e = 1$

! Reminder : With the Rademacher complexity :

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

Proof of the theorem :
— Massart's lemma.
— Bound on the growth function using the Rademacher complexity.
— Sauer's lemma.

Lemma 1. *Massart's lemma :*
Let $A \subseteq \mathbb{R}^n$ *and* $\mathcal{E}_1, ..., \mathcal{E}_n$ *independent.*
Rademacher variables ($\mathbb{P}(\mathcal{E}_i = +1) = \mathbb{P}(\mathcal{E}_i = -1) = \frac{1}{2}$)
Let $\gamma := \sup\limits_{a \in A} \| a \|_2$ *then,*

$$\mathbb{E}_{\mathcal{E}_1,...,\mathcal{E}_n}[\sup_{a \in A} \frac{1}{n} \sum \mathcal{E}_i a_i] \leq \gamma \frac{\sqrt{2\ln|A|}}{n}$$

$\mathcal{E}_i a_i$ *is the i-th component of a.*

Proof :

$$\begin{aligned}
exp(\lambda \mathbb{E}_{\mathcal{E}}[\sup_{a \in A} \sum_{i=1}^n \mathcal{E}_i a_i]) &\leq \mathbb{E}_{\mathcal{E}}[exp(\lambda \sup_{a \in A} \sum_{i=1}^n \mathcal{E}_i a_i)] \\
&\quad \text{(Convexity of exp and property of Jensen inequality)} \\
&= \mathbb{E}_{\mathcal{E}}[\sup_{a \in A} exp(\lambda \sum_{i=1}^n \mathcal{E}_i a_i)] \quad \text{(exp is increasing)} \\
&\leq \mathbb{E}_{\mathcal{E}}[\sum_{a \in A} exp(\lambda \sum_{i=1}^n \mathcal{E}_i a_i)] \\
&= \sum_{a \in A} \mathbb{E}_{\mathcal{E}} exp(\lambda \sum_{i=1}^n \mathcal{E}_i a_i) \\
&= \sum_{a \in A} \mathbb{E}_{\mathcal{E}}[\prod_{i=1}^n exp(\lambda \mathcal{E}_i a_i)] \\
&= \sum_{a \in A} \prod_{i=1}^n \mathbb{E}_{\mathcal{E}_i} exp(\lambda \mathcal{E}_i a_i)
\end{aligned}$$

$$= \sum_{a \in A} \prod_{i=1}^{n} [\frac{1}{2} exp(-\lambda a_i) + \frac{1}{2} exp(\lambda a_i)]$$

$$= \sum_{a \in A} \prod_{i=1}^{n} [\frac{1}{2} exp(-\lambda a_i) + \frac{1}{2} exp(\lambda a_i)]$$

$$\leq \sum_{a \in A} \prod_{i=1}^{n} exp(\frac{\lambda^2 a_i^2}{2}) \quad [as \frac{e^x + e^- x}{2} \leq e^{\frac{x^2}{2}}]$$

$$= \sum_{a \in A} exp(\frac{\lambda^2}{2} \sum_{i=1}^{n} a_i^2)$$

$$= \sum_{a \in A} exp(\frac{\lambda^2}{2} \parallel a \parallel^2$$

$$\leq \sum_{a \in A} exp(\frac{\lambda^2}{2} \gamma^2)$$

$$= |A| exp(\frac{\lambda^2}{2} \gamma^2)$$

We thus have,

$$exp(\lambda \mathbb{E}_{\mathcal{E}} [\sup_{a \in A} \sum_{i=1}^{n} \mathcal{E}_i a_i]) \leq |A| exp(\frac{\lambda^2}{2} \gamma^2)$$

$$\Rightarrow \mathbb{E}_{\mathcal{E}} [\sup_{a \in A} \sum_{i=1}^{n} \mathcal{E}_i a_i]) \leq \frac{\ln |A|}{\lambda} + \lambda \gamma$$

The right-hand side is minimized when, $\lambda = \sqrt{\frac{2 \ln |A|}{\gamma}}$
which is the result stated in the theorem.

Lemma 2. $\hat{Rad}(\mathcal{F}, \mathcal{S}) \leq \sqrt{\frac{1 \ln \prod_{\mathcal{F}}(n)}{n}}$ *[with* $\hat{Rad}(\mathcal{F}, \mathcal{S}) := \mathbb{E}_{\sigma_1, \ldots \sigma_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum \sigma_i f(x_i)$*]*

Proof : $S = \{x_1, ..., x_n\}$

$$\hat{Rad}(\mathcal{F}, \mathcal{S}) := \mathbb{E}_{\sigma} \sup_{a \in \mathcal{F}_{\mathcal{S}}} \frac{1}{n} \sum \sigma_i a_i$$

Since $\mathcal{F} \subseteq \{-1, +1\}^{\mathcal{X}} : \forall a \in \mathcal{F}_{\mathcal{S}} : \parallel a \parallel = \sqrt{n}, \quad \sqrt{\sum (a-i)^2}$ and $\forall i (a_i)^2 = 1$
We can use Massart's lemma on,

$$\hat{Rad}(\mathcal{F}, \mathcal{S}) := \mathbb{E}_{\sigma} \sup_{a \in \mathcal{F}_{\mathcal{S}}} \frac{1}{n} \sum \sigma_i a_i$$

$$\leq \sqrt{n} \frac{\sqrt{2 \ln(|\mathcal{F}_{\mathcal{S}}|)}}{n} \quad \text{(Massart's lemma)}$$

$$= \sqrt{\frac{2 \ln(|\mathcal{F}_{\mathcal{S}}|)}{n}}$$

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

$$\Pi_{\mathcal{F}}(n) \leq \sum_{i=1}^{d} \binom{n}{i} \leq (\frac{en}{d})^d$$

Here,

$$\binom{n}{i} = C_n^i = \frac{n!}{i!(n-i)!} = i \text{ choose } n$$

2.4 Expansion :

$$\begin{aligned}\hat{Rad}(\mathcal{F},\mathcal{S}) &= \mathbb{E}_{\sigma_1,\ldots\sigma_n} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=n}^{n} \sigma_i f(x_i) \\ &= \mathbb{E}_{\sigma_1,\ldots\sigma_n} \sup_{a\in\mathcal{F}_\mathcal{S}} \frac{1}{n}\sum \sigma_i a_i \qquad \textbf{(A)}\end{aligned}$$

1. $\mathcal{F} \subseteq \{-1,+1\}^{\mathcal{X}}$ is set beforehand. We want to measure the capacity of this given set of functions.
2. **Remember :** $\mathcal{F}_\mathcal{S} := \{(f_{x_1}, ..., f_{x_n})|f \in \mathcal{F}\}$.
   This is a set of binary vectors. Obviously $|\mathcal{F}| \leq 2^n$, there exists at most 2^n binary vectors of size n.
   For instance, we may consider that, $n = 5$, and that,

$$\begin{aligned}\mathcal{F}_\mathcal{S} = \{&(-1,-1,+1,-1,+1),\\ &(-1,+1,-1,+1,+1),\\ &(+1,+1,+1,+1,+1),\\ &(-1,+1,+1,+1,+1),\\ &(+1,+1,-1,+1,+1)\}\end{aligned}$$

   $|\mathcal{F}_\mathcal{S}| = 6$
3. Getting back to **(A)** :

$$\hat{Rad}(\mathcal{F},\mathcal{S}) = \mathbb{E}_{\sigma_1,\ldots\sigma_n} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)$$

   but for any f, the vector $(f(x_1), ..., f(x_n))$ is necessarily in $\mathcal{F}_\mathcal{S}$, by delimeter of $\mathcal{F}_\mathcal{S}$. Therefore, the only vectors to be looked at in the definition of the Rademacher Complexity are exactly those in $\mathcal{F}_\mathcal{S}$, or, if we expand the things a bit,

$$\begin{aligned}\hat{Rad}(\mathcal{F},\mathcal{S}) &= \mathbb{E}_{\sigma_1,..,\sigma_n} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i) \\ &= \mathbb{E}_{\sigma_1,\ldots\sigma_n} \sup_{a\in\{(f(x_1),..,f(x_n))|f\in\mathcal{F}\}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i a_i\end{aligned}$$

$$= \mathbb{E}_{\sigma_1,..,\sigma_n} \sup_{a \in \mathcal{F}_{\mathcal{S}}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i$$