Part 1: Machine learning theory

1. **Learning framework**
2. **Uniform convergence**
3. **Learnability of infinite size hypotheses set**
    1. **No-Free-Lunch theorem**
    2. **Infinite hypothesis class: Exemple**
    3. **Classification: VC dimension**
    4. **Regression: Covering number**
4. **Tradeoff Bias/Variance**

Reminder:

If $\boldsymbol{S}$ is $\boldsymbol{\varepsilon}$- representative $\Longrightarrow$ $\boldsymbol{H}$ is UC learnable $\Longrightarrow$ $\boldsymbol{H}$ is APAC learnable $\Longrightarrow$ $\boldsymbol{H}$ is PAC learnable

Learning PAC (targget f exist): $m_H^{PAC}(\varepsilon,\delta)$ **If** $|H| < \infty$, $m_H^{PAC}(\varepsilon,\delta) = \left\lceil \frac{ln\left(\frac{|H|}{\delta}\right)}{\varepsilon} \right\rceil$

- $\forall \varepsilon, \delta \in [0,1]^2$, *and* $\forall\, \mathcal{D}$ *over* Z, $\exists m_H^{PAC}(\varepsilon,\delta)$ ***such that*** $\forall m > m_H^{PAC}(\varepsilon,\delta)$ ***we have***

$$P_{S \sim \mathcal{D}^m}[S_x, L_{\mathcal{D},f}(h_S) > \varepsilon] \leq \delta$$

Learning APAC: $m_H^{APAC}(\varepsilon,\delta)$**If** $|H| < \infty$,$m_H^{APAC}(\varepsilon,\delta) \approx m_H^{CU}\left(\frac{\varepsilon}{2},\delta\right) \approx \left\lceil \frac{2\ ln\left(\frac{2|H|}{\delta}\right)}{\epsilon^2} \right\rceil$

- $\forall \varepsilon, \delta \in [0,1]^2$ *and* $\forall\, \mathcal{D}$ *over* Z, $\exists m_H(\varepsilon,\delta)$ ***such that*** $\forall m > m_H^{APAC}(\varepsilon,\delta)$ ***we have***

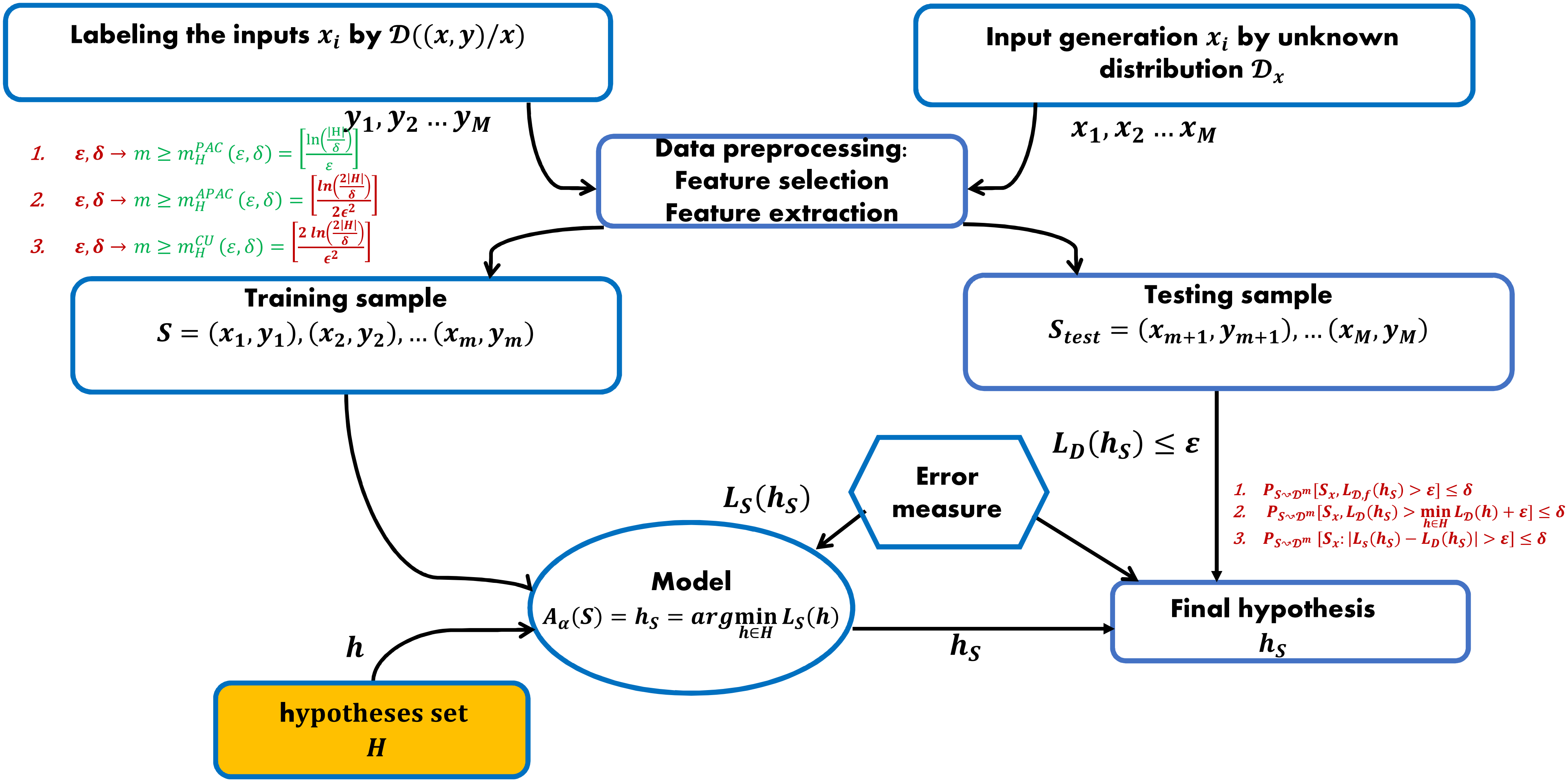
$$P_{S \sim \mathcal{D}^m}[S_x, L_{\mathcal{D}}(h_S) > \min_{h \in H} L_{\mathcal{D}}(h) + \varepsilon] \leq \delta$$

Learning UC: $m_H^{CU}(\varepsilon,\delta)$ **If** $|H| < \infty$, $m_H^{CU}(\varepsilon,\delta) \approx \left\lceil \frac{ln\left(\frac{2|H|}{\delta}\right)}{2\epsilon^2} \right\rceil$

- $\forall \varepsilon, \delta \in [0,1]^2$, *and* $\forall\, \mathcal{D}$ *over* Z, $\exists m_H^{CU}(\varepsilon,\delta)$ ***such that*** $\forall m > m_H^{CU}(\varepsilon,\delta)$ ***we have***

(S **is** ε**–representative**) $\forall h \in H \quad P_{S \sim \mathcal{D}^m}[S_x, |L_S(h) - L_D(h)| > \varepsilon] \leq \delta$

Supervised Learning Passive Offline Algorithm (SLPOA)



Reminder

Definition: Markov Inequality

Let θ be a positive random variable, such that $E[\theta] = \mu$.

So:

$$\forall a > 0 \qquad \mathbf{1} - \boldsymbol{F_{\theta}}(\boldsymbol{a}) = \mathrm{P}(\theta > \mathrm{a}) \leq \frac{\mu}{a}$$

Lemme:

Let θ be a random variable that takes values [0,1] such that $E[\theta] = \mu$.

So:

$$\forall a \in \,]0,1[\qquad \mathrm{P}(\theta > 1 - \mathrm{a}) \geq \frac{\mu - (1-a)}{a}$$

$$\forall a \in \,]0,1[\qquad \mathrm{P}(\theta > \mathrm{a}) \geq \frac{\mu - a}{1-a} \geq \mu - a$$

Proof:

Take $\overline{\boldsymbol{\theta}} = \mathbf{1} - \boldsymbol{\theta}$

Motivation

Objectives:

1- Is there a universal algorithm to solve all types of tasks without having prior knowledge on the task to solve?

The No-Free-Lunch Theorem: Choosing the Right Distribution.

2- The finite size of $\boldsymbol{H}$ is a sufficient condition, but is not necessary for PAC learning (PAC or APC).

when we have $|H| = \infty$

- VC dimension for classification. (Projet multiclass)
- Covering number for regression.

3.1 No-Free-Lunch theorem

Theorem:

Let H be a class of all functions from $X \subseteq \mathbb{R}^n \longrightarrow y = \{0,1\}$ $\mathbf{(}|H| = \infty, \mathrm{h} \in H = \{\, h(x) = a^T x + b, (a,b) \in \mathbb{R}^n \times \mathbb{R}\} \Leftrightarrow H = \{(a,b) \in \mathbb{R}^n \times \mathbb{R}\}),$

- $\forall\ A_\alpha$ and $\forall\ S$ of sample size $|S| \leq \frac{|X|}{2}$
- $\exists\ \boldsymbol{D}$ a distribution on $X \times \{0,1\}$ and $\exists f: X \longrightarrow \{0,1\}$ such that $L_D(f) = 0$. using ERM to find $A_\alpha(S) = h_S$,

Then if we take $\boldsymbol{\varepsilon = \frac{1}{8}, \delta = \frac{1}{7}}$, But:

$$\boldsymbol{P_{S \sim D^m}\ (L_D(h_S) > \frac{1}{8}) \geq \frac{1}{7}}$$

- $\forall D, \forall \varepsilon, \delta > 0\ :\ PAC\colon P_{S \sim D^m}\ (L_D(h_S) > \varepsilon) \leq \delta$

3.1 No-Free-Lunch theorem

Corollary:

Let X be an finite domaine and H the set of all functions from X to $\{0,1\}$. $|H| = \infty$
So $\exists$ $\boldsymbol{D}$ a distribution on $X \times \{0,1\}$ such that H is not PAC learning.

Proof:

Tool: No-Free-Lunch theorem

We will use absurd reasonning.

Therefore, we are going to suppose that H is a class of hypothesis that is PAC learnable.

And, we are going to select a random ε and δ in $[0,1]$, such that:

$$\varepsilon < \frac{1}{8}$$

And:

$$\delta < \frac{1}{7}$$

3.1 No-Free-Lunch theorem

Proof: (continu)

According to PAC definition, there exist an algorithm A and a number $m_H(\varepsilon, \delta)$, such that:

Whatever the distribution that generates the data on $X \times \{0,1\}$ and $\forall f: X \rightarrow \{0,1\}$ such that the realizability assumption is respected.

If we execute the algorithm$A_\alpha$on $m \geq m_H(\varepsilon, \delta)$ sampled $(i.i.d.)$ by D, A will generate a hypothesis such that:$h_S = A_\alpha(S)$

$$L_D(h_S) \leq \varepsilon$$

If we apply the NFL theorem, such that $|X| \geq 2m$

Whatever the algorithm is (in particular A_α), there exist a distribution D such that with a probability $\geq \frac{1}{7}$, we have:

$$L_D(h_S) > \frac{1}{8} > \varepsilon \qquad \text{which is absurd}$$

So, H is not PAC learnable. $No\ PAC$: $P_{S \sim D^m}\left(L_D(h_S) > \frac{1}{8}\right) \geq \frac{1}{7}$

3.1 No-Free-Lunch theorem

Notice:

- The theorem states that whatever the model A_α, there exists a certain distribution D where it fails.
- To avoid this bad distribution, it is necessary to use prior knowledge.
- This prior knowledge implies a restriction on the class of hypotheses H.

How to choose a good class?

$\Longrightarrow$ We should avoid this bad distribution.

$\Longrightarrow$ We should use prior knowledge of H. $H(S)$ with $|S| < \infty$

$\Longrightarrow$ We must apply a restriction on H: instead of working on the whole set X, we will work on another set S $\subset X$.

3.1 No-Free-Lunch theorem

It has been shown from the other chapters that:

1- $|H| < \infty \Longrightarrow H \text{ is } PAC$

2- $\begin{cases} X \text{ is an infinite domain} \\ H = \{h, h: X \to \{0,1\}\} \end{cases} \Longrightarrow H \text{ can be not } PAC$

- What makes a class $\boldsymbol{H}$ PAC and other non PAC?
- Are the infinite classes PAC?
- What determines the complexity of the sample for an infinite class?

$$|\boldsymbol{H}(\boldsymbol{S})| < \infty$$

3.2 Infinite hypothesis class

Example 1:

Let H_s be a set of threshold hypothesis, such that the threshold a belongs to a real set:

$$H_s = \{h_a, a \in \mathbb{R}\}, \qquad |H_s| = \infty$$

Let: $\mathrm{X} = \mathbb{R}$

and

$$h_a: \quad \mathbb{R} \longrightarrow \{0,1\}$$

$$x \longmapsto h_a(x) \quad = \mathbb{1}_{[x<a]} = \begin{cases} 0 \;\; if \;\; x < a \\ 1 \;\; if \;\; x \geq a \end{cases}$$

H_s has a infinite size because $a \in \mathbb{R}$.

Lemma 1:

H_s is PAC by ERM_H, such that the sample complexity is:

$$m_{H_s}(\varepsilon, \delta) \leq \frac{ln(\frac{2}{\delta})}{\varepsilon}$$

3.2 Infinite hypothesis class

Example 2:

Let: $\mathrm{X} = \mathbb{R}$, $H_S = \{h_A = \mathbb{1}_A, A \subseteq \mathbb{R}\} \cup \mathbb{1}_{\mathbb{R}} = \{A : A \subseteq \mathbb{R}\}$

and

$$\begin{aligned} h_A: \quad & \mathbb{R} \longrightarrow \{0,1\} \\ & x \longmapsto h_A(x) = \mathbb{1}_A(x) = \begin{cases} 1 \; if \; x \in A \\ 0 \; if \; x \notin A \end{cases} \end{aligned}$$

Such that A is a finite set.

H_S has a infinite size because $\mathrm{A} \subseteq \mathbb{R}$.

Lemma 2:

H_S is not PAC by ERM_H.

$$A = \{1,2,4,5\}, \qquad h_A(7) = h_A(3) = 0, h_A(5) = h_A(2) = h_A(1) = h_A(4) = 1$$

$$\begin{aligned} h_A: \quad & \mathbb{R} \longrightarrow \{0,1\} \\ & x \longmapsto h_A(x) = \begin{cases} 1 \; if \; x \in A \\ 0 \; otherwise \end{cases} \end{aligned}$$

3.3 Classification: Vapnik-Chervonenkis Dimension

- $\boldsymbol{d_{vc}}$: V-C dimension.
- Growth function (d_{vc}).
- (PAC: Generalisation bound) of infinite H.
- Fundamental theorems of learning.

Shuttering

Definition: shuttering

Let $\boldsymbol{H}$ be a set of functions from X to {0,1} and S⊆X a finite set. $|H| = \infty$

We say that $\boldsymbol{H}$ shutters S if the restriction of H over S is of finite cardinality:

$$|H(S)| = 2^{|S|}$$

Such that:

$$H(S) = \{h(a_1), \dots, h\left(a_{|S|}\right): h \in H\}$$

Example 1:

Let $X = \mathbb{R}$; $H = H_s = \{h_a = \mathbb{1}_{[x<a]}: a \in \mathbb{R}, x \in X\}$ and $S = \{7, 8\}$.

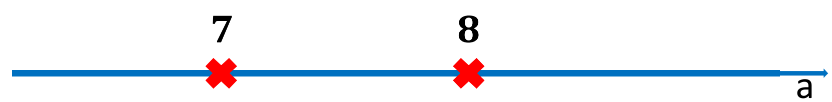
Is S shuttered by H? No

Shuttering

Example 1: answer

We notice that $h_a = \mathbb{1}_{[x<a]}$ has four behaviors of $\{0,1\}$ in S $|S|$=2:

$$H(S) = \{h_a(7), h_a(8) : a \in \mathbb{R}\}$$



1	1
1	0
0	0
0	1

But the hypotheses $h \in H(S)$ do capture only three behaviors.

Then, A is not shuttered by $H(S)$, because we have $|H(S)| = 3 \neq 2^2$.

1. $\forall\, a < 7(< 8)\ h_a(7) = h_a(8) = 0$
2. $\forall\, a < 8(\geq 7)\ h_a(7)$ =1 & $h_a(8)$=0
3. $\forall\, a \geq 8(\geq 7)\ h_a(8) = h_a(7)$ =1

Shuttering

Example 2:

Let $X = \mathbb{R}$ and $H = \{h_a(x) = \mathbb{1}_{[x<a]}: a \in \mathbb{R}, x \in X\}$. If $|S| = 3$ (for example S = {6,7,8}).

Is S shuttered by H?

Example 3:

Let $X = \mathbb{R}^2$ and $H = \{B_{(x,r)}: x \in \mathbb{R}^2 et\ r \in \mathbb{R}^+\}$ such that:

$$B_{(x,r)} = \{y: \|y - x\| \leq r\}$$

If $|S| = 2$.

Is S shuttered by H?

Example 4:

Let $X = \mathbb{R}^2$ and $H = \{B_{(x,r)}: x \in \mathbb{R}^2 et\ r \in \mathbb{R}^+\}$ such that:

$$B_{(x,r)} = \{y: \|y - x\| \leq r\}$$

If $|S| = 3$.

Is S shuttered by H?

Shuttering

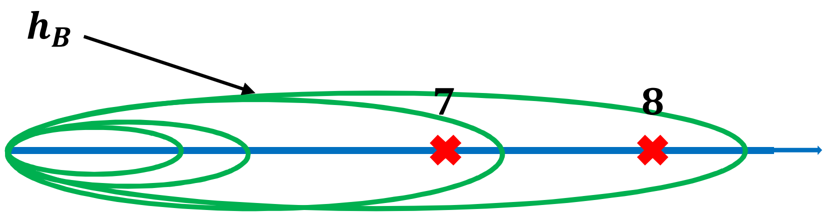
Example 2: answer

Let $X = \mathbb{R}$, $H = \{h_a(x) = \mathbb{1}_{[x<a]}: a \in \mathbb{R}, x \in X\}$ and $A = \{7{,}8\}$.

There exist four subsets in A:

$$\{\emptyset\}, \{7\}, \{8\}, \{7; 8\}$$

Here, the subsets of H have the following form :



By intersection between elements of H and the set A, we can obtain only three subsets of A:

$$\{\emptyset\}, \{7\}, \{7; 8\}$$

Hence, A is not shuttered by H.

Shuttering

Example 3: **answer** $|H(S)| = 2^{|2|} = 4$

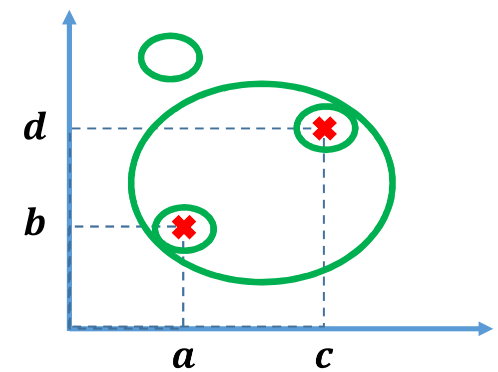
Let $X = \mathbb{R}^2$ and $H = \{B_{(x,r)}: x \in \mathbb{R}^2 and\ r \in \mathbb{R}^+\}$ such that:

$$B_{(x,r)} = \{y: \|y - x\| \leq r\}$$

We have $|S| = 2$.

Let $S = \{(a, b); (\mathrm{c}, \mathrm{d})\}$.

There are four subsets in S:
$\{\emptyset\}, \{(a, b)\}, \{(c, d)\}, \{(a, b); (\mathrm{c}, \mathrm{d})\}$



Here, the subsets of H are cercles.

By intersection betwen the elements of H and the set S,

we can capture all the subsets of S.

So, A is shuttered by H.

Shuttering

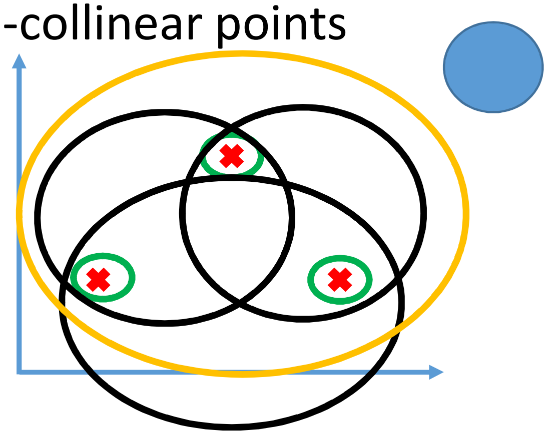
Example 4: **answer** $|H(S)| = 2^{|S|} = 8$

Let $X = \mathbb{R}^2$ and $H = \{B_{(x,r)}: x \in \mathbb{R}^2 and\ r \in \mathbb{R}^+\}$ such that:

$$B_{(x,r)} = \{y: \|y - x\| \leq r\}$$

We have that $|S| = 3$, this implies that S contains 8 subsets.

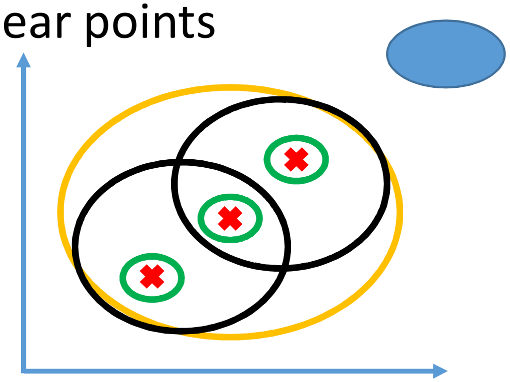
Case 1: non-collinear points



All subsets of S are captured by the elements of H. $|H(S)| = 8 = 2^3$

So, S is shuttered by H.

Case 2: collinear points



Only seven subsets of S are captured by the elements of H. $|H(S)| = 7 \neq 8 = 2^3$

So, S is not shuttered by H.

VC Dimension

Definition: VC Dimension

The VC dimension is a property of H which measures the maximum size of a set $S \subset X$ to be shuttered by H:

$$d_{VC}(H) = \begin{cases} \max\{|S|, S \text{ is shuttered by } H\} \\ +\infty \text{ there is no maximum for } S \end{cases}$$

$S \text{ is shuttered by } H \Leftrightarrow H(S) = 2^{|S|}$

Lemma-L.S.:

PLA: For linear seperators: $d_{VC}(H) = n + 1$ with n is the number of features.

VC Dimension

Examples:

What is the VC dimension of the following sets:

1- $H = H_s = \{h_a : a \in \mathbb{R}\}$, such that:

$$h_a(x) = \mathbb{1}_{[x<a]} \text{ and } X = \mathbb{R} \Longrightarrow d_{VC}(H_s) = 1$$

2- $H = H_S = \{h_A : A \subset \mathbb{R}\}$, such that:

$$h_A(x) = \begin{cases} 1\ si\ x \in A \\ 0\ sinon \end{cases} \text{ and } X = \mathbb{R}$$

VC Dimension

Example 1: answer

Let $X = \mathbb{R}$

And $H = H_s = \{h_a : a \in \mathbb{R}\}$, such that: $h_a(x) = \mathbb{1}_{[x<a]}$

We had proved that $\forall S$ of size ≥ 2, it is not shuttered by H_s.

Finally, we have :

$$d_{VC}(H_s) = 1$$

VC Dimension

Example 2: answer

Let $X = \mathbb{R}$

And $$H = H_S = \{h_A : A \subset \mathbb{R}\}, \text{ such that: } h_A(x) = \begin{cases} 1 \; if \; x \in A \\ 0 \; otherwise \end{cases}$$

We notice that ∀ the size of S, it is always shuttered by H_S, because there is no maximum for S.

Hence:

$$d_{VC}(H_S) = +\infty$$

Conclusion:

We have : $\mathrm{d_{VC}}(H_s) = 1$ and $\mathrm{d_{VC}}(H_S) = +\infty$ and the two sets are infinite.

Therefore, we just proved that the VC dimension is a good measure to make the difference between the infinite sets.

VC Dimension

- $H = H_S = \{h_A : A \subseteq \mathbb{R}\}$, such that: $h_A(x) = \begin{cases} 1 \; if \; x \in A \\ 0 \; if \; x \notin A \end{cases}$

- $S = \{1,2,3,4\}, H_S(S), h_A(x), \quad \forall x \in S, \forall A \subseteq \mathbb{R} \backslash \mathrm{S} \quad h_A(x) = 0$

Tous les h_A possible:

- $h_A : \; A = S$
- $h_A : \; A = \{1\}, A = \{2\}, A = \{3\}, A = \{4\}$
- $h_A : \; A = \{1,2\}, A = \{1,3\} \; A = \{1,4\}, A = \{2,3\}, A = \{2,4\}, A = \{3,4\}$
- $h_A : A = \{1,2,3\}, A = \{1,2,4\}, A = \{2,3,4\}, A = \{1,3,4\}$
- $h_A : A \subseteq \mathbb{R} \backslash \mathrm{S}$

Then $|H_S(S)| = 16 = 2^4$........ If $|S| = n \Longrightarrow |H_S(S)| = 2^n$ then we don'thave a maximum $\Longrightarrow d_{VC}(H_S) = +\infty$

VC Dimension

Corollary: No Free Lunch for $\boldsymbol{B \subset X}$

Let H be a class of all hypotheses from X to $\{0,1\}$. $|\boldsymbol{H}| = \infty$

Let's suppose that there exist $B \subset X$, such that **$\boldsymbol{B}$ is shuttered by $\boldsymbol{H}$, $|\boldsymbol{H}(\boldsymbol{B})| = \mathbf{2}^{|\boldsymbol{B}|} < \infty$ and $|\boldsymbol{B}| = \mathbf{2}\boldsymbol{m}$**.

For any algorithm A_α and for any sample S of size:

$$|S| = \frac{|B|}{2} = m$$

There exist a certain distribution $\boldsymbol{D}$ on $X \times \{0,1\}$ such that:

- $\exists f : X \longrightarrow \{0,1\} : L_D(f) = 0.$
- $\exists \varepsilon = \frac{1}{8}, \exists \delta = \frac{1}{7}, P_{S \sim D^m}\left(L_D\big(A_\alpha(S)\big) > \varepsilon = \frac{1}{8}\right) \geq \delta = \frac{1}{7}$

VC Dimension

Theorem:

Let H be a class of hypotheses , $if\ d_{VC}(H) = +\infty \Longrightarrow H\ is\ not\ PAC.$

Proof: (Theorem)

We have $d_{VC}(H) = +\infty$.

So, for any sample S of size m, there exist a class $A \subset X$ of size $|A| = 2m$ such that A is shuttered by H.

According to the above corollary:

$\forall A_\alpha, \exists D$ on $X \times \{0,1\}$ and $h \in H$ such that $L_D(h) = 0$ but:

$$P_{S \sim D^m}\ (L_D\big(A_\alpha(S)\big) > \frac{1}{8}) \geq \frac{1}{7}$$

Therefore, H is not PAC.

Growth function

Definition:

Let H be a class of hypothesis, the growth function of H is $\Pi_H : \mathbb{N} \longrightarrow \mathbb{N}$, such that:

$$\Pi_H(m) = \max_{\substack{A \subset X \\ |A| = m}} |H_A = H(A)| \qquad H(A) \text{ is the restriction of } H \text{ on } A.$$

Notice:

- $\forall H$ and $\forall m$, $\Pi_H(m) \leq 2^m$
- If H **shutters** the class of size m, $|H_A|= 2^m$ So:
$$\Pi_H(m) = 2^m$$
- If $d_{VC}(H) < m$, So:
$$\Pi_H(m) < 2^m$$

Results

Lemma 4: Sauer

Let H be a class of hypotheses such that:

$$\boldsymbol{d_{VC}(H) \leq d < +\infty}$$

Then:

$$\forall \boldsymbol{m} \;,\quad \boldsymbol{\Pi_H(m) \leq \sum_{i=0}^{d} C_m^i \Longrightarrow log(\Pi_H(m)) \leq log\left(\sum_{i=0}^{d} C_m^i\right)}$$

$$\Longrightarrow \frac{\boldsymbol{4 + \sqrt{\log\left(\Pi_H(2m)\right)}}}{\boldsymbol{\delta\sqrt{2m}}} \leq \frac{\boldsymbol{4 + \sqrt{log\left(\sum_{i=0}^{d} C_m^i\right)}}}{\boldsymbol{\delta\sqrt{2m}}}$$

In particular, if $\boldsymbol{m > d + 1}$, so:

$$\Pi_H(m) \leq \left(\frac{me}{d}\right)^d$$

Generalization bound of infinite H (classification)

Theorem: Generalization bound of VC(C.U)

Let H be a class of hypotheses and Π_H is its growth function. So, for any D and for any $\delta \in [0,1]$:

$$\boldsymbol{P_{S \sim D^m}\left(|L_D(h) - L_S(h)| \leq \varepsilon = \frac{4 + \sqrt{\log\big(\Pi_H(2m)\big)}}{\delta\sqrt{2m}}\right) \geq 1 - \delta}$$

$$P_{S \sim D^m}\left(|L_D(h) - L_S(h)| > \frac{4 + \sqrt{\log\big(\Pi_H(2m)\big)}}{\delta\sqrt{2m}}\right) \leq \delta$$

Such that:

$$\boldsymbol{\varepsilon = \frac{4 + \sqrt{log\big(\Pi_H(2m)\big)}}{\delta\sqrt{2m}}}$$

Fundamental Theorems of Learning

Theorem 1:

Let H be a class of hypotheses in $X \times \{0,1\}$.

Let l be the classification loss function.

We have equivalence between:

1. H follows a uniform convergence.
2. H is APAC learnable by ERM.
3. H is APAC learnable.
4. H is PAC learnable.
5. H is PAC learnable by ERM.
6. $d_{VC}(H)$ is finite.

Notice:

The VC dimension is a tool characterizing the PAC learning.

Fundamental Theorems of Learning

Theorem 2:

Let H be a class of hypotheses in $X \rightarrow \{0,1\}$. Let l a classification loss function.

Let's suppose that $d_{VC}(H) = d < +\infty$. So, there exist two constants C_1 and C_2 such that:

1. H follows a uniform convergence having the sample complexity:

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2} \leq m_H^{CU}(\varepsilon, \delta) \leq C_2 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2}$$

2. H is agnostic PAC learnable having the sample complexity:

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2} \leq m_H^{APAC}(\varepsilon, \delta) \leq C_2 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2}$$

3. H is PAC learnable having the sample complexity:

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon} \leq m_H^{PAC}(\varepsilon, \delta) \leq C_2 \frac{d \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})}{\varepsilon}$$

Notice: The VC dimension allows to determine the sample complexity.

$\boldsymbol{S}$ is a sample of size $\boldsymbol{m}$ $\boldsymbol{A_\alpha(S) = h_S}$,(ch1, Ch2,Ch3)

- **PAC**

$\forall D, \forall(\varepsilon, \delta) \in [0,1]^2, \exists m_H^{PAC}(\varepsilon, \delta), \textit{ such that } \forall m \geq m_H^{PAC}(\varepsilon, \delta)$

$$P_{S \rightsquigarrow D^m}\left(L_D(h_S) > \varepsilon\right) \leq \delta \Leftrightarrow P_{S \rightsquigarrow D^m}\left(L_D(h_S) \leq \varepsilon\right) > 1 - \delta$$

- **APAC**

$\forall D, \forall(\varepsilon, \delta) \in [0,1]^2, \exists m_H^{APAC}(\varepsilon, \delta), \textit{ such that } \forall m \geq m_H^{APAC}(\varepsilon, \delta)$

$$P_{S \rightsquigarrow \mathcal{D}^m}\left[L_{\mathcal{D}}(h_S) > \min_{h \in H} L_{\mathcal{D}}(h) + \varepsilon\right] \leq \delta \Leftrightarrow P_{S \rightsquigarrow D^m}\left(L_D(h_S) \leq \min_{h \in H} L_{\mathcal{D}}(h) + \varepsilon\right) > 1 - \delta$$

- **Uniform Convergence**

$\forall D, \forall \varepsilon, \delta \in [0,1], \exists m_H^{CU}(\varepsilon, \delta), \textit{ such that } \forall m \geq m_H^{CU}(\varepsilon, \delta)$

$$P_{S \rightsquigarrow \mathcal{D}^m}\left[|L_s(h_S) - L_D(h_S)| > \varepsilon\right] \leq \delta \Leftrightarrow P\left[|L_s(h_S) - L_D(h_S)| \leq \varepsilon\right] \geq 1 - \delta$$

- $|H| < \infty$
  - With the Realizabilty hypotheses we have PAC
  - Without we have APAC (tool: Uniform Convergence $\Longrightarrow APAC$)

$$|\boldsymbol{H}| = \infty, \qquad |\boldsymbol{S}| = \boldsymbol{m}$$

1. **Binary Classification**

$$d_{VC}(H) = \begin{cases} \max\{|S|, S \textit{ is shuttered by } H\} \\ +\infty \textit{ there is no maximum for } S \end{cases}$$

- $S \textit{ is shuttered by } H \Leftrightarrow H(S) = 2^{|S|}$
- PLA: For linear seperators: $d_{VC}(H) = n + 1$ with n is the number of features.
- APAC learnable $\Leftrightarrow$ PAC learnable $\Leftrightarrow$ CU learnable $\Leftrightarrow d = d_{VC}(H) < \infty$
- $\begin{cases} \boldsymbol{m \leq d \Longrightarrow \Pi_H(m) \leq \sum_{i=0}^{d} C_m^i} \\ \boldsymbol{m > d + 1} \Longrightarrow \Pi_H(m) \leq \left(\frac{me}{d}\right)^d \end{cases}$
- $\boldsymbol{P_{S \sim D^m}\left(|L_D(h) - L_S(h)| \leq \varepsilon = \frac{4 + \sqrt{\log(\Pi_H(2m))}}{\delta\sqrt{2m}}\right) \geq 1 - \delta}$

3.4 Regression: Covering number

- Background
- **Covering numbers** in a general metric space
- **Covering numbers** in Euclidean space
- **Uniform covering numbers** for a real-valued function class

- $H = \left\{h_{a,b,c}(x) = ax^2 + bx + c\colon (a,b,c) \in \mathbb{R}^3\right\} \Longrightarrow |H| \approx \infty$
- $S = \{(x_i, y_i)\} \Longrightarrow |H(S_x)|\colon S_x = \{x_i\}$
- $h_s(x_i) = y_i \in \mathbb{R}\ and\ x_i \in \mathbb{R}$

Background

Definition: Metric space

(M, d) is called a metric space that consists of a set M together with a metric $d: M \times M \to [0, \infty)$ that satisfies the following for all $x, y, z \in M$:

- $d(x, y) = 0 \quad \Longrightarrow \quad x = y.$
- $d(x, y) = d(y, x).$
- $d(x, z) \leq d(x, y) + d(y, z).$

Definition: Open d-ball

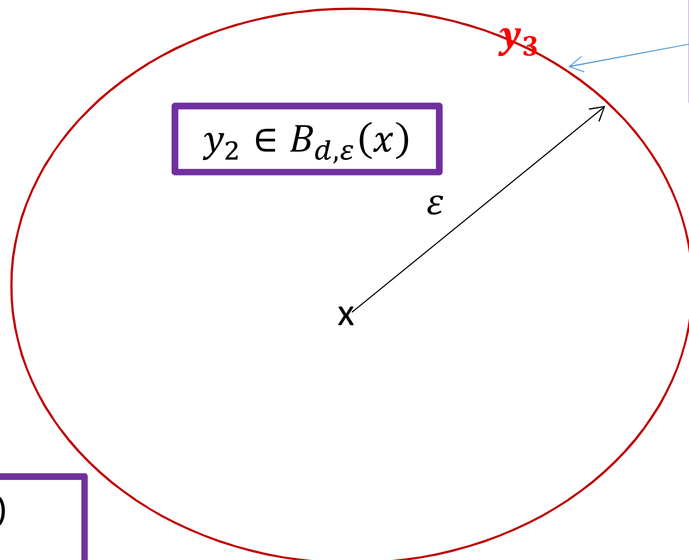
An open d-ball centered at $x \in M$ is defined as:

$$\boldsymbol{B_{d,\varepsilon}(x)} = \{y \in M \mid d(x, y) < \varepsilon\}$$

Open d-ball : Space M

- $d(x, y_1) > \varepsilon$
- $d(x, y_2) < \varepsilon$
- $d(x, y_3) = \varepsilon$
- $\boldsymbol{y_3} \in F_{d,\varepsilon}(x)$

$y_3 \in F_{d,\varepsilon}(x) = \{y \in M \mid d(x,y) = \varepsilon\}$
$F_{d,\varepsilon}(x) \neq B_{d,\varepsilon}(x)$



$y_1 \notin B_{d,\varepsilon}(x) \wedge y_1 \notin F_{d,\varepsilon}(x)$

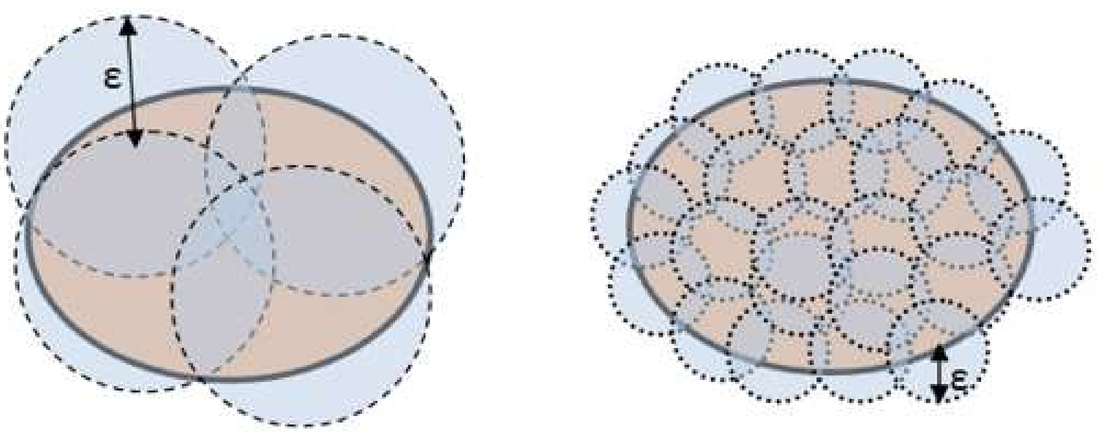
Covering numbers in a general metric space

Definition: ε-cover

Let (M, d) be a metric space. 1: Let $W \subseteq M$ and let $\varepsilon > 0$. A set $C \subseteq W$ is said to be ε-cover of W with respect to d **if $(\forall \boldsymbol{w} \in \boldsymbol{W})(\exists \boldsymbol{c} \in \boldsymbol{C})$ such that:**

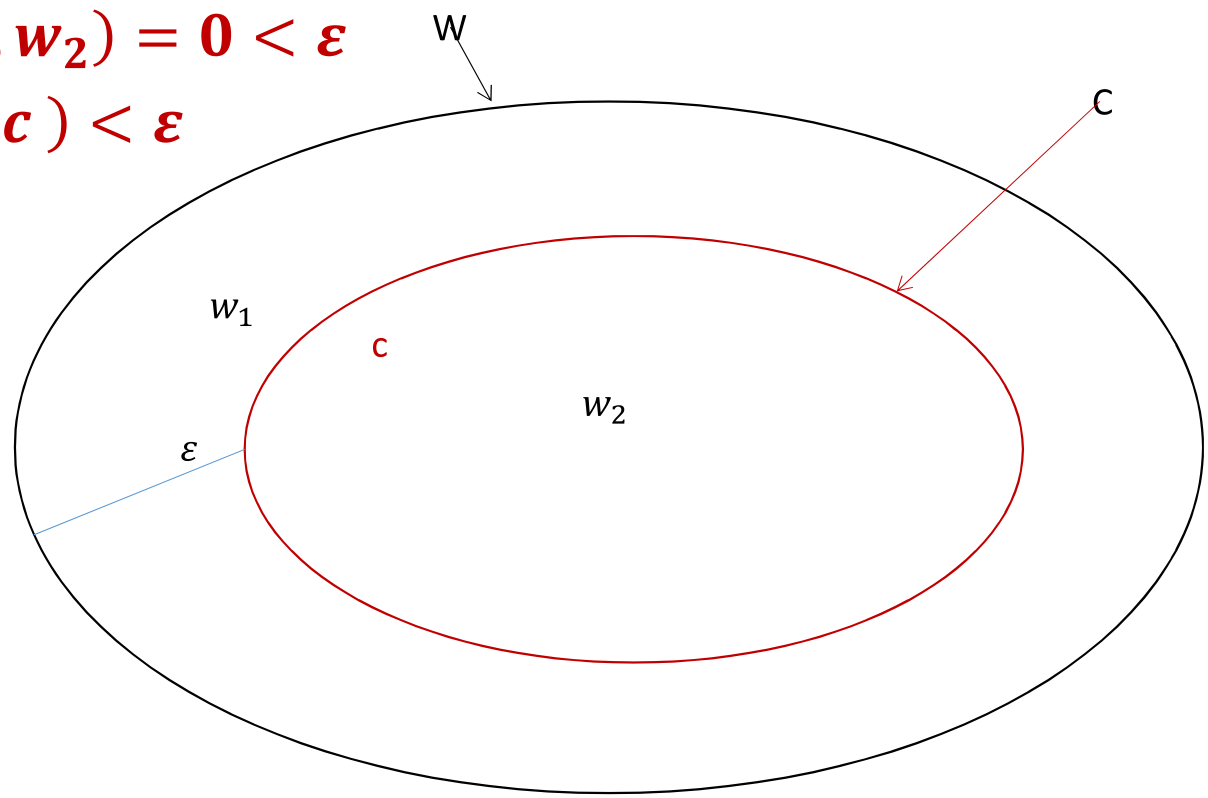
$$\boldsymbol{d(w, c) < \varepsilon}$$

2: In other words, $C \subseteq W$ is an ε-cover of W with respect to d if the union of (open) d-balls of radius $\boldsymbol{\varepsilon}$ centered at points in C contains W: $\bigcup_{c \in C} B_{d,\varepsilon}(c) \supseteq W$



Let $\boldsymbol{W} \subseteq \boldsymbol{M}$ and let $\boldsymbol{\varepsilon} > \mathbf{0}$. A set $\boldsymbol{C} \subseteq \boldsymbol{W}$ is said to be $\boldsymbol{\varepsilon}$-cover of $\boldsymbol{W}$ with respect to $\boldsymbol{d}$ if $(\forall \boldsymbol{w} \in \boldsymbol{W})(\exists \boldsymbol{c} \in \boldsymbol{C})$ such that: $\boldsymbol{d}(\boldsymbol{w}, \boldsymbol{c}) < \boldsymbol{\varepsilon}$

- $\boldsymbol{w_2} \in \boldsymbol{C} \Longrightarrow \boldsymbol{d}(\boldsymbol{w_2}, \boldsymbol{w_2}) = \mathbf{0} < \boldsymbol{\varepsilon}$
- $\boldsymbol{w_1} \notin \boldsymbol{C} \Longrightarrow \boldsymbol{d}(\boldsymbol{w_1}, \boldsymbol{c}) < \boldsymbol{\varepsilon}$

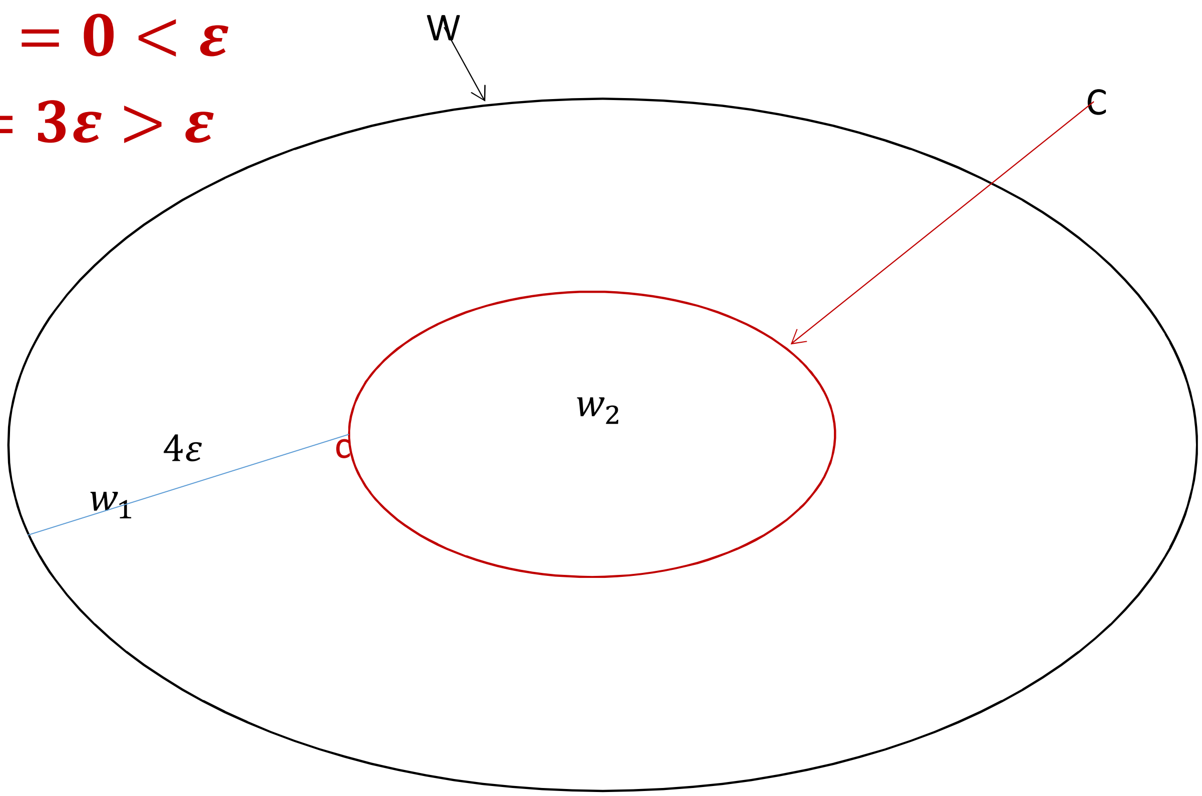


Let $\boldsymbol{W} \subseteq \boldsymbol{M}$ and let $\boldsymbol{\varepsilon} > \mathbf{0}$. A set $\boldsymbol{C} \subseteq \boldsymbol{W}$ is said to be $\boldsymbol{\varepsilon}$-cover of $\boldsymbol{W}$ with respect to $\boldsymbol{d}$ if $(\forall \boldsymbol{w} \in \boldsymbol{W})(\exists \boldsymbol{c} \in \boldsymbol{C})$ such that: $\boldsymbol{d}(\boldsymbol{w}, \boldsymbol{c}) < \boldsymbol{\varepsilon}$

- $\boldsymbol{w_2} \in \boldsymbol{C} \implies \boldsymbol{d}(\boldsymbol{w_2}, \boldsymbol{w_2}) = \mathbf{0} < \boldsymbol{\varepsilon}$
- $\boldsymbol{w_1} \notin \boldsymbol{C} \implies \boldsymbol{d}(\boldsymbol{w_1}, \boldsymbol{c}) = \mathbf{3}\boldsymbol{\varepsilon} > \boldsymbol{\varepsilon}$

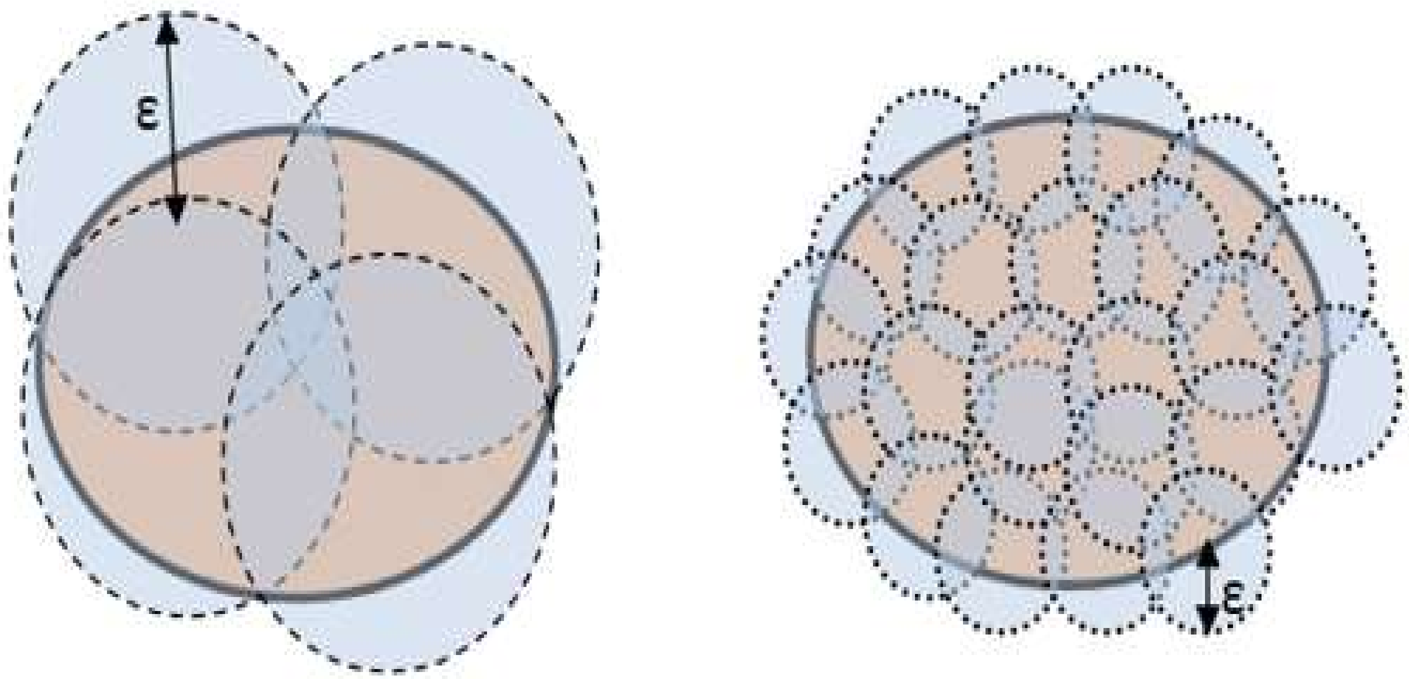
C isn't said to be ε-cover of W

But C is said to be 4ε-cover of W

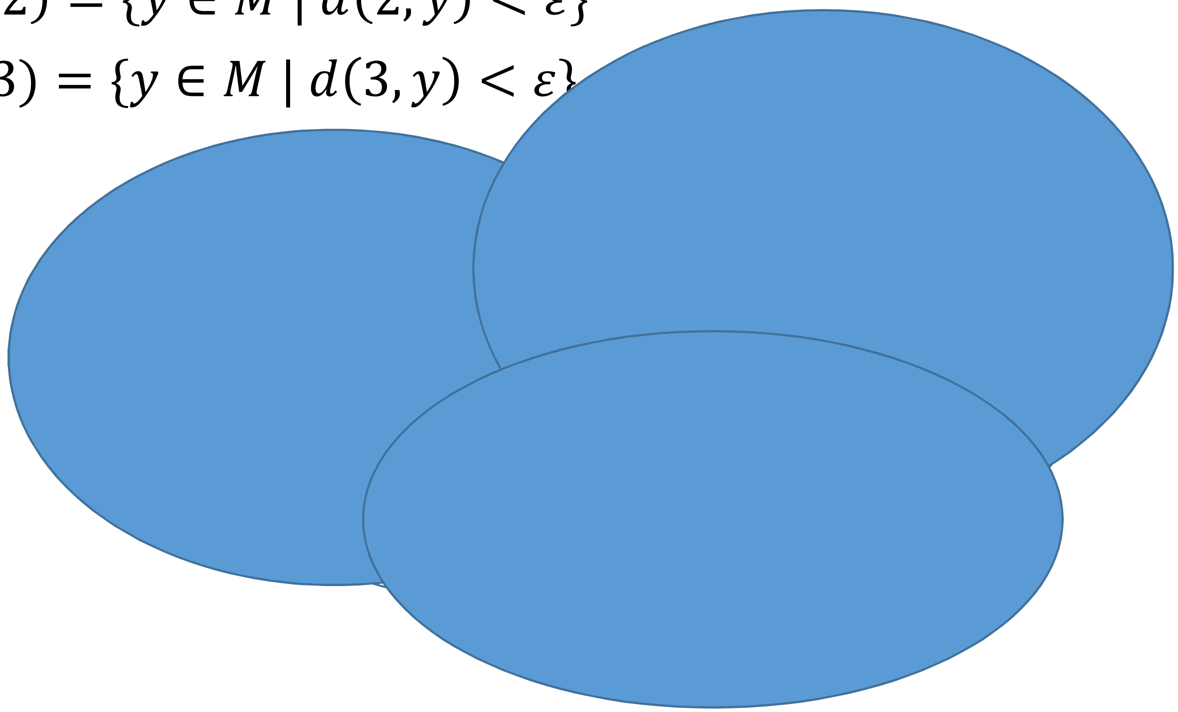


$C \subseteq W$ is an ε-cover of W with respect to d

- if the union of (open) d-balls of radius $\boldsymbol{\varepsilon}$ centered at points in C contains W: $\bigcup_{c \in C} B_{d,\varepsilon}(c) \supseteq W$



- $c = \{1,2,3\}$
- $\Longrightarrow B_{d,\varepsilon}(1) = \{y \in M \mid d(1, y) < \varepsilon\}$
- $\Longrightarrow B_{d,\varepsilon}(2) = \{y \in M \mid d(2, y) < \varepsilon\}$
- $\Longrightarrow B_{d,\varepsilon}(3) = \{y \in M \mid d(3, y) < \varepsilon\}$



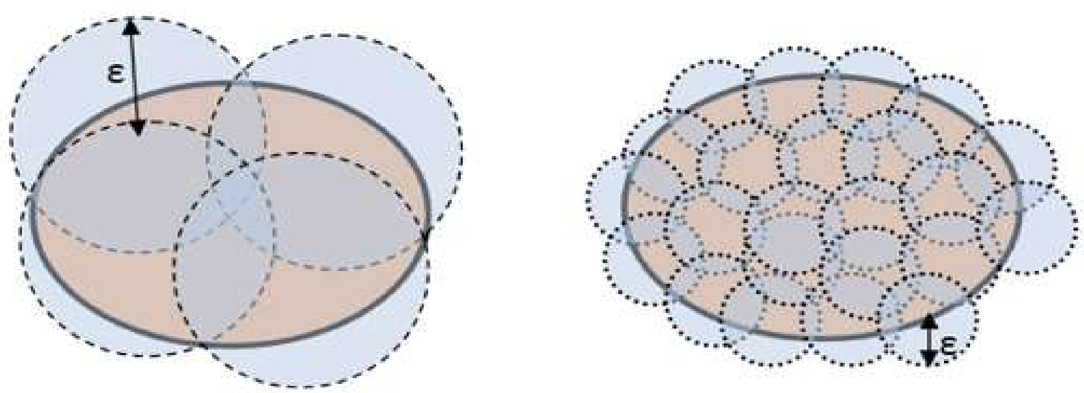
Covering numbers in a general metric space

Definition: ε-covering number

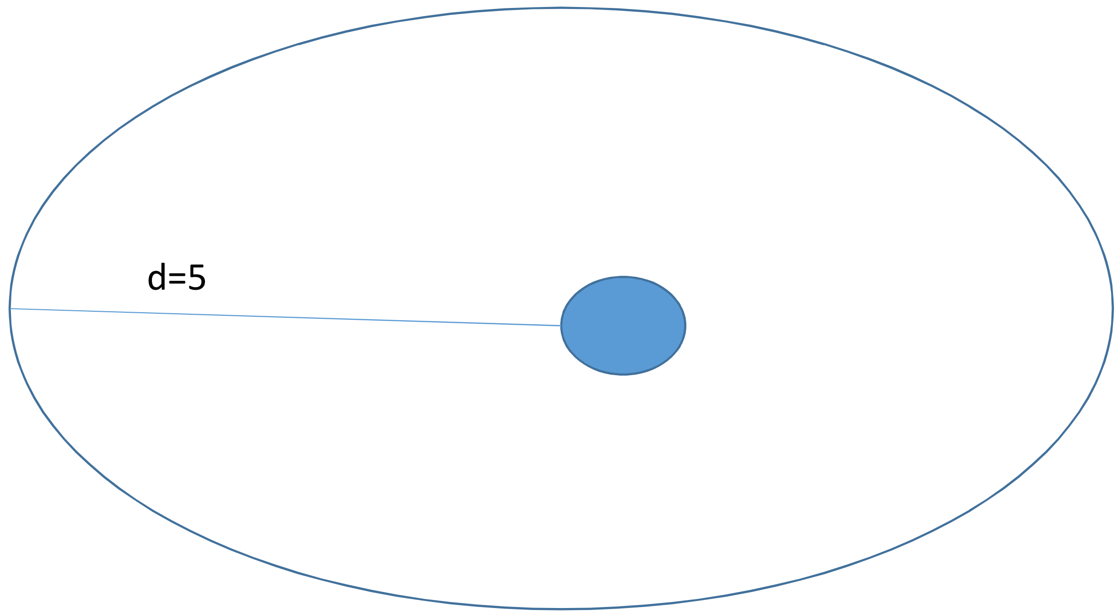
The ε-covering number $\mathcal{N}(\varepsilon, W, d)$ of W with respect to d is defined as the cardinality of the smallest ε-cover of W if W has a finite ε-cover with respect to d. Otherwise, if W does not have a finite ε-cover with respect to d, ε-covering number is equal to infinity.

$$C \subseteq W$$

$$\mathcal{N}(\varepsilon, W, d) = \begin{cases} \min\{|C|, C \; is \; an \; \varepsilon - \text{cover} \; of \; W \; with \; respect \; to \; d\} \\ \infty \qquad if \; W \; does \; not \; have \; a \; finite \; \varepsilon - \text{cover} \end{cases}$$



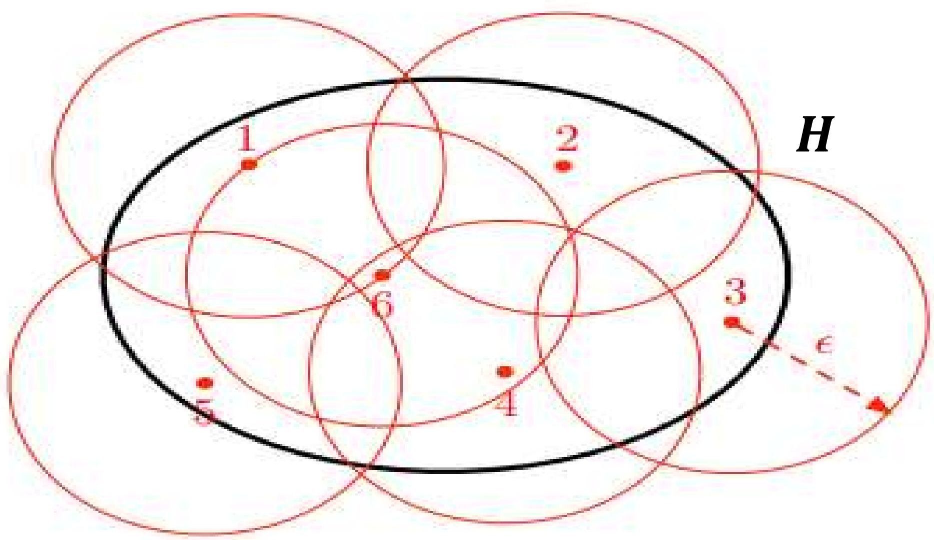
- $\varepsilon \in [0,1]$



Covering numbers in a general metric space

Example:

For instance, for the H shown in the figure the set of points {1, 2, 3, 4, 5, 6} is a covering. However, the covering number is 5 as point 6 can be removed from the set C and the resulting points are still a covering.

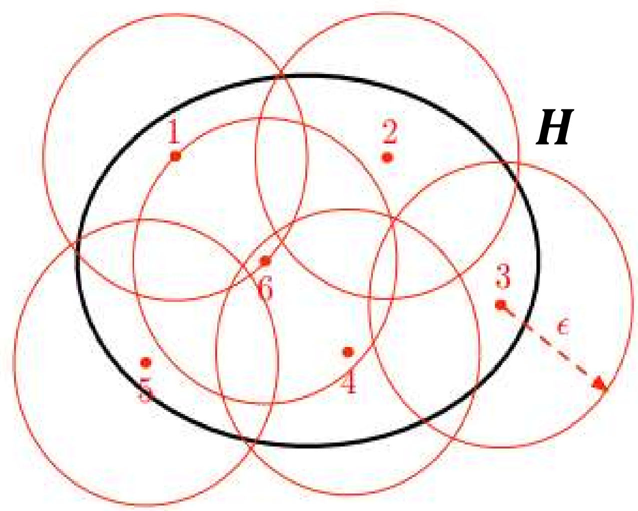


Covering numbers in a general metric space

Example:

$\{1, 2, 3, 4, 5, 6\} \Longrightarrow$

- $B_{d,\varepsilon}(1) = \{y \in M \mid d(1, y) < \varepsilon\}, B_{d,\varepsilon}(2) = \{y \in M \mid d(2, y) < \varepsilon\}$
- $B_{d,\varepsilon}(3) = \{y \in M \mid d(3, y) < \varepsilon\}, B_{d,\varepsilon}(4) = \{y \in M \mid d(4, y) < \varepsilon\}$
- $B_{d,\varepsilon}(5) = \{y \in M \mid d(5, y) < \varepsilon\}$
- $B_{d,\varepsilon}(6) = \{y \in M \mid d(6, y) < \varepsilon\} \subset \bigcup_{x=1,..,5} B_{d,\varepsilon}(x)$
- $\Longrightarrow \bigcup_{x=1,..,5} B_{d,\varepsilon}(x) = \bigcup_{x=1,..,6} B_{d,\varepsilon}(x)$



Covering numbers in Euclidean space

Consider now $M = \mathbb{R}^n$. We can define a number of different metrics on $\mathbb{R}^n$, including in particular the following:

$$d_1(x, x') = \frac{1}{n}\sum_{i=1}^{n} |x_i - x_i'|$$

$$\mathrm{d}_2(\mathrm{x}, \mathrm{x}') = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x}_\mathrm{i} - \mathrm{x}_\mathrm{i}')^2}$$

$$d_\infty(x, x') = \max_i |x_i - x_i'|$$

Covering numbers in Euclidean space

Accordingly, for any $W \subseteq \mathbb{R}^n$, we can define the corresponding covering numbers $\mathcal{N}(\varepsilon, W, d)$ for $p = 1,2,\infty$.

It is easy to see that:

$$\boldsymbol{d_1(x,x') \leq d_2(x,x') \leq d_\infty(x,x') \leq \varepsilon}$$

Therefore, the corresponding covering numbers satisfy the relation:

$$\boldsymbol{\mathcal{N}(\varepsilon, W, d_1) \leq \mathcal{N}(\varepsilon, W, d_2) \leq \mathcal{N}(\varepsilon, W, d_\infty)}$$

$$\boldsymbol{d_1(x,x') \leq \varepsilon' \not\Rightarrow d_2(x,x') \leq \varepsilon'}$$

Uniform covering numbers for a real-valued function class

Definition: uniform covering number

Let H be a class of real-valued functions on X:

$$H = \{h \in H \mid h: X \longrightarrow \mathbb{R}\} \Longrightarrow |H| \approx \infty$$

And let $A = \{x_1, \dots, x_m\} \subset X$. Then the $H_A = H(A) = \{h(x_1), \dots, h(x_m): h \in H\} \subseteq \mathbb{R}$.

For any $\varepsilon > 0$ and $m \in N$, the uniform d_p covering numbers of H for $p = 1,2,\infty$ are defined as:

$$\mathcal{N}_p(\varepsilon, H, m) = \begin{cases} \max\limits_{A \subset X} \mathcal{N}(\varepsilon, H_A, d_p) & if\ \mathcal{N}\left(\varepsilon, H_A, d_p\right)\ is\ finite\ for\ all\ A \subset X\ |A| = m \\ \infty & otherwise \end{cases}$$

Notice: The number of “uniform” refers to the maximum over all $A \subset X$. It has no relationship with uniform convergence.

$$|H| = \infty$$

1. Regression

- **ε-covering number**

$$\mathcal{N}(\varepsilon, W, d) = \begin{cases} \min\{|C|, C \; is \; an \; \varepsilon - \text{cover} \; of \; W \; with \; respect \; to \; d\} \\ \infty \qquad if \; W \; does \; not \; have \; a \; finite \; \varepsilon - \text{cover} \end{cases}$$

- **Uniform Covering Number**

$$\mathcal{N}_p(\varepsilon, H, m) = \begin{cases} \max_{A \subset X} \mathcal{N}(\varepsilon, H_A, d_p) \; if \; \mathcal{N}\left(\varepsilon, H_A, d_p\right) is \; finite \; for \; all \; A \subset X \; |A| = m \\ \infty \qquad otherwise \end{cases}$$

- $H_A = \mathrm{H(A)}$
- APAC learnable $\Leftrightarrow$ PAC learnable $\Leftrightarrow$CU learnable $\Leftrightarrow$ $\mathcal{N}_p(\varepsilon, H, m) < \infty$

Uniform convergence in a Real-valued Function class *H*

Let's assume that H takes values in some set $\hat{Y} \subseteq \mathbb{R}$, so that $H \subseteq \hat{Y}^X$.

We will require the loss function l to be bounded. we will assume $\exists B > 0$ such that:

$(\forall y \in Y)(\forall \hat{y} \in \hat{Y}) \qquad 0 \leq l(y, \hat{y}) \leq B$ and $l: Y \times \hat{Y} \longrightarrow [0, B]$

Definition: The loss function class $\overline{H} = l_H \Longrightarrow |l_H| = \infty\ because |H| = \infty$

We will find it useful to define for any function class $H \subseteq \hat{Y}^X$ and loss $l: Y \times \hat{Y} \longrightarrow [0, B]$ the loss function class $l_H \subseteq [0, B]^{X \times Y}$ given by:

$$\overline{H} = l_H = \{l_h: X \times Y \longrightarrow [0, B] \mid l_h(x, y) = l(y, h(x))\ for\ some\ h \in H\}$$

- $\hat{Y}^X = \{h: X \longrightarrow \hat{Y} \subseteq \mathbb{R}\}, \hat{Y} = \{h(x), x \in X, h \in \hat{Y}^X\}$
- $(x, y),\ d\big(y, h(x)\big) = \ l(y, \hat{y})$
- the loss function $l(: Y \times \hat{Y} \rightarrow \mathbb{R})$ to be bounded :
  - $(\forall y \in Y)(\forall \hat{y} \in \hat{Y}) \qquad 0 \leq l(y, \hat{y}) \leq B$
  - $l: Y \times \hat{Y} \longrightarrow [0, B]$

Uniform convergence in a Real-valued Function class H

Theorem: generalization bound

Let the sets $Y, \hat{Y} \subseteq \mathbb{R}$. Let $H \subseteq \hat{Y}^X$, and let $l: Y \times \hat{Y} \longrightarrow [0, B]$.

Let D be any distribution on $X \times Y$.

For any $\varepsilon > 0$:

$$\underset{S \sim D^m}{\mathrm{P}}\left(\sup_{h \in H} |L_D(h) - L_S(h)| \geq \varepsilon\right) \leq \delta = 4\,\mathcal{N}_1\left(\frac{\varepsilon}{8}, l_H, 2m\right) e^{-m\varepsilon^2/32B^2}$$

Uniform convergence in a Real-valued Function class H

Lemma: L-Lipschitz loss

Let $Y, \hat{Y} \subseteq \mathbb{R}$.

Let $H \subseteq \hat{Y}^X$, and let $l: Y \times \hat{Y} \longrightarrow [0, B]$.

l is Lipschitz in its second argument with Lipschitz constant $L > 0$, if and only if:

$$|l(y, \hat{y}_1) - l(y, \hat{y}_2)| \leq L|\hat{y}_1 - \hat{y}_2| \quad \forall y \in Y, \hat{y}_1, \hat{y}_2 \in \hat{Y} = h(X)$$

Then for any $m \in N$

$$\mathcal{N}_1(\varepsilon, l_F, m) \leq \mathcal{N}_1(\frac{\varepsilon}{L}, l_H, m)$$

l_F **is Lipshitz with L**

Uniform convergence in a Real-valued Function class H

Corollary: generalization bound

Let $Y, \hat{Y} \subseteq \mathbb{R}$.

Let $H \subseteq \hat{Y}^X$, and let $l: Y \times \hat{Y} \longrightarrow [0, B]$ such that l is Lipchitz in its second argument with Lipschitz constant $L > 0$.

Let D be any distribution on $X \times Y$.

For any $\varepsilon > 0$:

$$\underset{S\sim D^m}{\mathrm{P}}\left(\sup_{h\in H}|L_D(h) - L_S(h)| \geq \varepsilon\right) \leq \delta = 4\,\mathcal{N}_1\left(\frac{\varepsilon}{8L}, l_H, 2m\right) e^{-m\varepsilon^2/32B^2} \leq \mathcal{N}_2 \leq \mathcal{N}_\infty$$