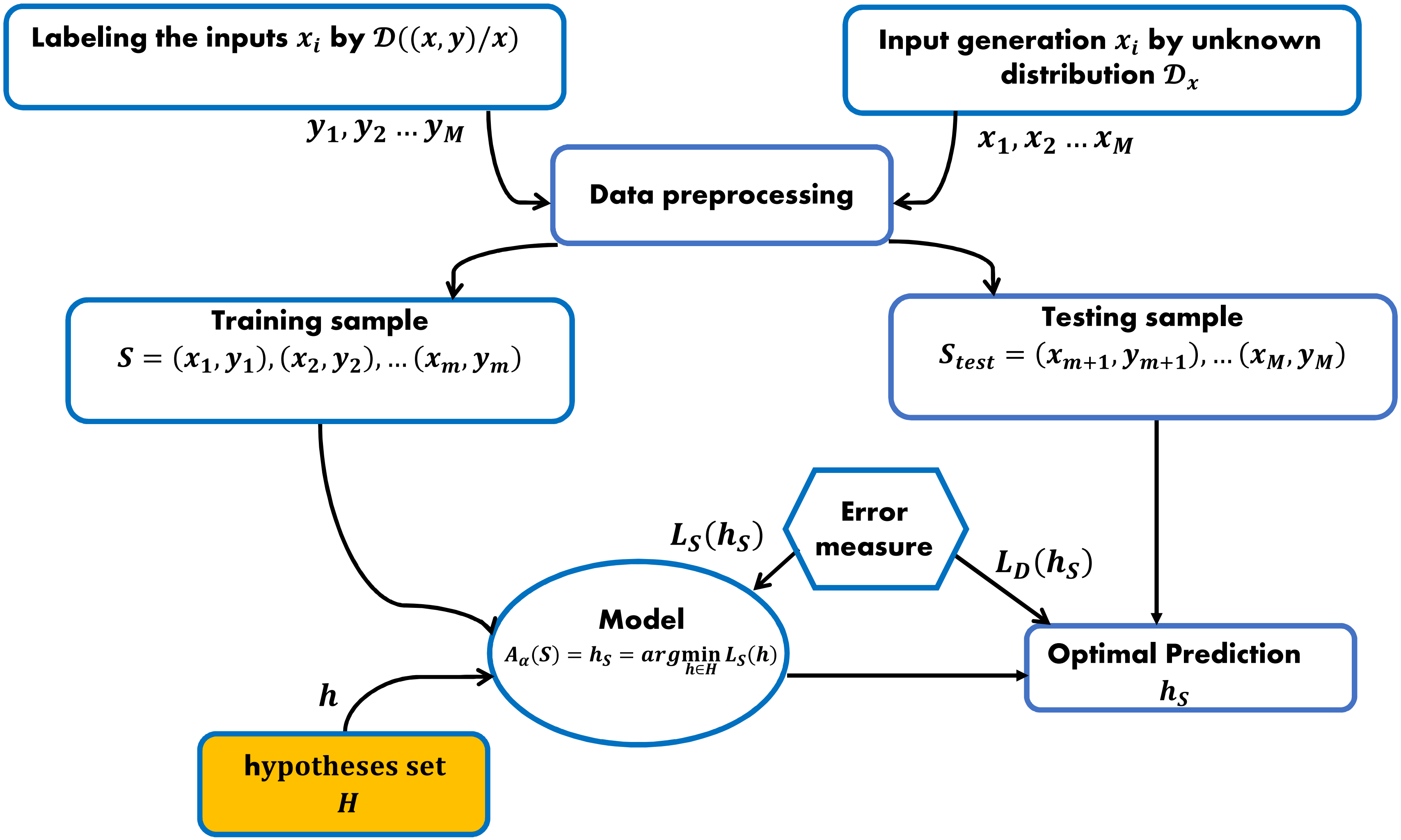
Part 1: Machine learning theory

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

Supervised Learning Passive Offline Algorithm (SLPOA)



Reminder

Definition: ε-representative sample

The sample S is ε-representative with respect to $(Z, H, l, \mathcal{D})$ if :

$$\forall h \in H \qquad |L_s(h) - L_D(h)| \leq \varepsilon$$

Definition: Uniform convergence

We say that H has the uniform convergence property with respect to (Z, l), if there exist:

- a function $m_H^{CU}(\varepsilon, \delta)$: $[0,1]^2 \longrightarrow \mathbb{N}$, such that: $\forall(\varepsilon, \delta) \in [0,1]^2 and\ \forall\ \mathcal{D}$ over Z.
- S is a sample of size $m \geq m_H^{CU}(\varepsilon, \delta)$, whose points are drawn $(i.i.d.)$ by $\mathcal{D}$, such that with probability of at least $(1 - \delta)$, S is ε-representative:

$$P[|L_s(h) - L_D(h)| \leq \varepsilon] \geq 1 - \delta$$

Reminder

Definition: Markov Inequality

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

So:

$\forall a > 0$ P(θ>a) $\leq \frac{\mu}{a}$

Lemme:

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

So:

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

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

Proof:

Take $\boldsymbol{\bar{\theta} = 1 - \theta}$

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.

- VC dimension for classification.
- Covering number for regression.

3.1 No-Free-Lunch theorem

Theorem:

Let H be a class of all functions from $X \longrightarrow \{0,1\} \Longrightarrow |H| \approx \infty$

$\forall \boldsymbol{A_\alpha}$ and $\forall$ S of sample size $|S| \leq \frac{|X|}{2}$

$\exists$ $\boldsymbol{D}$ a distribution on $\boldsymbol{X} \times \{\boldsymbol{0}, \boldsymbol{1}\}$ and **$\exists \boldsymbol{f}: \boldsymbol{X} \longrightarrow \{\boldsymbol{0}, \boldsymbol{1}\}$ such that $\boldsymbol{L_D(f) = 0 \Longrightarrow f} \in H$.**

But:

$$\boldsymbol{P_{S \sim D^m}\left(S\text{: } L_D\big(A_\alpha(S)\big) > \varepsilon = \frac{1}{8}\right) \geq \delta = \frac{1}{7} \Leftrightarrow No\ PAC\ Learning}$$

3.1 No-Free-Lunch theorem

Proof: Let $C \subset X$ such that $|C| = 2m$.

➢**Intuition:**

Let's consider that the algorithm receives the sample S, such that $|S| = m$.

Let H_{2m} be the set of all possible hypotheses in C:

$$H_{2m} = \{f, f : C \to \{0,1\}\} = \{f_1, f_2, \dots, f_T\}$$

We notice that :

$$|H_{2m}| = 2^{2m} = T$$

For each hypothesis f_i such that $i \in \{1, \dots, T\}$, let D_i be the probability distribution on $C \times \{0,1\}$ defined by:

$$\boldsymbol{D_i(\{\mathrm{x}, y\})} = \begin{cases} \dfrac{\mathbf{1}}{|\boldsymbol{C}|} \ \textbf{if } \mathbf{y = f_i(x)} \\ \mathbf{0} \qquad \boldsymbol{otherwise} \end{cases}$$

We know that the training and the testing set follow the same distribution, so:

$$L_{D_i}(f_i) = 0$$

3.1 No-Free-Lunch theorem

➢ **Objective of the proof:**

Step 1: Prove that $\forall A$ and $\forall S \subset \{C \times \{0,1\}\}$, of size m, $A(S): C \rightarrow \{0,1\}$ such that:

$$\max_{i \in \{1,\dots,T\}} \mathop{E}_{S \sim D_i^m} [L_{D_i}(A(S))] \geq \frac{1}{4} \qquad \textbf{Eq.1}$$

Step 1 implies that:

$\forall A$ and $\forall S \subset \{X \times \{0,1\}\}$, of size m, $A(S): X \rightarrow \{0,1\}$, $\exists f: X \rightarrow \{0,1\}$ and a distribution D on $\mathrm{X} \times \{0,1\}$, such that: $L_D(f) = 0$

and

$$\mathop{E}_{S \sim D^m} [L_D(A(S))] \geq \frac{1}{4} \qquad \textbf{Eq.2}$$

Step 2: Prove that $\forall A$ and $\forall S \subset \{X \times \{0,1\}\}$, of size m, $A(S): X \rightarrow \{0,1\}$

$$P\big(L_D\big(A(S)\big) > 1/8\big) \geq \frac{1}{7}$$

3.1 No-Free-Lunch theorem

Step 1:

From C we can extract $k = (2m)^m$ possible samples of size m:

$$S_1, S_2, \dots, S_k$$

Let the sample $S_j = (x_1, \dots, x_m)$ such that S_j^i is a sample of the following form:

$$S_j^i = ((x_1, f_i(x_1)), \dots, (x_m, f_i(x_m)))$$

If the points are sampled by a distribution D_i, so the algorithm A can receive the following training sets:

$$S_1^i, \dots, S_k^i$$

So:

$$\underset{S \sim D_i^m}{\mathrm{E}}\left[L_{D_i}(A(S))\right] = \frac{1}{k}\sum_{j=1}^{k} L_{D_i}\left(A\left(S_j^i\right)\right) \qquad \textbf{Eq.3}$$

3.1 No-Free-Lunch theorem

So:

$$\max_{i\in\{1,\dots,T\}} \mathop{\mathrm{E}}_{S\sim D_i^m}\left[L_{D_i}\big(A(S)\big)\right] = \max_{i\in\{1,\dots,T\}} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}\left(A\left(S_j^i\right)\right) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{D_i}\left(A\left(S_j^i\right)\right)$$

Then:

$$\max_{i\in\{1,\dots,T\}} \mathop{\mathrm{E}}_{S\sim D_i^m}\left[L_{D_i}\big(A(S)\big)\right] \geq \min_{j\in\{1,\dots,k\}} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}\left(A\left(S_j^i\right)\right) \quad \textbf{Eq.4}$$

Now, let's fix $j \in \{1, \dots, k\}$.

Let $S_j = (x_1, \dots, x_m)$ and $(\vartheta_1, \dots, \vartheta_p)$ be the examples of C that do not belong to S_j.

3.1 No-Free-Lunch theorem

So, it is clear that $p \geq m$.

So $\forall h: C \rightarrow \{0,1\}$ and $\forall i \in \{1, \ldots, T\}$, we have that:

$$L_{D_i}(h) = \frac{1}{2m} \sum_{x \in C} 1_{[h(x) \neq f_i(x)]} \geq \frac{1}{2m} \sum_{r=1}^{p} 1_{[h(\vartheta_r) \neq f_i(\vartheta_r)]} \geq \frac{1}{2p} \sum_{r=1}^{p} 1_{[h(\vartheta_r) \neq f_i(\vartheta_r)]}$$

And:

$$h = A\left(S_j^i\right)$$

Then:

$$\frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(A\left(S_j^i\right)\right) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} 1_{\left[A\left(S_j^i\right)(\vartheta_r) \neq f_i(\vartheta_r)\right]} = \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} 1_{\left[A\left(S_j^i\right)(\vartheta_r) \neq f_i(\vartheta_r)\right]}$$

3.1 No-Free-Lunch theorem

So:

$$\frac{1}{T}\sum_{i=1}^{T} L_{D_i}\left(A\left(S_j^i\right)\right) \geq \frac{1}{2}\min_{r\in\{1,\dots,p\}}\frac{1}{T}\sum_{i=1}^{T} 1_{\left[A\left(S_j^i\right)(\vartheta_r)\neq f_i(\vartheta_r)\right]} \quad \textbf{Eq.5}$$

Now, let's fix $\mathrm{r} \in \{1, \dots, p\}$.

We can partition the functions $f_1, f_2, \dots, f_T$ on $\frac{T}{2}$ disjoint pairs.

Such that for one pair $(f_i, f_{i'})$, we have $\forall c \in C$:

$$f_i(c) \neq f_{i'}(\mathrm{c}) \text{ if and only if } (\mathrm{c} = \vartheta_r)$$

This means that :

$$S_j^i = S_j^{i'} \text{ for all } (f_i, f_{i'})$$

3.1 No-Free-Lunch theorem

So:

$$1_{[A(S_j^i)(\vartheta_r) \neq f_i(\vartheta_r)]} + 1_{[A(S_j^{i'})(\vartheta_r) \neq f_{i'}(\vartheta_r)]} = 1$$

Then:

$$\frac{1}{T}\sum_{i=1}^{T} 1_{[A(S_j^i)(\vartheta_r) \neq f_i(\vartheta_r)]} = \frac{1}{2} \quad \textbf{Eq.6}$$

By combining equations **3**, **4**, **5** and **6**, we get:

$$\max_{i\in\{1,\ldots,T\}} \mathop{\mathrm{E}}_{S\sim D_i^m}[L_{D_i}(A(S))] \geq \frac{1}{4}$$

3.1 No-Free-Lunch theorem

Step 2:

Since $L_D(A(S))$ takes values in [0,1], and according to step 1 :

$$\mathop{\mathrm{E}}_{S\sim D^m}[L_D(A(S))] \geq \frac{1}{4} \Longrightarrow \mathop{\mathrm{E}}_{S\sim D^m}\left[L_D\left(A(S)\right)\right] - \frac{1}{8} \geq \frac{1}{4} - \frac{1}{8} = \frac{1}{8}$$

Let's prove that:

$$P(L_D(A(S)) > \frac{1}{8}) \geq \frac{1}{7}$$

By Markov inequality, with $a = \varepsilon$, and:

$$\mu = \mathop{\mathrm{E}}_{S\sim D^m}[L_D(A(S))]$$

We have that:

$$\forall \varepsilon \in [0,1] \quad P\left(L_D\left(A(S)\right) > \frac{1}{8}\right) \geq \frac{\mu - \frac{1}{8}}{1 - \frac{1}{8}} = \frac{\mu - \frac{1}{8}}{\frac{7}{8}} \geq \frac{\frac{1}{8}}{\frac{7}{8}} = \frac{1}{7}$$

3.1 No-Free-Lunch theorem

Corollary:

Let X be an infinite domaine and H the set of all functions from X to $\{0,1\}$.
So H is not PAC learning.

Proof:

We will use absurd reasoning.

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 on $m \geq m_H(\varepsilon, \delta)$ sampled $(i.i.d.)$ by D, A will generate a hypothesis such that:

$$L_D(A(S)) \leq \varepsilon$$

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

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

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

So, H is not PAC learnable.

3.1 No-Free-Lunch theorem

Notice:

- The theorem states that whatever the algorithm A, 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.

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

3.1 No-Free-Lunch theorem

It has been shown from the other chapters that:

1- $|H| < \infty \Longrightarrow H \ is \ PAC$

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

- 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?

- $|S| = m < \infty \Longrightarrow H(S) < \infty$

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:

$$H_s = \{h_a, a \in \mathbb{R}\} \Longrightarrow |H| \approx \infty$$

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

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

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

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:

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

3.2 Infinite hypothesis class

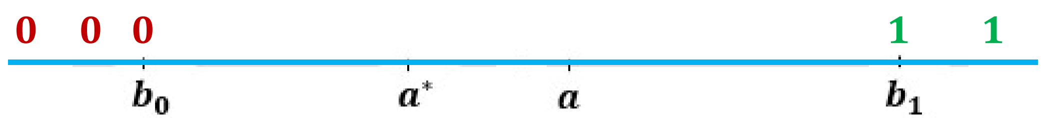
Proof: (Lemma 1)

Consider the algorithm $\boldsymbol{A}$ that reveives a sample $\boldsymbol{S} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_m}\}$ such that:

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

Let's prove that H_S is PAC by ERM.

Consider: $\begin{cases} b_0 = \max\limits_x \{x: (x, 0) \in S\} \\ b_1 = \min\limits_x \{x: (x, 1) \in S\} \end{cases}$



Let the threshold a generated by the hypothesis h_a, so: $a \in [b_0, b_1]$.

Let the threshold a^* generated by the optimal hypothesis h_{a^*} such that:

$$h_{a^*}(x) = \mathbb{1}_{[x \geq a^*]} \quad \text{and} \quad L_D(h_{a^*}) = 0$$

3.2 Infinite hypothesis class

Proof: (Lemma 1)

Background:

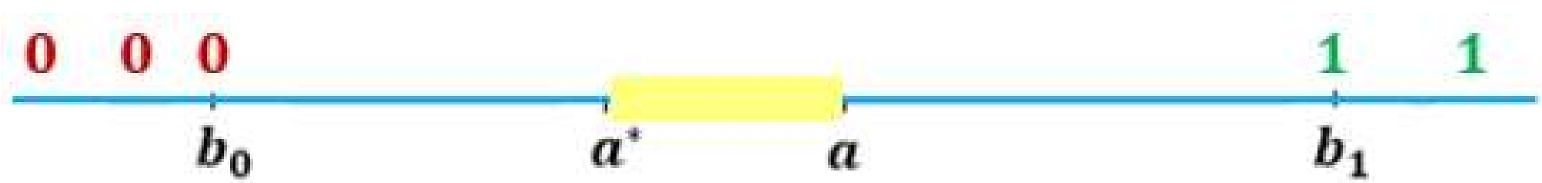
- $1 + x \leq e^x$, $\forall x \geq 0$.
- $\boldsymbol{P(A \cup B) \leq P(A) + P(B)}$ such that A and B are independants.
- If $A \subseteq B$ then $P(A) \leq P(B)$.

Objective of the proof:

For all $m \geq m_{H_s}(\varepsilon, \delta)$: $P_{S \sim D^m}(L_D(h_a) > \varepsilon) \leq \delta$

Let's note that:

- All the points in $[a, a^*]$ will be labeled differently by h_a and h_{a^*}.
- All the point in $[a, +\infty[$ and $]-\infty, a^*]$ will be labeled identically by h_a and h_{a^*}.



3.2 Infinite hypothesis class

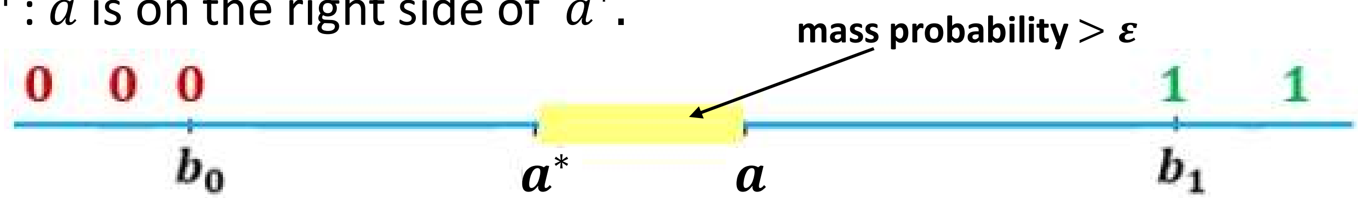
Proof: (Lemma 1)

Then the generalization error $L_{D,f}(h_a)$ is the mass probability between a and a^*.

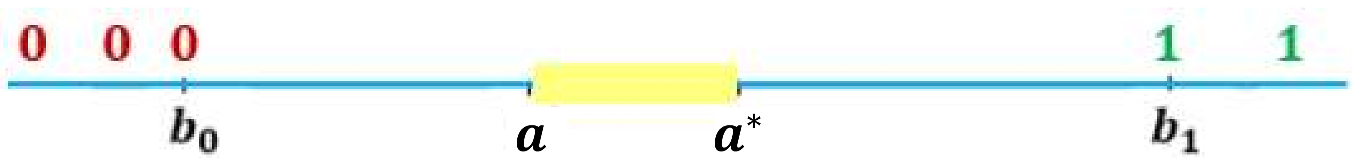
We have two events that realizes the following condition:

$$L_{D,f}(h_a) > \varepsilon$$

The event B^+: a is on the right side of a^*.



The event B^-: a is on the left side of a^*.



Then:

$$L_{D,f}(h_a) > \varepsilon \Longrightarrow B^+ \cup B^-$$

3.2 Infinite hypothesis class

Proof: (Lemma 1)

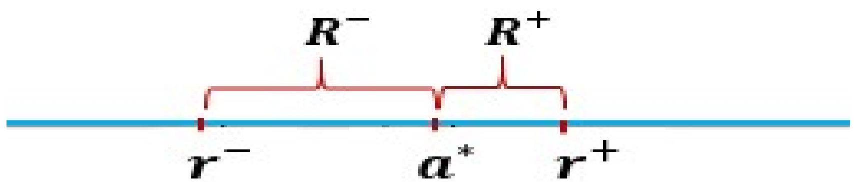
So:

$$P_{S\sim D^m}(L_{D,f}(h_a) > \varepsilon) \leq P_{S\sim D^m}(B^+ \cup B^-) \leq P(B^+) + P(B^-)$$

Let's determine $P(B^+)$ and $P(B^-)$.

Consider:

- r^+ a point in the right side of a^* such that the mass probability of $R^+ = [a^*, r^+]$ is ε.
- r^- a point in the left side of a^* such that the mass probability of $R^- = [r^-, a^*]$ is ε.



- The event B^+ occurs if a is on the right side of r^+ $(a > r^+)$.

Then $b_1 \geq a > r^+$, so $\forall x \in S : \mathrm{x} \notin R^+$

- The event B^- occurs if a is on the left side of r^- $(a < r^-)$.

Then $b_0 \leq a < r^-$, so $\forall x \in S : \mathrm{x} \notin R^-$

3.2 Infinite hypothesis class

Proof: (Lemma 1)

Then:

$$P(B^+) = P\big((x_1 \notin R^+\big) \wedge (x_2 \notin R^+) \wedge \cdots \wedge (x_m \notin R^+)) = (1-\varepsilon)^m \leq e^{-\varepsilon m}$$
$$P(B^-) = P\big((x_1 \notin R^-\big) \wedge (x_2 \notin R^-) \wedge \cdots \wedge (x_m \notin R^-)) = (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

Therefore:

$$P_{S \sim D^m}(L_{D,f}(h_a) > \varepsilon) \leq e^{-\varepsilon m} + e^{-\varepsilon m} = 2e^{-\varepsilon}$$

We have that: $$m_H(\varepsilon, \delta) \leq \frac{ln(\frac{2}{\delta})}{\varepsilon}$$

So: $$m \geq \frac{ln(\frac{2}{\delta})}{\varepsilon}$$

$$2e^{-\varepsilon} \quad \leq \delta$$

Hence:

$$P_{S \sim D^m}(L_D(h_a) > \varepsilon) \leq \delta$$

Finally H_s is PAC.

3.2 Infinite hypothesis class

Example 2:

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

$$H_S = \{h_A, A \subseteq \mathbb{R}\} \cup \mathbb{1}_{\mathbb{R}} \Longrightarrow |\boldsymbol{H_S}| \approx \infty$$

and

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

$$x \longmapsto h_A(x) \quad = \begin{cases} 1 \; if \; x \in A \\ 0 \; otherwise \end{cases}$$

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.

3.2 Infinite hypothesis class

Proof: (Lemma 2)

To prove that H_S is not PAC, we should prove that:

$\forall m_H(\varepsilon,\delta)$, $\exists(m \geq m_H(\varepsilon,\delta))$ such that:

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

Let $\mathbb{P}$ be a uniform distribution on $[0,1]$.

Consider the labeling la fonction $\mathbb{1}$: $(\forall x, \mathbb{1}(x) = 1)$.

And $A = \{x_1, \dots, x_m\}$.

Let's prove that H_S is not PAC by ERM.

Consider a sample of size m, $(S \sim P^m)$, $S = ((x_1, 1), \dots, (x_m, 1))$.

3.2 Infinite hypothesis class

Proof: (Lemma 2)

The algorithm ERM can select a hypothesis h_A such that:

$$L_S(h_A) = 0$$

We have that the probability of that the points $\{x_1, \dots, x_m\}$ figure in the test sample is zero (all the points have the same probability).

Therefore, $\forall x \notin S$, we have:

$$h_A(x) = 0$$

The label function is $\mathbb{1}$.

So:

$$L_{\mathbb{P},f}(h_A) = 1$$

Then h_A fails in all the testing points.

3.2 Infinite hypothesis class

Proof: (Lemma 2)

So, we have :

$$L_S(h_A) = 0 \text{ et } L_{\mathbb{P},f}(h_A) = 1$$

Hence the overfitting problem.

Finally, H_A is not PAC by ERM.

Notice:

According to example 1 and 2, H_s and H_S are two sets of the same size which is infinite. But:

- H_s is PAC according to ERM.
- H_S is not PAC according to ERM.

So, the size of H is not a necessary condition for PAC learning. (sufficient condition).