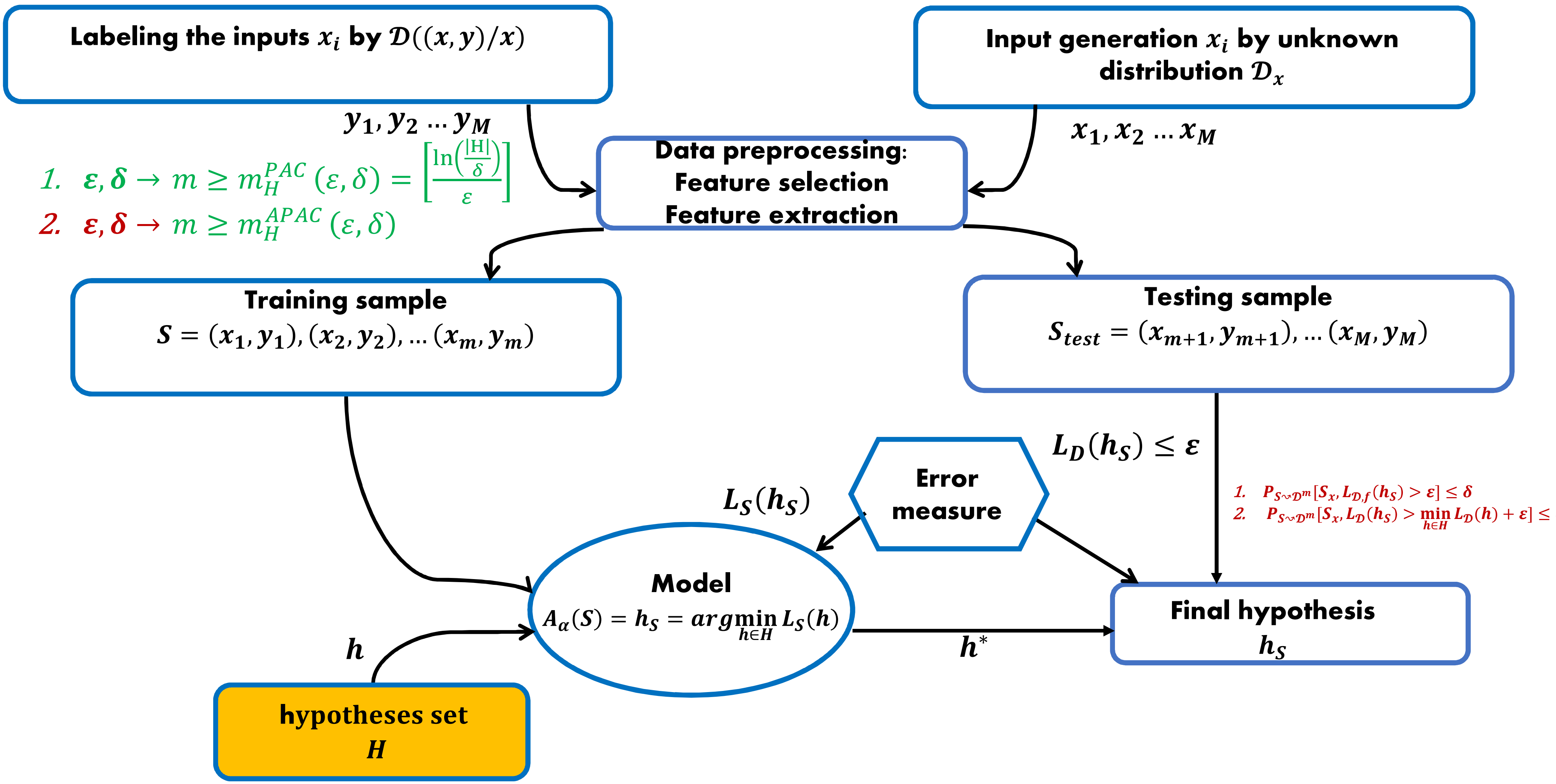
Part 1: Machine learning theory

1. Learning framework
2. Uniform convergence:
    1. ε-representative sample.
    2. Uniform convergence.
3. Learnability of infinite size hypotheses classes
4. Tradeoff Bias/Variance
5. Non-Uniform learning.

Reminder



Labeling the inputs x_i by $\mathcal{D}((x,y)/x)$
Input generation x_i by unknown distribution $\mathcal{D}_x$
$y_1, y_2 \dots y_M$
$x_1, x_2 \dots x_M$
Data preprocessing: Feature selection Feature extraction
1. $\varepsilon, \delta \rightarrow m \geq m_H^{PAC}(\varepsilon, \delta) = \left\lceil \frac{\ln\left(\frac{|H|}{\delta}\right)}{\varepsilon} \right\rceil$
2. $\varepsilon, \delta \rightarrow m \geq m_H^{APAC}(\varepsilon, \delta)$
Training sample
$S = (x_1, y_1), (x_2, y_2), \dots (x_m, y_m)$
Testing sample
$S_{test} = (x_{m+1}, y_{m+1}), \dots (x_M, y_M)$
$L_D(h_S) \leq \varepsilon$
1. $P_{S \sim \mathcal{D}^m}[S_x, L_{\mathcal{D},f}(h_S) > \varepsilon] \leq \delta$
2. $P_{S \sim \mathcal{D}^m}[S_x, L_{\mathcal{D}}(h_S) > \min_{h \in H} L_{\mathcal{D}}(h) + \varepsilon] \leq \delta$
$L_S(h_S)$
Error measure
Model
$A_\alpha(S) = h_S = \underset{h \in H}{argmin}\, L_S(h)$
h
h^*
Final hypothesis
h_S
hypotheses set
H

Reminder

- **PAC** if hypotheses of realizability: targget f exist
- **APAC** if there isn't hypotheses of realizability f with probability
- **If $|H| < \infty$ and hypotheses of realizability holds then we have PAC** $\forall\, m \geq m_H^{PAC}(\varepsilon, \delta) = \left\lceil \frac{\ln\left(\frac{|H|}{\delta}\right)}{\varepsilon} \right\rceil$

Learning PAC : $m_H^{PAC}(\varepsilon, \delta)$

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

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

Learning APAC: $m_H^{APAC}(\varepsilon, \delta)$

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

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

Motivation

Our aim in this chapter is to proove this proposition(**there isn't a realizability hypotheses**):

if H is a finite class of hypotheses, then H follows an Agnostic PAC learning.

Tool:

Uniform convergence.

$|H| < \infty \Longrightarrow$ **Learning Uniforme convegrence** $\Longrightarrow$ **Agnostic PAC learning**

Learning Uniforme convegrence: $m_H^{CU}(\varepsilon, \delta)$***: S is*** **ε-representative**

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

$$\boldsymbol{P[S_x: |L_s(h_S) - L_D(h_S)| > \varepsilon] \leq \delta \Leftrightarrow P[S_x: |L_s(h_S) - L_D(h_S)| \leq \varepsilon] \geq 1 - \delta}$$

- $\boldsymbol{|L_s(h_S) - L_D(h_S)| \approx 0\ and\ \ L_D(h_S)\ \approx 0}$

2.1. ε-representative sample

Definition:

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

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

Notice:

If S is ε-representative, so ERM_H is a good learning strategy.

- $\boldsymbol{|L_s(h) - L_D(h)| \approx 0}$
- $\boldsymbol{L_D(h) \approx 0}$

2.1. ε-representative sample

Lemma:

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

$$\boldsymbol{L_D(h_s) \le \min_{h \in H}\big(L_D(h)\big) + 2\varepsilon}$$

Such that:

$$ERM_H(\mathrm{S}) = \boldsymbol{A_\alpha(S)} = \mathrm{h_S} \in \underset{h \in H}{\operatorname{argmin}}\{L_S(h)\}$$

2.1. ε-representative sample

proof:

Let S be ε-representative, then:

$\forall h \in H$, we have:

$$|L_s(h) - L_D(h)| \leq \varepsilon \Longrightarrow L_S(h) \leq L_D(h) + \varepsilon$$

We know that h_S is the output of $ERM_H(S)$, so:

$$\mathrm{h_S} \in \underset{h \in H}{\operatorname{argmin}}\{L_S(h)\}$$

So, $\forall h \in H$, we have:

$$L_S(h_S) \leq L_S(h)$$

Since S is ε-representative, then for $\mathrm{h} = h_S$, We also have:

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

2.1. ε-representative sample

proof:

$$L_D(h_S) \leq L_S(h_S) + \varepsilon \leq L_S(h) + \varepsilon \leq L_D(h) + 2\varepsilon$$

So, $\forall h \in H$:

$$L_D(h_S) \leq L_D(h) + 2\varepsilon$$

Then :

$$L_D(h_s) \leq \min_{h \in H}\big(L_D(h)\big) + 2\varepsilon$$

2.2. Uniform convergence

Definition:

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

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

$$\boldsymbol{eq\ 1}: \quad \boldsymbol{P[S_x: |L_s(h_S) - L_D(h_S)| \leq \varepsilon] \geq 1 - \delta \Leftrightarrow P[S_x: |L_s(h_S) - L_D(h_S)| > \varepsilon] \leq \delta}$$

$\boldsymbol{P}[\boldsymbol{S_x}: |\boldsymbol{L_s}(\mathrm{h_S}) - \boldsymbol{L_D}(\mathrm{h_S})| > \boldsymbol{\varepsilon}] \leq \boldsymbol{\delta}$ bad hypothesis(there isn't a ε-representative)

Such that $ERM_H(\mathrm{S}) = \boldsymbol{A_\alpha(S)} = \mathrm{h_S} \in \underset{h \in H}{\operatorname{argmin}}\{L_S(h)\}$

Notice:

If H has the uniform convergence property $\Longrightarrow$ H is called « **Glivenko-Cantelli class** ».

H: hypotheses set

PAC: $m_H^{PAC}(\varepsilon,\delta)$, ε_{PAC}

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

APAC $m_H^{APAC}(\varepsilon,\delta)$

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

CU: $m_H^{CU}(\varepsilon_{CU},\delta)$

$$P_{S\sim\mathcal{D}^m}[|L_s(h) - L_D(h)| > \varepsilon] \le \delta$$

$S\ is$ **ε-representative**

2.2. Uniform convergence

Definition: **Sample complexity** $\boldsymbol{m_H^{CU}(\varepsilon,\delta) \Rightarrow m_H^{APAC}(2\varepsilon,\delta) \Longrightarrow m_H^{PAC}(?\,,\delta)}$

The function $m_H^{CU}(\varepsilon,\delta) : [0,1]^2 \longrightarrow \mathbb{N}$ enables to determine the minimal number of data for which H follows a uniform convergence with accuracy ε and confidance δ.

Theorem1: (Lemma 1 and definition Of CU)

If H follows a uniform convergence with complexity $m_H^{CU}(\varepsilon,\delta)$, then H follows Agnostic PAC learning with complexity $m_H^{APAC}(\varepsilon,\delta)$ such that:

$$\boldsymbol{m_H^{APAC}(\varepsilon,\delta) \le m_H^{CU}(\varepsilon/2,\delta) \Longrightarrow m_H^{APAC}(\varepsilon,\delta) \approx m_H^{CU}(\varepsilon/2,\delta)}$$

Moreover, ERM_H succeeds in the agnostic PAC learning of H.

2.2. Uniform convergence

General Objective

Objective of proof:

Prove that if H is a finite class of hypotheses $\Longrightarrow$ H is agnostic PAC learnable.

But, we know that:

Remarks

If S is ε- representative $\Longrightarrow$ H has the uniform convergence property $\Longrightarrow$ H is agnostic PAC learnable.(theorem 1 and Lemma)

Objective: reformulation

We should simply prove that, if $|H| < \infty$ and **we have sufficient amount of data, S is ε-representative (avec probability 1- δ).**

Then H is agnostic PAC learnable(Remarks)

2.2. Uniform convergence

Proof strategy of the general objective

Step 1:

Consider that H owns one hypothesis h, and let's prove that S is ε- representative:

$$L_{\mathcal{D}}(h) \approx L_S(h)$$

Step 2:

Let's supppose that H owns many hypotheses, and let's prove using Boole's inequality that S is ε- representative:

$$\forall h \in H, \qquad L_{\mathcal{D}}(h) \approx L_S(h)$$

Step 3:

Let's determine the necessary number of data so that S can be ε- representative.

2.2. Uniform convergence

Definition: Law of large scale numbers

Let $(\theta_1, \theta_2, \dots, \theta_m)$ be a random variables $(\boldsymbol{i.i.d.})$, such that μ is their real mean and $\frac{1}{m}\sum_{i=1}^{m}\theta_i$ their empirical mean. $\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| \longrightarrow 0$ when $m \longrightarrow \infty$

So, if $m \longrightarrow \infty$, the empirical mean converges converge to the real mean, with probability equal to 1.

The measures of concentration inqualities are statistical tools that allow to quantify the deviation between the empirical mean and the real mean when m is finite.

Among these inqualities, there exists **''Hoeffding's inequality''**.

2.2. Uniform convergence

Definition: Hoeffding's Inequality

Let's suppose that $(\theta_1, \theta_2, \dots, \theta_m)$ are random variables $(i.i.d.)$, having the real mean μ, such that these variables have values in $[a, b]$. So:

$$P\left[\left|\mu - \frac{1}{m}\sum_{i=1}^{m}\theta_i\right| > \epsilon\right] \leq 2e^{\frac{-2m\epsilon^2}{(b-a)^2}}$$

Real mean
($L_{\mathcal{D}}$ general)

L_S **Empirical mean**

Notice:

This probability decreases exponentially if the size m of the sample increases.

$$P\left[\left|(\mu = L_D(h)) - (\frac{1}{m}\sum_{i=1}^{m}\theta_i = L_s(h))\right| > \varepsilon\right] \leq \delta$$

2.2. Uniform convergence

Proof - Step 1

Let's suppose that $H = \{h\}$, and let's prove that S is ε- representative:

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$$

This implies to prove that :

$$\underset{S \sim \mathcal{D}^m}{\mathrm{P}}[|L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon] \quad \text{is small}$$

To bound this inequality we will use the Hoeffding's Inequality.

We have:

$$\underset{S \sim \mathcal{D}^m}{\mathrm{P}}[|L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon] = P\left[\left|\underset{z \sim \mathcal{D}}{\mathrm{E}}[l(h,z)] - \frac{1}{m}\sum_{i=1}^{m} l(h,z_i)\right| > \epsilon\right]$$

2.2. Uniform convergence

Proof - Step 1

In that case we have:

$\theta_i = l(h, z_i) \in [0,1]$, $\mu = \underset{z \sim \mathcal{D}}{\mathrm{E}}[l(h,z)]$, $\frac{1}{m}\sum_{i=1}^{m} \theta_i = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$

Then:

$$P\left[\left|\underset{z \sim \mathcal{D}}{\mathrm{E}}[l(h,z)] - \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)\right| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

$$\underset{S \sim \mathcal{D}^m}{\boldsymbol{P}}[|\boldsymbol{L}_{\mathcal{D}}(\boldsymbol{h}) - \boldsymbol{L}_{\boldsymbol{S}}(\boldsymbol{h})| > \boldsymbol{\varepsilon}] \leq \boldsymbol{2e^{-2m\epsilon^2}} \quad \textbf{Eq.1}$$

So:

$\exists h \in H,\ \ |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$ if m is sufficiently large.

2.2. Uniform convergence

Proof - Step 2

Now, let's generalize **Eq.1** for all hypotheses $h \in H$.

Let's suppose that H owns several hypotheses, and let's prove that using the Boole's inequality that S is ε- representative:

$$\forall h \in H, \qquad L_{\mathcal{D}}(h) \approx L_S(h)$$

We have:

$(\forall h \in H, L_{\mathcal{D}}(h) \approx L_S(h)) \Leftrightarrow (\forall h \in H$, the probability of failure of h is small)

We have that h fails if S is not ε- representative.

This implies to prove that $\forall h \in H$:

$$P[S \text{ is not } \varepsilon\text{– representative with respect to } H] \text{ is small}$$

2.2. Uniform convergence

Proof - Step 2

$P[S$ is not $\varepsilon-$ representative with respect to $H]$

$$= P[\exists h \in H, |L_D(h) - L_S(h)| > \varepsilon]$$

We have:

$$P[\exists h \in H, |L_D(h) - L_S(h)| > \varepsilon] \leq P[\bigcup_{h \in H} \{h, |L_D(h) - L_S(h)| > \varepsilon\}]$$

According to **Boole's inequality:**

$$P[\bigcup_{h \in H} \{h, |L_D(h) - L_S(h)| > \varepsilon\}] \leq \sum_{h \in H} \underset{\boldsymbol{S \sim \mathcal{D}^m}}{\boldsymbol{P}}[|\boldsymbol{L_{\mathcal{D}}(h) - L_S(h)}| > \boldsymbol{\varepsilon}]$$

According to Hoeffding inequality, we have:

$$\underset{S \sim \mathcal{D}^m}{P}[|L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon] \leq 2e^{-2m\epsilon^2}$$

2.2. Uniform convergence

Proof - Step 2

So:

$$\sum_{h \in H} \underset{S \sim \mathcal{D}^m}{P}[|L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon] \leq \sum_{h \in H} 2e^{-2m\epsilon^2}$$

Then we will have that:

$$P[S \text{ is not } \varepsilon\text{– representative with resppect to } H] \leq |H| 2e^{-2m\epsilon^2}$$

So:

$$\boldsymbol{\forall h \in H \ \underset{S \sim \mathcal{D}^m}{P}[|L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon] \leq |H| 2e^{-2m\epsilon^2}} \quad \textbf{Eq.2}$$

Finally:

$\forall h \in H,\ L_{\mathcal{D}}(h) \approx L_S(h)$ if m is sufficiently big.

2.2. Uniform convergence

Proof - Step 3

Let's determine the necessary number of data so that S can be ε- representative.

We know that $m_H^{CU}(\varepsilon, \delta)$ is the minimal number of data so that S can be ε-representative with probability $\geq 1 - \delta$.

So, we want that $P[S \text{ is not } \varepsilon\text{– representative with respect to } H]$ be $\leq \delta$

So:

$$\underset{S \sim \mathcal{D}^m}{P}[|L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon] \leq |H| 2e^{-2m\epsilon^2} \leq \delta$$

Hereby, the necessary number of data is:

$$\boldsymbol{m \geq \frac{ln\left(\frac{2|H|}{\delta}\right)}{2\epsilon^2}}$$

2.2. Uniform convergence

Theorem2:

Let H be a finite class of hypotheses, let Z be a set of data and $l: H \times Z \longrightarrow [0,1]$ the cost function.

So, H follows a uniform convergence learning with sample complexity:

$$\boldsymbol{m_H^{CU}(\varepsilon, \delta) \leq \left\lceil \frac{ln\left(\frac{2|H|}{\delta}\right)}{2\epsilon^2} \right\rceil \Longrightarrow m_H^{CU}(\varepsilon, \delta) \approx \left\lceil \frac{ln\left(\frac{2|H|}{\delta}\right)}{2\epsilon^2} \right\rceil} (by\ proof\ above\ )$$

Moreover, H follows agnostic PAC learning with ERM_H algorithm, with sample complexity:

$$m_H^{APAC}(\varepsilon, \delta) \leq m_H^{CU}\left(\frac{\varepsilon}{2}, \delta\right) \leq \left\lceil \frac{2\ln\left(\frac{2|H|}{\delta}\right)}{\epsilon^2} \right\rceil (by\ th1)$$

- $\boldsymbol{\Longrightarrow m_H^{APAC}(\varepsilon, \delta) \approx m_H^{CU}\left(\frac{\varepsilon}{2}, \delta\right) \approx \left\lceil \frac{2\ ln\left(\frac{2|H|}{\delta}\right)}{\epsilon^2} \right\rceil}$

Supervised Learning Passive Offline Algorithm (SLPOA)