MACHINE LEARNING

giovanni.neglia@inria.fr
othmane.marfoq@inria.fr

EVALUATION

(1) EXAM (40%)
(2) HOMEWORK (30%)
(3) QUESTION AT EVERY LECTURE (30%)

2020/2021

16/09

INTRODUCTION

MACHINE LEARNING

- way of extract rules from data

DATA → [ML] → RULES/PROGRAM

ISSUES

- To large set of rules (REM Pidgeon)
- To small set of rules (REM Rat)

& TOO MUCH reduces learning flexibility

INDUCTIVE PRIOR

- A PRIORI KNOWLEDGE to prevent useless conclusion

INDUCTIVE INFERENCE

- Ability to proceed from general examples to a broadere generalization

KINDS OF LEARNING

SUPERVISED / UNSUPERVISED

- In supervised learning output label are given; Unsupervised instead finds unexpected correlation. (SPAM MAIL vs ANOMALITY DETECTION)

REINFORCHENT

- Labels are given but after a while

COOPERATIVE / INDIFFERENT / ADVERSARIAL TEACHER

- Change the way the data is given from a teacher

BATCH / ONLINE

- Data can be given all together or PASSO PASSO

! ⇒

- we'll se SUPERVISE STATISTICAL BATCH LEARNING WITH PASSIVE TEACHER

STATISTIC & ML

1. Statistic wants to check an HYPOTHESIS (■ smoke affects heart?)
   ML wants to find the HYPOTHESIS (■ what affects heart?)
2. Statistic starts from the Prob. Distr. of data.
   ↳ ASYNTOTICS: if you look at enough data you find a GAUSSIAN
   ML doesn't know the distribution
   ↳ FINITE SAMPLES: you get the distribution you have

STATISTICAL LEARNING FRAMEWORK

- X input space
- Y output space

} (vector of features)

- S dataset of size m (# SAMPLES)

---

f and D are unknown to the learner

- f correct labeling function ; D distribution over X
- $A : S \subset (X \times Y)^m \longrightarrow \{\text{functions } X \to Y\}$
  $\Rightarrow A(S) = h$

h is the output of an ML Algo. (is a PREDICTION RULE)

- $h : X \to Y$ called HYPOTESIS or PREDICTOR or CLASSIFIER it can be seen as $A(S)$, so the output of the ML ALGO. A given the DS S

INDICATOR FUNCTION: 0 if prediction is ok it is 0 else is 1

EXPECTED LOSS

- $L_D(h) = \mathbb{E}_D\left(\mathbb{1}_{h(x) \neq y}\right) \in [0,1]$

EMPIRICAL LOSS

- $L_S(h) = \dfrac{\sum_{i=1}^{m} \mathbb{1}_{h(x_i) \neq y_i}}{m}$

$\Rightarrow$ for large number theory, EMPIRICAL LOSS will converge to EXPECTED LOSS.

---

note

- we can write $y = f(x)$ so we can rewrite $L_{D,f}(h) = \mathbb{E}_D(\mathbb{1}_{h(x) \neq f(x)})$; note also that our hope is to find h such that $L_{D,f}(h) < \varepsilon$ with Prob $> 1 - \delta$ because in the best case we can have $L_{D,f}(h) = 0$ (that is obv. impossible)

APPROXIMATE WITH HIGH PROBABILITY

ERP - EMPIRICAL RISK MINIMIZATION

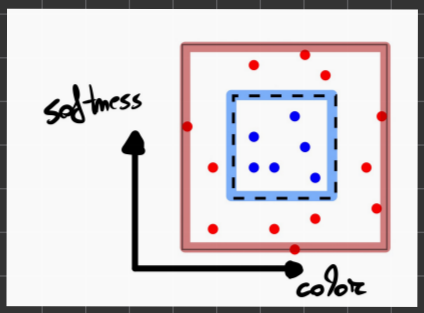
Def

- ERP is the process of finding a PREDICTOR h which minimizes the EMPIRICAL Loss $L_S(h)$.

OVERFITTING

- When a predictor is excellent on the training set but very poor on the distribution

ERM is bad?



- Uniform D
- AREA( ) = 1
- AREA( ) = 2
- $f(x)$ { 1 if in , 0

Overfitting example

- let $S =$ and define:

$$h_S(x) = \begin{cases} y, & \forall x \in S \\ 0 \end{cases}$$

it means that $h_S(x)$ is perfect for the training set, so $L_S(h_S) = 0$; but, having $P = 1/2$ of taking an element $x \notin S \Rightarrow L_D(h_S) = 1/2$, so it's bad for the distribution D

- MEMORIZATION ALGO.

other OVERFITTING EXAMPLE

- Suppose: $h_M(x) = \begin{cases} 1 \text{ if } x = x_i \in S \\ \text{TOSS A COIN} \end{cases}$
- the P of taking an already known SAMPLE from all over the distribution D is ϕ, so you always toss a coin.

$\Rightarrow$ it perfectly explain the dataset, but is shit in general

ERM WITH INDUCTIVE BIAS

⇒ ERM is not bad but learns in a too large set of HYPOTESIS, so it depends from the set of HYPOTESIS you apply it.

Def
- The IDEA is to reduce the SEARCH SPACE of ERP;
- we define, before looking at the data, a set of predictor H called HYPOTESIS CLASS, and we use ERP rule to find $h_{ERM_H} = ERM_H(S) \in \underset{h \in H}{\operatorname{argmin}} L_S(h)$.

INDUCTIVE BIAS
- mean biasing a LEARNER to a finet set of predictor.

?
HOW TO REDUCE THE SEARCH SPACE H?
- UPPERBOUNDING IT'S SIZE

REALIZABILITY ASSUMPTION
- There exist alway $h^* \in H \mid L_{D,f}(h^*) = 0$, it means $h^* = f$

⇒ it implies that, if the set H is finite, then for every S taken from D $L_S(h^*) = 0$, so ERM_H will alway find it.

ISSUE
- We are interested in $L_{D,f}(h_{ERM_H})$, so to be well done S should be enough representative of the underlying dist. D

↓

i.i.d. ASSUMPTION
- All sample of S are INDIPENDENTLY and IDENTICALLY DISTRIBUTED according to D

↳ it implies that the bigger is S the more it is representative of D and f

---

NOTE

- $L_{D,f}(h_S)$ is a random variable because the choice of S is random, so can always happen that S would be not representative of D

CONFIDENCE PARAMETER

- δ is the PROB. of getting NON-REPR. S
- $(1-\delta)$

ACCURACY PARAMETER

- ε is needed when to address the QUALITY of a Prediction

⇒ if $L_{D,f}(h_S) > \varepsilon$ FAIL; else if $L_{D,f} \leq \varepsilon$ we got an APPROX. CORRECT PREDICTOR h_S

? IDEA

Can we upp. bound the prob. of sampling S from D in a way that S want be MISLEADING?

We can find the right size $m = |S|$ such that ERM_H won't choose a $h_{ERM_H} = h_S$

Some Defs

- BAD HYPOTHESIS SET → $H_B := \{h \in H \mid L_{D,f}(h) > \varepsilon\}$
- POSSIBLY MISLEADING S →

$$S_p := \{S \mid \exists \underbrace{h_S \in H_B, L_S(h_S) = 0}_{\text{BAD HYPOT. THAT LOOK GOOD ON } S}\}$$

- MISLEADING S →

$$S_M \in \{S \mid \underbrace{L_{D,f}(h_S) > \varepsilon}_{\text{IT HAPPENS ONLY IF } h_S \in H_B}\} = \{S \mid h_S \in H_B\} \subseteq S_p$$

- we can rewrite $S_P = \bigcup_{h_S \in H_B} \{S \mid L_S(h_S) = 0\}$

Prob. of picking S_M

- we now calculate $\mathbb{P}_{D^m}(S_M)$

$$\mathbb{P}_{D^m}(S_M) \leq \mathbb{P}_{D^m}(S_P) = \mathbb{P}_{D^m}\left(\bigcup_{h_S \in H_B} \{S \mid L_S(h_S) = 0\}\right) \leq$$

$$\leq \sum_{h_S \in H_B} \mathbb{P}_{D^m}(\{S \mid L_S(h_S) = 0\}) \dots$$

- Notice that $L_S(h_S) = 0$ if $\forall (x_i, y_i) \in S \; h_S(x_i) = y_i \Rightarrow$

$$\Rightarrow \mathbb{P}(h_S(x_i) = y_i) = 1 - L_{D,f}(h_S) \leq 1 - \varepsilon \Rightarrow$$

($L_{D,f}(h_S) > \varepsilon$ (searching S_M!))

$$\Rightarrow \mathbb{P}(h_S(x_i) = y_i \; \forall i = 1, \dots, m) = \mathbb{P}_{D^m}(\{S \mid L_S(h_S) = 0\}) \leq (1-\varepsilon)^m$$

$$\dots \sum_{h_S \in H_B} \mathbb{P}_{D^m}(\{S \mid L_S(h_S) = 0\}) \leq \sum_{h_S \in H_B} (1-\varepsilon)^m =$$

$$= |H_B|(1-\varepsilon)^m \leq |H| e^{-\varepsilon m}$$

(obv $H_B \subseteq H$)

$\Rightarrow$ - in conclusion $\boxed{\mathbb{P}_{D^m}(S_M) \leq |H_B| e^{-\varepsilon m}}$

BOUND

- we want $\mathbb{P}_{D^m}(S_M) \leq \delta$ so we should choose m carefully:

$$|H_B| e^{-\varepsilon m} \leq \delta \Rightarrow \boxed{m \geq \frac{\log(|H|/\delta)}{\varepsilon}}$$

Conclusion

COROLLARY 2.3 *Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$ and let m be an integer that satisfies*

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}.$$

Then, for any labeling function, f, and for any distribution, $\mathcal{D}$, for which the realizability assumption holds (that is, for some $h \in \mathcal{H}$, $L_{(\mathcal{D},f)}(h) = 0$), with probability of at least $1 - \delta$ over the choice of an i.i.d. sample S of size m, we have that for every ERM hypothesis, h_S, it holds that

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon.$$

The preceeding corollary tells us that for a sufficiently large m, the $\text{ERM}_{\mathcal{H}}$ rule over a finite hypothesis class will be *probably* (with confidence $1-\delta$) *approximately* (up to an error of ϵ) correct.

21/09/21

PAC - LEARNABILITY

PAC LEARNABILITY Def 1

. (Given REALIZ. ASS and 0-1 loss funct $\mathbb{1}_{h \neq y}$) H is PAC-LEARNABLE if:

- $\exists A : (X \times Y)^m \to H$
- $\exists m_H : (0,1)^2 \to \mathbb{N}$ // SAMPLE COMPLEXITY how many samples are required to guarantee an approx. correct solution?

They are such that:

$\forall D$ over X, $\forall f$, $\forall \delta \in (0,1)$, $\forall \varepsilon \in (0,1)$ if we get S of m i.i.D. samples according to D from X such that $|S| = m \geq m_H(\varepsilon, \delta)$, then:

$$L_{D,f}(A(S)) < \varepsilon \quad \text{w.p. } (1-\delta)$$

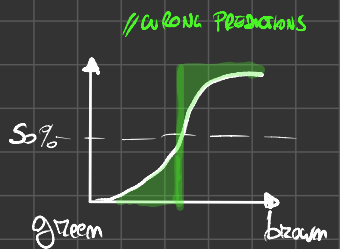
weak points of Def 1

1. REALIZABILITY ASSUMPTION → $\exists f$, $f \in H$
2. 0-1 loss (only binary classification)

=>

- $f(x)$ is wasted and we'll use $P(y|x)$
- We need a THRESHOLD for $P(y|x)$, and put it to 50% is the choiche that reduces $L_{D,P}(h)$

$$h(x) = \begin{cases} 1 & , P(1,x) = 0,5 \\ 0 \end{cases}$$



MARGINAL DIST

- $P(x)$

CONDITIONAL LABELLING PROB

- $P(y|x) = \dfrac{P(x,y)}{P(x)}$

LOSS GENERALIZATION

- There exist different kinds of loss other than the (0-1)-loss:

some losses:

- $\ell(h,(x,y)) \in \mathbb{R}$
- $\ell(h,(x,y)) = \mathbb{1}_{h(x)\neq y}$ ((·)² cause we want $\ell \geq 0$)
- $\ell(h,(x,y)) = (h(x)-y)^2$

- we can generalize them by $\Rightarrow L_{\mathcal{D}}(h) = \mathbb{E}[\ell(h,(x,y))]$

REAL RISK EVALUATION

- Until now, with realizability ass., we were sure that the best possible $L_{\mathcal{D}}(h)$ was 0 (because $\exists h^* \in H \mid h^* = f$); relaxing this assumption we need the to compare our loss to be the nearest possible to the BEST ONE we can achieve in H.

$$L_{\mathcal{D}}(h) < \min_{h' \in H} L_{\mathcal{D}}(h') + \varepsilon$$

AGNOSTIC PAC LEARNABILITY

Def2

- ~~(Given realiz. ass on 0-1 loss funct $\mathbb{1}_{h\neq y}$)~~ H is AGNOSTIC PAC-LEARNABLE, if:
  - $\exists A : (X \times Y)^m \to H$
  - $\exists m_H : (0,1)^2 \to \mathbb{N}$

w.r.t. the loss ℓ

They are such that:

$\forall \mathcal{D}$ over $(X \times Y)$, ~~$\forall f$~~, $\forall \delta \in (0,1)$, $\forall \varepsilon \in (0,1)$ if we get S of m i.i.d. sampled according to $\mathcal{D}$ from $(X \times Y)$ such that $|S| = m \geq m_H(\varepsilon, \delta)$, ~~then~~:

~~$L_{\mathcal{D},f}(A(S)) < \varepsilon$ w.p. $(1-\delta)$~~

$$L_{\mathcal{D}}(A(S)) \leq \min_{h} L_{\mathcal{D}}(h) + \varepsilon \quad \text{w.p.} \geq 1-\delta$$

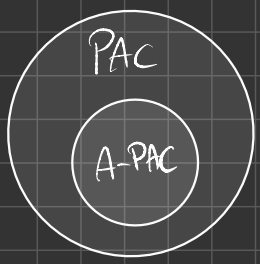
BEST POSSIBLE ERROR OVER H. w/ REALIZABILITY WAS 0 BECAUSE OF $f \in H$

NB

- What is true?

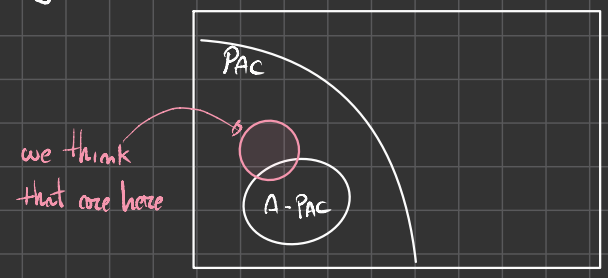
① H is A-PAC $\Rightarrow$ H is PAC

② H is PAC $\Rightarrow$ H is A-PAC



$\Rightarrow$ ①: if H satisfy PAC LEARN. constr, it's the obvious that satisfy also AGNOSTIC PAC LEARN. constr.

Where are finite class of HIPOTHESIS? we know are PAC, but no who they are



ε-REPRESENTATIVE

ε-REPRESENTATIVE DATASET

- S is ε-REPRESENTATIVE if $|L_D(h) - L_S(h)| < \varepsilon \quad \forall h \in H$
- if you have S ε-REPR, then ERM_H finds a good predictor
  - let $h_S = ERM_H(S)$
  - let S be $\varepsilon/2$-REPRESENTATIVE
  - we know by def that

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

↕ BY DEF

$$\leq L_D(h) + \varepsilon/2 + \varepsilon/2 \quad \forall h$$
$$\leq \min_{h \in H} L_D(h) + \varepsilon$$

↑ BY A-PAC LEARNABILITY

UNIFORM CONVERGENCE PROPERTY (UC)

- H has UC prop. w.r.t. the loss ℓ $\exists m_H^{UC}: (0,1)^2 \to \mathbb{N}$ $\forall D$, $\forall \varepsilon, \delta \in (0,1)$, if you draw a dataset S i.i.d from D with $|S| = m \geq m_H^{UC}(\varepsilon, \delta)$ then S is ε-REPR. w.p. $\geq 1-\delta$

NOTE

- If H has UC prop. then S is ε-repr. with high prob., so $L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon$, so H is AGNOSTIC PAC LEARNABLE, and $ERM_H(S)$ an AGNOSTIC PAC LEARNER for H

TH

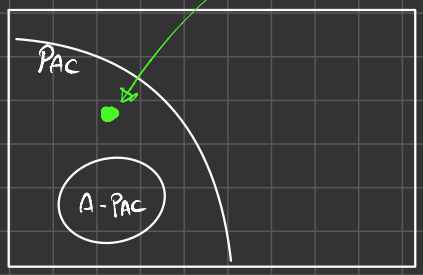
- If H is finite $\Rightarrow$ H has UC prop

Proof

Postponed...

TH

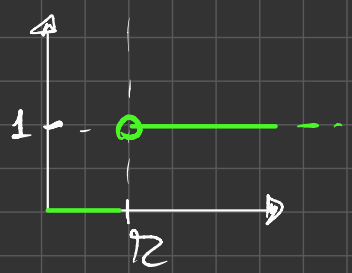
- There is at least an INFINITE PAC CLASS H



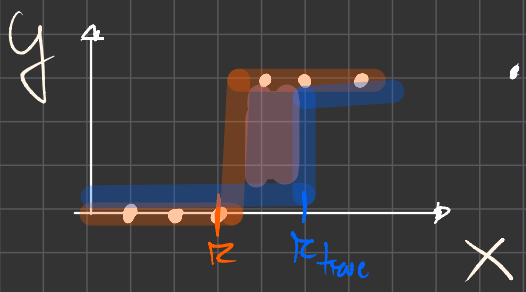
Proof

- Let H be a threshold function

$H = \{ h_\pi, \pi \in [0,1], h_\pi : [0,1] \to \{0,1\}, h_\pi(x) = \begin{cases} 1 & \forall x > \pi \\ 0 & \forall x \leq \pi \end{cases} \}$



- Given the dataset

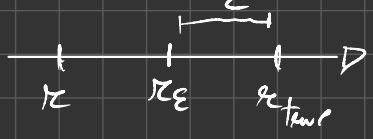


- ERM_H gives me h_z where $z = \max\{x_i \mid y_i = 0\}$

- suppose the $\mathcal{D}$ every time we observe a point in the PINK area we have a loss of 1

$$L_D(h_z) = \mathbb{P}(x \in [z, z_{true}])$$

- now we look a point in $[z, z_{true}]$ such that the error on it is ε



- z_ε: $\mathbb{P}(x \in [z, z_{true}]) = \varepsilon$

- if $z < z_\varepsilon \Rightarrow L_D(h_z) > \varepsilon$ (1)
- if $z \geq z_\varepsilon \Rightarrow L_D(h_z) < \varepsilon$ (2)

$\mathbb{P}(z < z_\varepsilon) =$

$\Rightarrow \mathbb{P}(x \notin [z, z_{true}]) = 1 - \varepsilon$

$= \mathbb{P}(x_i \notin [z, z_{true}] \ \forall i = 1, \dots, m) = \prod_{i=1}^{m} \mathbb{P}(x \notin [z, z_{true}]) =$

$= \prod_{i=1}^{m} (1 - \varepsilon) = (1-\varepsilon)^m \Rightarrow$ we want $, t \leq \delta$

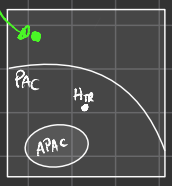
$\Rightarrow$ if we take $m \geq \frac{\log(\delta)}{\log(1-\varepsilon)}$ then $L_D(ERM_S) \leq \varepsilon$

w.p $\geq 1 - \delta$ so H is PAC

∎

28/09

- We will show that There is a class $H_{\{0,1\}^X} = \{$All binary partition over $X\} \mid H_{\{0,1\}^X} \notin PAC$



TH [NO FREE LUNCH]

A

- $\exists D$ over X and a labeling function f, $\forall A$ learning algorithm for binary classification, those are s.t. by picking S of $m < |X|/2$ i.i.d. according to D you have

$$L_{D,f}(A(S)) \geq 1/8 \quad \text{w.p.} \geq 1/7 \leftarrow \text{NON NEGLIGIBLE}$$

PAC NEGATION

- H is not pac if $\forall A$, $\forall m_H : (0,1)^2 \to \mathbb{N}$, $\exists D$ over X and $\exists f \in H$, $\exists \varepsilon_0, \delta_0 \in (0,1)$: if $\exists m \geq m_H(\varepsilon_0, \delta_0) \mid |S| = m$ then EVERY A learn with an error of:

$$L_{D,f}(A(S)) \geq \varepsilon_0 \text{ w.p } \geq 1-\delta_0$$

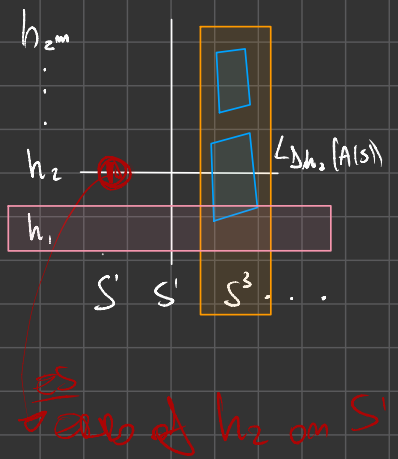
→ FROM NO FREE LUNCH TH we can enforce the ¬PAC definition

- $\exists f$ instead of $\exists f \in H$
- $\varepsilon_0 = 1/3$ and $\delta_0 = 1/7$
- the requirement $\exists m \geq m_H(\varepsilon_0, \delta_0)$ that became $\forall m$

free lunch proof

B

- Pick $D \mid D(x_i) = 1/2^m \ \forall i$
- Let's proof that $\exists f \mid \mathbb{E}_{S \sim D^m}[L_{D,f}(A(S))] \geq 1/4$
- Build a table of $|H|$ rows with every possible S on column



// pick 4 SAMPLES AT A TIME

// the average on each ☐ has an AVG ERROR of 1/4

// the set of a column ☐ has ... 1/4

// the AVG BS too has ERR 1/4

// AVG of elements on RAW ☐ 1/4

⇒ ∃ RAW with AVG ≥ 1/4, H's the RAW of f

↳ Proof if the whole set has AVG ≥ 1/4 there is at least one with AVG ≥ 1/4

- Now show that $A \Rightarrow B$ by showing $\neg A \Rightarrow \neg B$

(A) $L_{D,f}(A(S)) \geq 1/8$ w.p. $\geq 1/7$

(B) $\exists f \mid \mathbb{E}_{S \sim D^m}[L_{D,f}(A(S))] \geq 1/4$

How can $L_{D,f}(A(S)) = 1$? happens w.p. $< 1/7$ so

$$\mathbb{E}_{S \sim D^m}[L_{D,f}(A(S))] < 1 \cdot 1/7 + 1/8 \cdot 6/7 = 1/4$$

COROLLARY [NO FREE LUNCH]

- If $|X| = +\infty$, $H_{\{0,1\}^X}$ is not PAC learnable w.r.t. 0-1 loss

Proof

- Assume H is PAC and choose $\varepsilon < 1/8$ and $\delta < 1/7$
- By PAC def: $L_D(A(S)) \leq \varepsilon$ w.p. $> 1 - \delta$
- By NO-FREE LUNCH: $\exists D \mid L_D(A(S)) > 1/8 > \varepsilon$ w.p. $> 1/7 > \delta$ ⊥

VC DIMENSION

- Finitness of H is a SUFFICIENT but not NECESSARY condition for it's LEARNABILITY

why VC-DIMENSION?

- $PAC \equiv A\text{-}PAC \equiv UC \equiv \{$ everything we can learn via ERM$\} \equiv$
$\equiv VC\text{-}dim(H) < +\infty$

SHATTERING

- Given X, H over X and $A \subset X$ subset of X
- H shatters A if $\forall g: A \to \{0,1\}$ $\exists h \in H \mid h(x) = g(x)$ $\forall x \in A$

shattering example

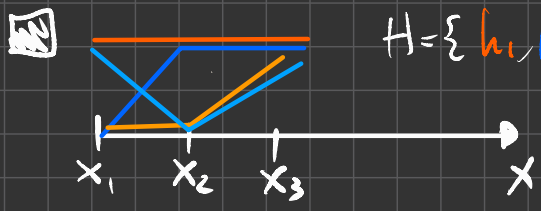
$H = \{h_1, h_2, h_3, h_4\}$

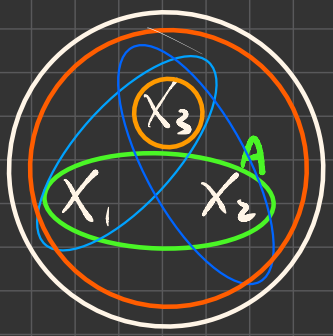
- $A = \{x_1\} \Rightarrow H$ shatter x_1
- $A = \{x_3\} \Rightarrow H$ DOESENT (you cant get $x_3 = 1$ with $h \in H$)
- $A = \{x_1, x_2\} \Rightarrow H$ shatter A (we have 00, 01, 10, 11 in h)

why shattering

- given a subset of Point, we can always define H in a way that shatter them, so H can be seen as a SUBSET of X:

$\Rightarrow H$ shatters A if $\forall g$ subset of A $\exists h \in H \mid h \cap g = 0$





- To get $\{\emptyset\}$ we get x_3 set cause $\{x_3\} \cap A = \{0\}$
- To get x_1 we got ◯
- To get x_2 we get ◯

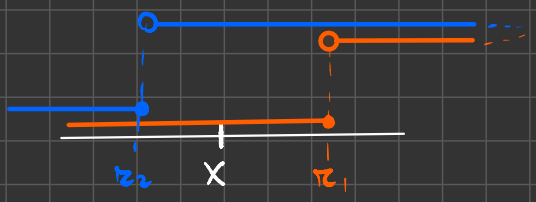
□

VC-dim

- $\text{VC-dim}(H) = \max\{|A| \mid H \text{ shatters } A\}$

Find VC-dim of $H_{\text{THRESH}} = \{h_{x_2} \mid h_{x_2}(x) \begin{cases} 1 & \text{if } x > x_2 \\ 0 & \end{cases}\}$

raw/bunch



real set

- we can pick him by taking the set $X \setminus \{x\}$ cause

$A \cap X \setminus \{x\} = \{\emptyset\}$

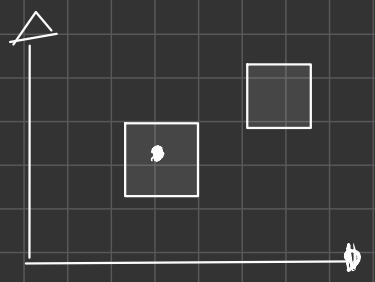
- Now we know $\text{VC-dim}(H) \geq 1$
- we can't do the same for 2 points because there is no way of picking only one of them

$\Rightarrow \text{VC-dim}(H) = 1$

find VC-dim of $H_{rect} = \{h_{a,b,c,d}$ s.t.

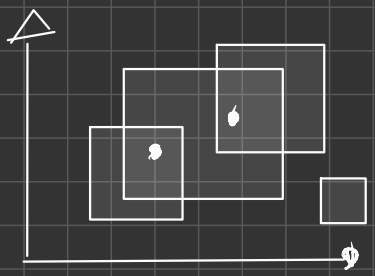
$h_{a,b,c,d} \begin{cases} 1 & \text{if } 0 \le a \le b \text{ and } 0 \le c \le b \\ 0 & \end{cases} \}$

1 point

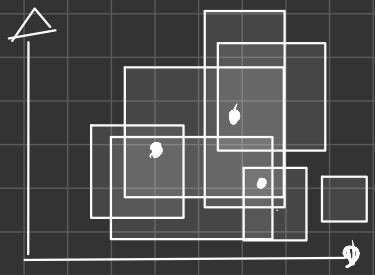


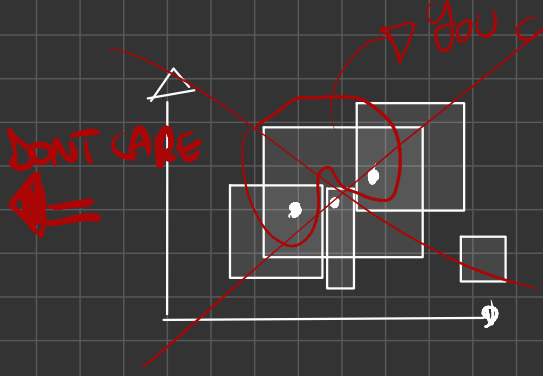
$\Rightarrow$ VC-dim(H_{rect}) ≥ 1

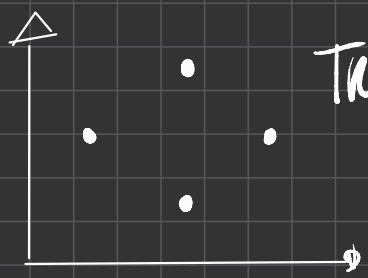
2 point



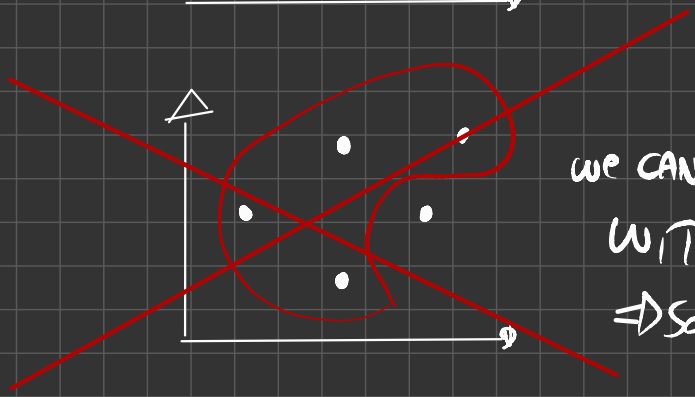
3 point







TRUST, WE CAN!



WE CANT SHATTER 4 POINTS WITHOUT THE 5TH !

$\Rightarrow$ So you proved also for more than 5 points

$\Rightarrow$ VC-dim(H) = 4

FUNDAMENTAL TH OF STATISTICAL LEARNING

TH FTSL

· X, 0-1 LOSS, H

(1) H IS PAC LEARNABLE

(2) ERM_H IS PAC LEARNER

(3) H IS AGNOSTIC PAC LEARNABLE

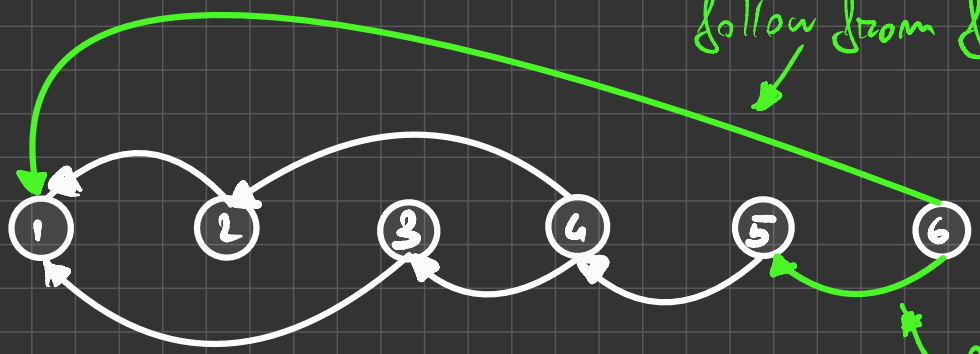
(4) ERM_H IS AGNOSTIC PAC LEARNER

(5) H HAS UC PROPERTY

(6) $VC\text{-}dim(H) < +\infty$

Proof

we don't Know



Codio, non lo vediamo che ce vole troppo

(6-1) Proof

TH. H is PAC $\Rightarrow$ $VC\text{-}dim(H) < +\infty$

Proof.

NEGATED TH:

$VC\text{-}dim(H) > +\infty \Rightarrow H$ is NOT PAC

... vedi Alessio, io non ce la faccio più a sentire Porcodio

5/10/21

TH QUANTITATIVE FTSL

// You can see how VC dim characterise the complexity of different class of learning

. Suppose $d = VCdim(H) < +\infty$, $\exists C_1, C_2 \in \mathbb{R}$ s.t.

(1) H is PAC-LEARNABLE with $m_H(\varepsilon, \delta)$:

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

(2) H is APAC-LEARNABLE with $m_H(\varepsilon, \delta)$:

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

(3) H has UC-PROPERTY with $m_H(\varepsilon, \delta)$:

$$C_1 \cdot // \leq m_H^{UC}(\varepsilon, \delta) \leq C_2 \cdot //$$

LINEAR PREDICTORS

- Linear predictors are a FAMILY of HYPOTHESIS CLASS

CLASSES	ALGORITHMS
HALFSPACES	LP & PERCEPTRON
LINEAR REGRESSION PREDICTORS	
LOGISTIC REGRESSION PREDICTORS	LEAST SQUARE

AFFINE FUNCTIONS

- $L_d = \{ h_{w,b} : X \to \langle w, x \rangle + b \mid w \in \mathbb{R}^d \wedge b \in \mathbb{R} \}$
- $h_{w,b}(x) = \langle w, x \rangle + b =$
- we can incorporate the bias b in w

$$w' = (b, w_1, \_, w_d) \in \mathbb{R}^{d+1}, \quad x' = (1, x_1, \_, x_d) \in \mathbb{R}^{d+1}$$

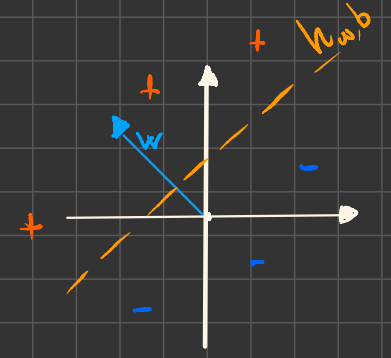
HALFSPACES

- Designed for BINARY CLASSIFICATION problems
- $X = \mathbb{R}^d$, $Y = \{-1, +1\}$

$$HS_d = L_d \circ SIGN = \{ X \to SIGN(h_{w,b}(x)) \mid h_{w,b} \in L_d \}$$

Ex. w/ $d = 2$

EACH HYPOTESIS FORM AN HYPERPLANE PERPENDICULAR TO w



SEPARABLE AND UNSEPARABLE CASES

- Separable mean that we can separate all POSITIVE samples from NEGATIVE with an HYPERPLANE.
- The NON-Separable is NP-HARD

⇒ We have 2 method for IMPLEMENTING ERM FOR HALFSPACES IN SEPARABLE CASE (ie. LP and PERCEPTRON)

LINEAR PROGRAMMING FOR HS

- ERM predictor should have 0 ERROR on the TS, so we looks for a vector w^* s.t.

$$SIGN(\langle w^*, x_i \rangle) = y_i \quad \forall i = 1, \_, m \Rightarrow$$

$$\Rightarrow y_i \langle w^*, x_i \rangle > 0 \quad \forall i = 1, \_, m$$

- Let $\varphi = \min_{i=1,\_,m} \{ y_i \langle w^*, x_i \rangle \}$ and $\bar{w} = \frac{w^*}{\varphi}$ we have

$$y_i \langle \bar{w}, x_i \rangle = \frac{1}{\varphi} y_i \langle w^*, x_i \rangle \geq 1 \quad \forall i = 1, \_, m$$

- So $\exists w \mid y_i \langle w, x_i \rangle \geq 1 \ \forall i = 1 \_ m$
- This is our ERM predictor

- So the ERM LP is:

$$\max_{w \in \mathbb{R}^d} \langle w, u \rangle \quad \text{s.t.} \quad Aw \geq v$$

where

- $A_{m \times d} \mid A_{i,j} = y_i \cdot x_{ij}$
- $v = (1, \_, 1) \in \mathbb{R}^m$
- $u = (0, \_, 0) \in \mathbb{R}^d$ // because every w that satisfy the constraints are equal candidates

---

PERCEPTRON FOR HS

- ERM wants 0 error on the TS so $\text{SIGN}(\langle w^*, x_i \rangle) = y_i \quad \forall i = 1 \_ m$
- At each step t if there is an $x_i \mid \text{SIGN}(\langle w^t, x_i \rangle) \neq y_i$ the algo. update w^t in such a way $w^{t+1} = w^t + y_i x_i$ to accomplish $y_i \langle w^{t+1}, x_i \rangle > 0$

Batch Perceptron

input: A training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
initialize: $\mathbf{w}^{(1)} = (0, \ldots, 0)$
  for $t = 1, 2, \ldots$
    if $(\exists\, i \text{ s.t. } y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0)$ then
    $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
    else
      output $\mathbf{w}^{(t)}$

- At the end of the execution all SAMPLES of TS will be correctly classified

TH

- Given a SEPARABLE TS
- Let $B = \min \{ \|w\| \mid y_i(\langle w, x_i \rangle) \geq 1 \quad \forall i = 1, \_, m \}$
- Let $R = \max_i \{ \|x_i\| \}$

$\Rightarrow (RB)^2$ ITERATIONS

- It stops with $y_i \langle w^t, x_i \rangle > 0 \quad \forall i = 1 \_ m$

VCdim(HS)

- **HOMOGENEOUS HALFSPACE**: is an halfspace that does contain the $\emptyset$-VECTOR

  ⇒ So the induced HYPERPLANE pass through the ORIGIN.

TH

- $VCdim(HS) = d$ for the class of HOMOGENEOUS HALFSPACES in $\mathbb{R}^d$

Proof

- Consider the set $(e_1, \_\_, e_d) \mid e_{is} = 0 \ \forall s \neq i \ \wedge \ e_{ii} = 1$
- This set is shattered by the Hom. HS class, by simply setting w as the labeling $w = (y_1, \_, y_d) \Rightarrow \langle w, e_i \rangle = y_i \ \forall i$
- Let $x_1, \_, x_d, x_{d+1} \in \mathbb{R}^d$ then must exist $a_1, \_, a_{d+1} \in \mathbb{R}$ not all zeros s.t. $\sum_{i=1}^{d+1} a_i \cdot x_i = 0$ (LINEAR DEPENDENCE)
- Let $I := \{i \mid a_i > 0\}$, $J := \{j \mid a_j < 0\}$
- Assume both non empty:

$$\sum_{i \in I} a_i x_i - \sum_{j \in J} |a_j| x_j = 0$$

- Suppose $x_1, \_, x_{d+1}$ are shattered then

$$\exists w \mid \langle w, x_i \rangle > 0 \ \forall i \in I \ \wedge \ \langle w, x_j \rangle < 0 \ \forall j \in J$$

it implies

$$0 < \sum_{i \in I} a_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \left\langle \sum_{i \in I} a_i \mathbf{x}_i, \mathbf{w} \right\rangle = \left\langle \sum_{j \in J} |a_j| \mathbf{x}_j, \mathbf{w} \right\rangle = \sum_{j \in J} |a_j| \langle \mathbf{x}_j, \mathbf{w} \rangle < 0,$$

- if I is empty

$$0 = \sum_{i \in I} a_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \left\langle \sum_{i \in I} a_i \mathbf{x}_i, \mathbf{w} \right\rangle = \left\langle \sum_{j \in J} |a_j| \mathbf{x}_j, \mathbf{w} \right\rangle = \sum_{j \in J} |a_j| \langle \mathbf{x}_j, \mathbf{w} \rangle < 0,$$

- if J is empty

$$0 < \sum_{i \in I} a_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \left\langle \sum_{i \in I} a_i \mathbf{x}_i, \mathbf{w} \right\rangle = \left\langle \sum_{j \in J} |a_j| \mathbf{x}_j, \mathbf{w} \right\rangle = \sum_{j \in J} |a_j| \langle \mathbf{x}_j, \mathbf{w} \rangle = 0,$$

TH • $VCdim(HS) = d+1$ for the class of NON-HOMOGENOUS HALFSPACES in $\mathbb{R}^{d+1}$

Proof • As before consider the set $(0, e_1, \dots, e_d)$ is shattered. Then suppose $x_1, \dots, x_{d+2} \in \mathbb{R}^{d+1}$, we can reach a $\perp$ as before.

ERROR DECOMPOSITION

$$L_D(A(S)) = \underbrace{L_D(A(S)) - \min_{h \in H} L_D(h)}_{\text{ESTIMATION ERROR}} + \underbrace{\min_{h \in H} L_D(h)}_{\text{APPROXIMATION ERROR}}$$

($\leq \varepsilon$ for PAC LEARNABILITY)

NOTE • if $H \subset H'$ we have

$$\varepsilon_{APPROX}(H') \leq \varepsilon_{APPROX}(H)$$

$$\varepsilon_{ESTIMATION}(H) \leq \varepsilon_{ESTIMATION}(H')$$

• The bigger is H the easier the OVERFITING PROBABILITY

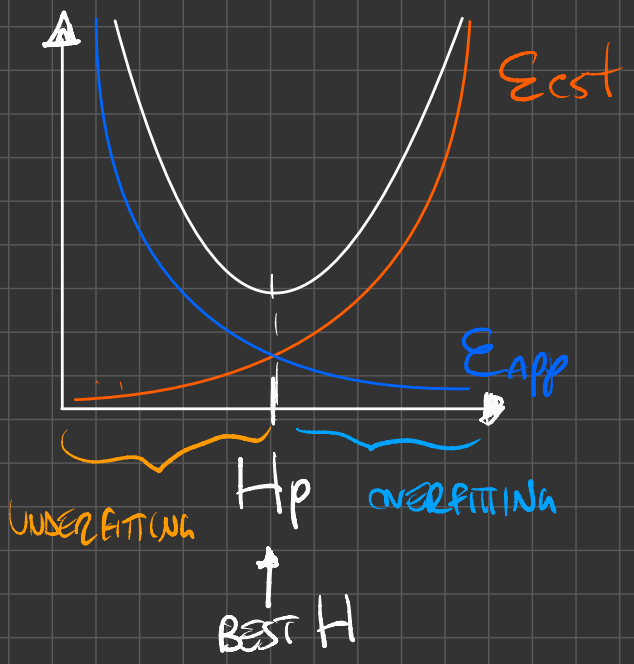
• From the QUANT. FTSL:

$$m_H(\varepsilon, d) \approx \frac{d_H + \ln(1/\delta)}{\varepsilon^2}$$

• so given m:

$$\varepsilon_{ESTIM} \approx \sqrt{\frac{d_H + \ln(1/\gamma)}{m}}$$

- Graphically:



- What is the best possible error?

$L_D(A(S)) =$

$$\underbrace{L_D(A(S)) - \min_{h \in H} L_D(h)}_{\text{ESTIMATION ERROR}} + \underbrace{\min_{h \in H} L_D(h) - \min_{h \in H} L_D(h)}_{\text{APPROXIMATION ERROR}} + \underbrace{\min_{h \in H} L_D(h)}_{\text{BIAS ERROR (IRREDUCIBLE)}}$$

WEAK LEARNABILITY, BOOSTING

- We look for a simpler LEARNABILITY definition leading to more efficient solution.

HOPE

- Something NP-HARD for PAC LEARN. becomes POLYNOMIAL for WEAK LEARN. (of course they fucked up)

CONCLUSION

- WL ≡ PAC L.
- SHAPIRE showed that a POLY. ALGO. to WEAKLY LEARN H can be reduced in polynomial time in a PAC LEARNER for H (che figa)

- The poly. reduction algorithm is called **Boosting**

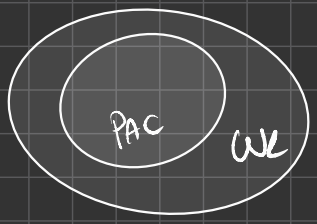
WEAK LEARNABILITY

- Assumed 0-1 loss and Realizability
- H is γ-WL if $\exists A: (X \times Y)^m \to H$ and $\exists m_H^{WL}: (0,1) \to \mathbb{N}$ s.t. $\forall D$ over X, $\forall f \in H$, $\forall \delta$ if $S \sim D^m$ i.i.d $|S| = m \geq m_H^{WL}(\delta)$ then:

$$L_{D,f}(A(S)) \leq \frac{1}{2} - \gamma \quad \text{w.p. } 1-\delta$$

$\gamma \in (0, \frac{1}{2})$ // we want the learner a bit better than a coin tosser

- Because WL require a stupid learner:



- But the truth is PAC-L ≡ WL, we will show why.

TH

- WL ≡ PAC

Proof

① PAC-L ⇒ WL : $\exists \gamma \in (0, \frac{1}{2}) \mid H$ is γ-WL?

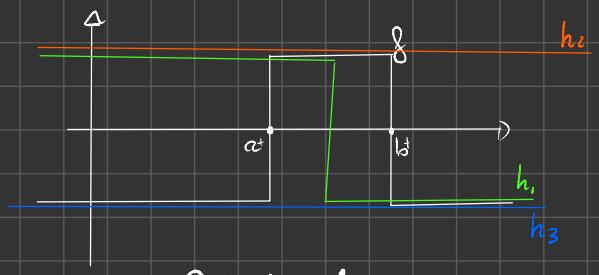
- if $H \in PAC \Rightarrow L_{D,f}(A(S)) < \varepsilon$ w.p. $\geq (1-\delta)$
- suppose $H \in PAC$ is also $H \in WL$:
- we can choose $\gamma = \frac{1}{4} \Rightarrow L_{D,f}(A(S)) \leq \frac{1}{4} - \frac{1}{2} - \gamma$

② WL ⇒ PAC-L : Proof by $\neg PAC \Rightarrow \neg WL$

- let H not PAC $\Rightarrow VCd(H) = +\infty$
- For quant. fcls $\forall \varepsilon, \delta \Rightarrow m_H(\varepsilon, \delta) \geq \frac{VCd(H) + \log(1/\delta)}{\varepsilon} \leq m_H(\varepsilon, \delta)$
- so $L_D(A(S)) < \varepsilon$ in particular $\varepsilon = \frac{1}{2} - \delta$

- let $H_3 = \{ h_{a,b} : \mathbb{R} \to \{-1,+1\} \mid a,b \in \mathbb{R} \cup \{-\infty,+\infty\} \}$, $h_{a,b}(x) = \begin{cases} +1 & \text{if } x \in [a,b] \\ -1 & \end{cases}$
- H_3 is W.L. by a learner that works over a smaller class $H'_{DS} \subset H_3$
- H'_{DS} = class of decision stumps in 1-d $= \{ h_{a,b} \mid \mathbb{R} \to \{-1,+1\},\ a \in \mathbb{R},\ b \in \{-1,1\} \}$, $h_{a,b}(x) = \begin{cases} b & \text{for } x \le a \\ -b & \text{for } x > a \end{cases}$

- $\forall D, f$
  - $\exists h \in H'_{DS} \mid L_{D,f}(h) \le 1/3$
  - if $f \in H_3$ AND $f \in H'_{DS}$ then $L_{D,f}(h) = 0 < 1/3$
  - if $f \in H_3 \setminus H'_{DS}$



function does an error:
- $L_{D,f}(h_3) = P(x \in [a_f, b_f])$
- $L_{D,f}(h_2) = P(x > b_f)$
- $L_{D,f}(h_1) = P(x < a_f)$

$\nexists P(h) \le 1/3$ *

- H'_{DS} IS APAC WITH VCdim = 2
- we can take $m_H(\varepsilon, \delta) = \dfrac{2 + \ln(1/\delta)}{\varepsilon^2}$ * to have

$$L_{D,f}(A(S)) \le \min_{h \in H'_{DS}} L_{D,f}(h) + \varepsilon \overset{*}{\le} 1/3 + \varepsilon$$

($\min_{h \in H'_{DS}} L_{D,f}(h)$: error of the best h in H_3)

- we can say that this is a WL for H_3 just by choosing ε, δ such that $\boxed{1/3 + \varepsilon = 1/2 - \gamma}$ so we choose $\varepsilon = 1/12 = \gamma$ $\left(1/3 + 1/12 = 1/2 - 1/12\right)$

CONCLUSION
- ERM over H'_{DS} is a $1/12$-Weak Learner for H_3

WHY? → Solving $ERM_{H'_{DS}}$ IS EASIER than ERM_{H_3}; we can then transform it in a PAC-LEARNER for H_3 BY BOOSTING ALGO.

BOOSTING ALGO

- $ERM_{H^d_{DS}}$ CAN be implemented in Polynomial time
- $H^d_{DS} = \{ h_{a,b,i} : \mathbb{R}^d \to \{-1,+1\} \mid a \in \mathbb{R}, b \in \{-1,+1\} \}$
- $h_{a,b,i} = \begin{cases} b & \text{if } x_i < a \\ -b \end{cases}$
- Given $S = \{ (x_i, y_i) \mid i = 1, \dots, m \}$
- $ERM_{H^d_{DS}} = \underset{h \in H^d_{DS}}{\operatorname{argmin}} L_S(h) = \underset{a,b,i}{\operatorname{argmin}} L_S(h_{a,b,i})$ ← for a we take just m+1 value

- The complexity is:

COMPLEXITY

$$\underbrace{(m-1)}_{\text{useful a}} \cdot \underbrace{2}_{|b|} \cdot \underbrace{d}_{i} \cdot \underbrace{m}_{\text{to compute } L_S(h_{a,b,i})} + \underbrace{d\, m \log m}_{\text{ORDER ALL } d \text{ sets of point}}$$

we can update it instead of recalculate at every step

$$\Rightarrow md + md \log m$$

- QUA BOH, NON SO PERCHÉ MA:

$$\operatorname{argmin} \sum_{i=1}^{m} D_i \, \mathbb{1}(h_i(x_i, y_i))$$

// D_i IS A DISTRIBUTION OF WEIGHTS

ADA BOOST

- To learn over H it uses a γ-WL over H_B WL

$$\underset{h \in H_B}{\operatorname{argmin}} \sum_{i=1}^{m} D_i \, \ell(h_i(x_i, y_i))$$

$D_i^{(1)} = 1/m$ for $i = 1 \dots m$

For $(t = 1 \dots T)$

$h_t = WL(D^{(t)}, S)$

$\varepsilon_t = \sum_{i=1}^{m} D_i^{(t)} \, \ell(h_i, (x_i, y_i))$

$w_t = 2 \ln\left(\frac{1}{\varepsilon_t} - 1\right)$

$$D^{t+1} = \frac{\exp(-w_t y_i h_t(x_i)) D_i^t}{\sum_{j=1}^{m} \exp(-w_t y_j h_t(x_j)) D_j^t}$$

// POSITIVE PAPAYA HAVE POSITIVE NUMERATO, ELSE NEGATIVE (ENFORCE THE WEIGHTS!)

RETURN $h = \operatorname{sign}\left(\sum_{t=1}^{T} w_t h_t\right)$

TH

- if WL outputs h with $\sum D_i^t \, \ell(h_i(x_i, y_i)) \leq \frac{1}{2} - \gamma$ w.p. δ then ADABOOST outputs h_{ADA} with $L_S(h_{ADA}) \leq e^{-2\gamma^2 T}$ w.p $(1 - \delta T)$