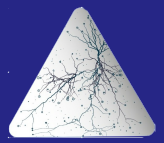


The Theory of Learning

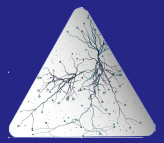
a special session

Hamza BA-MOHAMMED

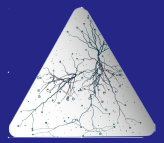
Former EAIC Deputy President, Former Training Head

ENSIAS AI Club

November 25, 2023

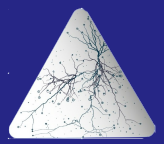


Summary



Questions

What is learning?
Could you measure learning? How so?



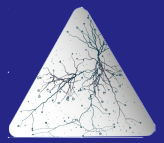
Terminology

domain set $\mathcal{X}$ (infinite)

label set $\mathcal{Y}$ (finite or infinite)

distribution $\mathcal{D}$ (unknown)

training data S (finite) where all elements are I.I.D. (independent and identically distributed)



Terminology

domain set $\mathcal{X}$ (infinite)

label set $\mathcal{Y}$ (finite or infinite)

distribution $\mathcal{D}$ (unknown)

training data S (finite) where all elements are I.I.D. (independent and identically distributed)

algorithm A

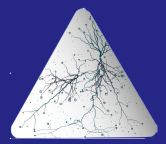
concept set $\mathcal{C}$ (infinite)

hypothesis h (output) also named classifier, predictor, or model

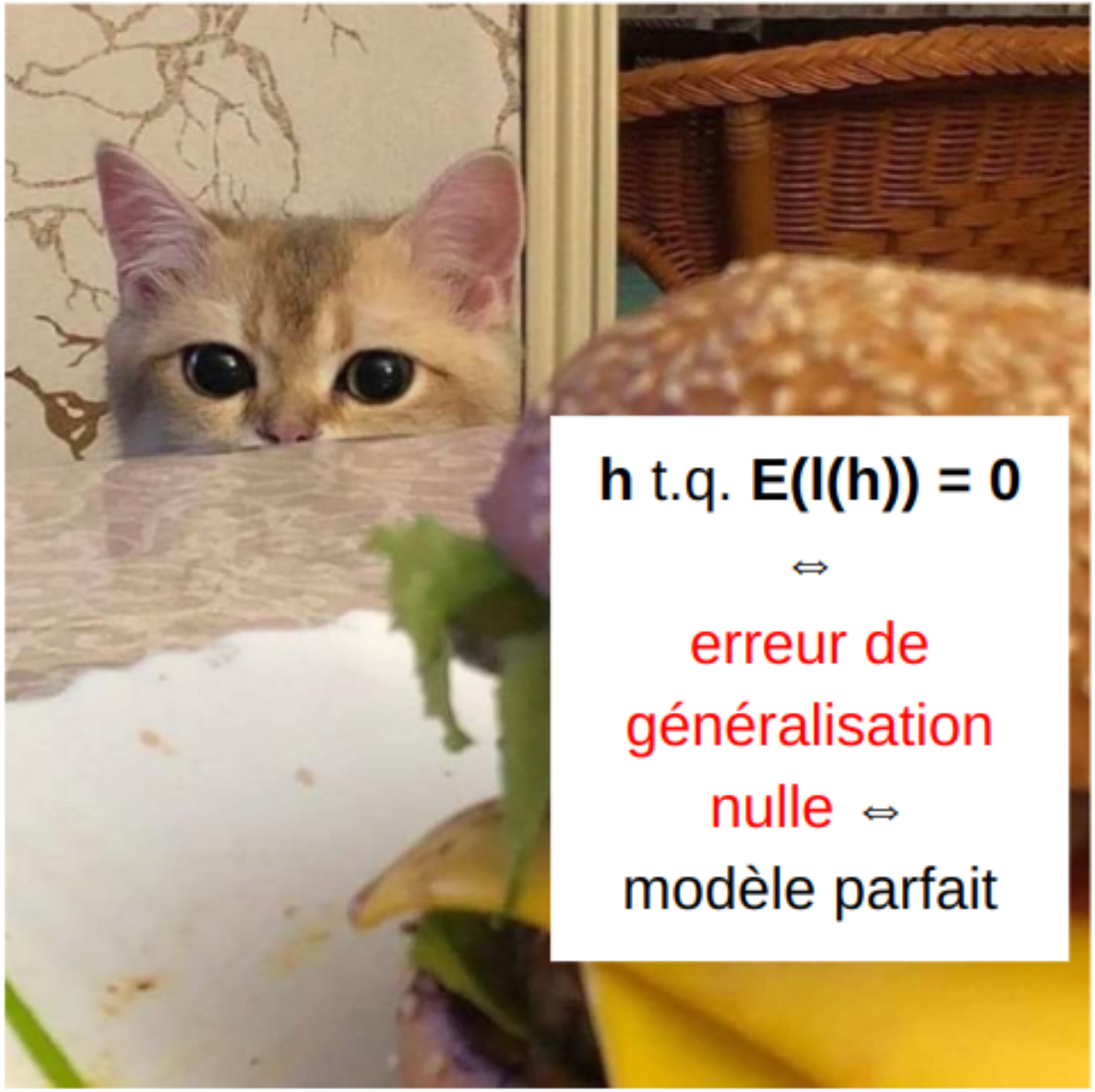
hypothesis class $\mathcal{H}$ (finite, infinite)

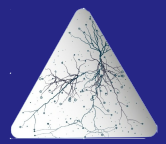
loss function $l(h(x), y)$ also named cost function

generalization error $err(h) = E(l(h)) = L_D(h)$



Objective



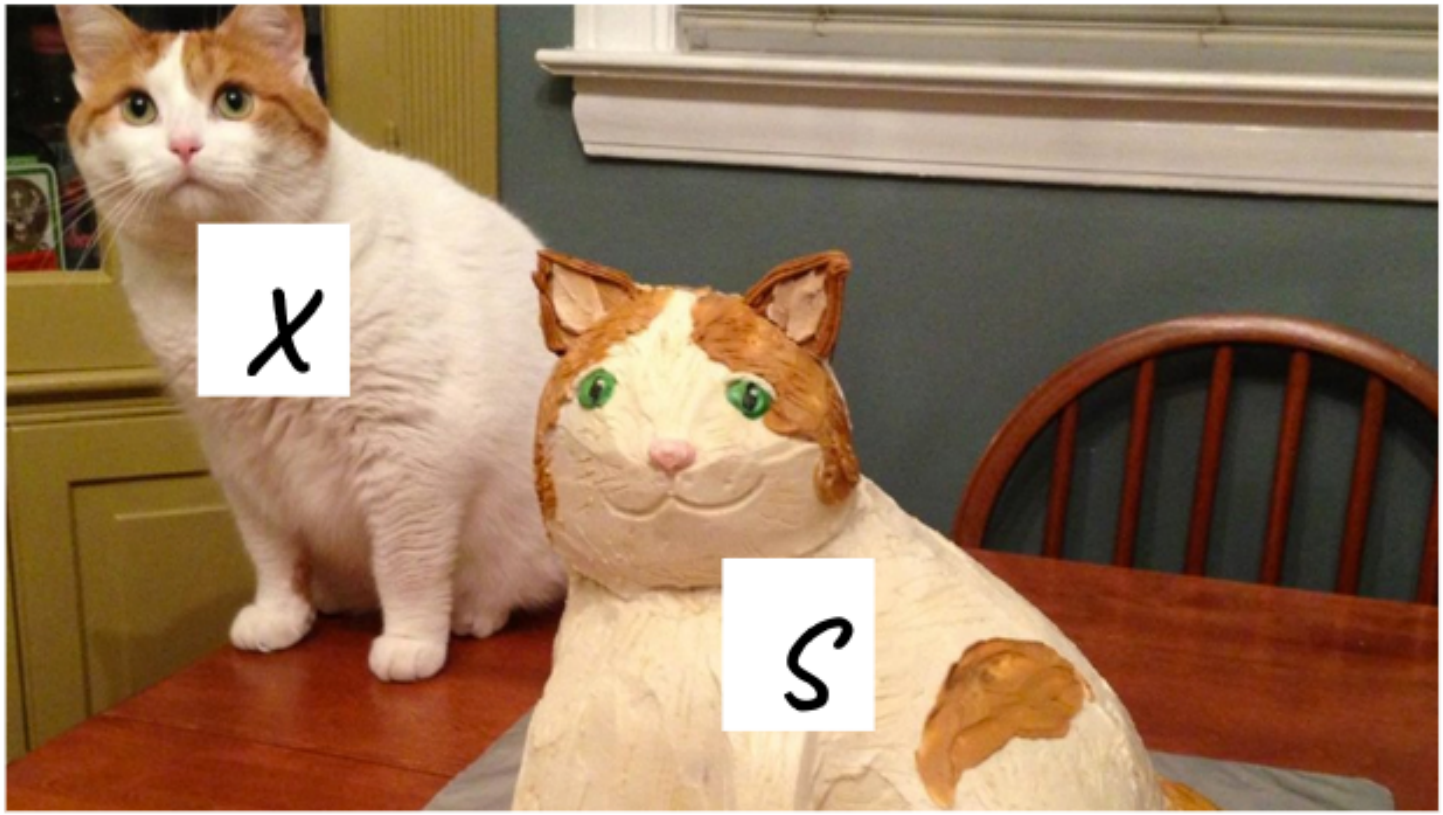


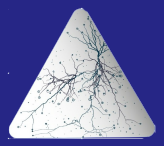
Challenges

S is a **finite** subpart of $\mathcal{X}$

S contains **stochastic noise** (measure imprecision)

$\Longrightarrow$ insufficient data for a perfect description of reality

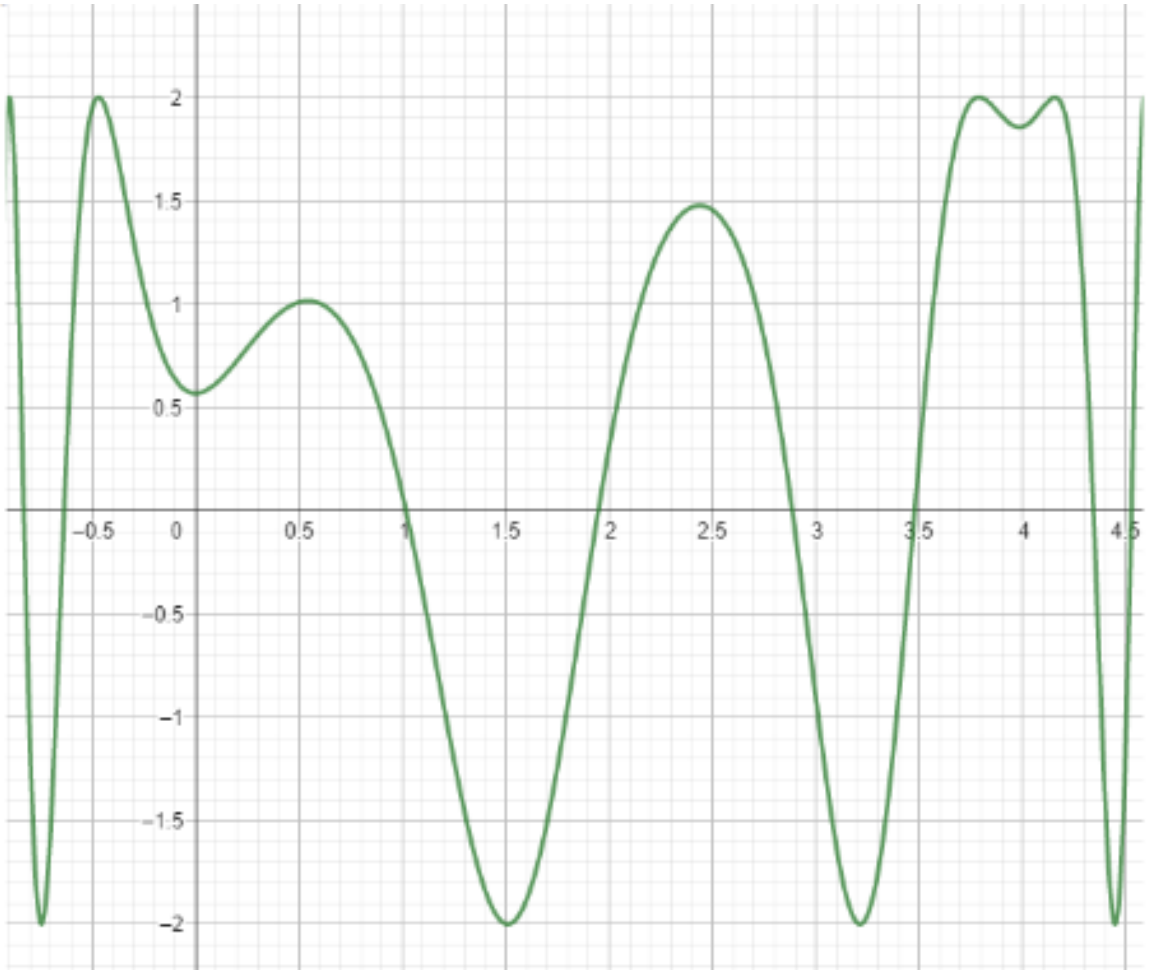




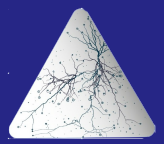
Challenges

Demo

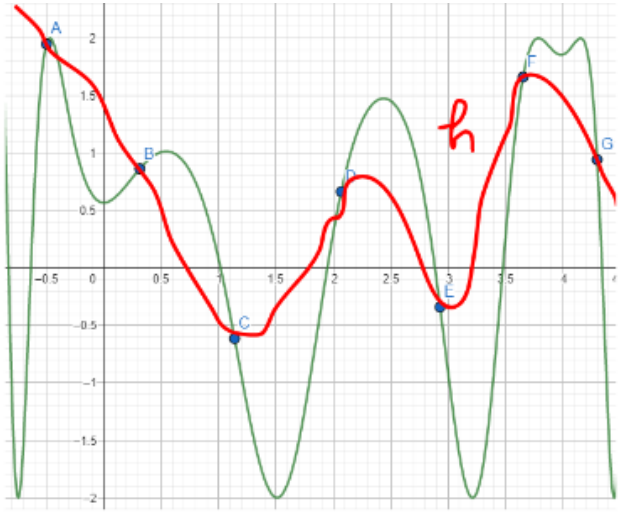
Suppose we have the above domain set $\mathcal{X}$:



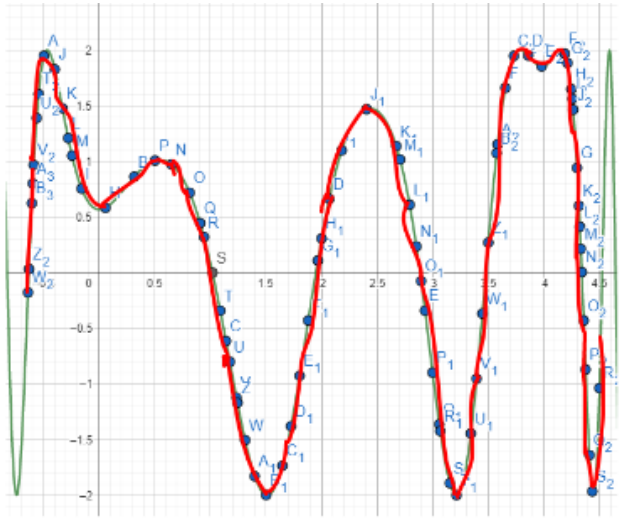
$$f(x) = 2\cos(5 + 3x^2 - 5.02x^3 + 2.01x^4 - 0.24x^5 + 0.0015x^6)$$

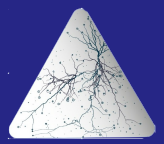


size of the sample S : the precision of our hypothesis is stronger the more we have data, but we can't keep on augmenting our data because:
a) material resources (CPU / GPU / RAM) are limited in their computation / storage capacity
b) the data itself is limited ($|S| < +\infty$)

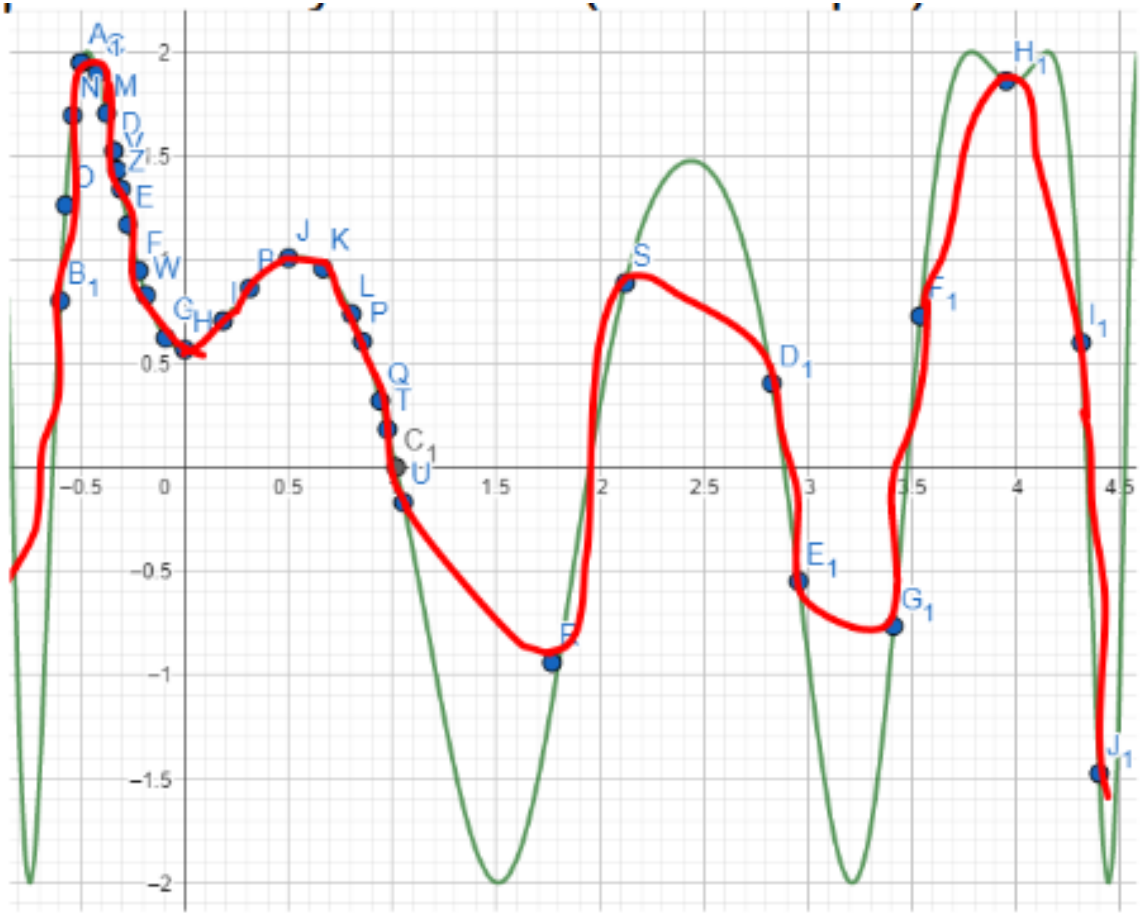


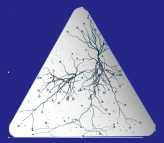
more data =>





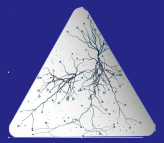
distribution of the data S: generally, we ignore the data distribution $\mathcal{D}$, so even a big sample could not be representative of the reality (bad sample)





Solution

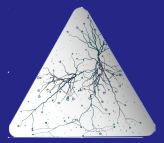
As in any scientific study, we must set a theoretical context that we call a hypothesis system, an axiomatic system, a policy, or a **framework**.
Since we can't find a perfect model, we should find a model which is **Approximately Correct**
$(err(h) \sim 0 \Longrightarrow err(h) < \epsilon)$
Let be $\epsilon \in [0, 1]$, we call it **the Precision**.
Since our error depends on the training sample, its value changes from one to another. And given that the model is imperfect, there exist at least a sample $S \in \mathcal{X}$ such that $err(h) > \epsilon$. We call it the **bad sample**.



Solution

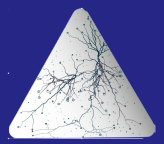
So, the event $A = \{h \text{ is approximately correct}\}$ is never certain ($P(A) \neq 1$). However, we look for maximizing this probability, e.i. minimizing the inverse probability of $B = \{h \text{ is incorrect}\}$ above a threshold δ, that we call **the confidence**
consequently, given ϵ and δ, we're looking for a concept set $\mathcal{C}$ where we can find a hypothesis h such that:

$$P(err(h) > \epsilon) < \delta$$



Solution

Despite that, we still need to limit a little bit the concept class which is so big for search, thus the relaxation of the situation using the **Hypothesis class H**.
So now, we are looking for an $h \in H$ such that
$\forall x \in S : h(x) = c(x)$ (with $c \in \mathcal{C}$, the real descriptor of the domain set $\mathcal{X}$)
Explicitly:
- in linear regression / single layer perceptron, $\mathcal{H}$ is the set of linear functions (ax + b)
- in polynomial regression of degree n, $\mathcal{H}$ is the set of polynomials of degree n
- in logistic regression, $\mathcal{H}$ is the set of Sigmoid functions, etc



Probably Approximately Correct Learning

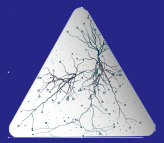
Finally, all we need to find out is the minimal size of the sample S that verifies the precision ϵ with a confidence level δ for an algorithm A and a hypothesis class $\mathcal{H}$, e.i. a function $m(\epsilon, \delta)$. We call it **the sample complexity of the algorithm** A.

Definition 1.1 ((realizable) PAC Learning). *A concept class $\mathcal{C}$ of target functions is PAC learnable (w.r.t to $\mathcal{H}$) if there exists an algorithm A and function $m_{\mathcal{C}}^{A} : (0,1)^2 \to \mathbb{N}$ with the following property:*

Assume $S = ((x_1, y_1), \ldots, (x_m, y_m))$ is a sample of IID examples generated by some arbitrary distribution D such that $y_i = h(x_i)$ for some $h \in C$ almost surely. If S is the input of A and $m > m_{\mathcal{C}}^{A}(\epsilon, \delta)$ then the algorithm returns a hypothesis $h_S^A \in \mathcal{H}$ such that, with probability $1 - \delta$ (over the choice of the $\mathbf{m}$ training examples):

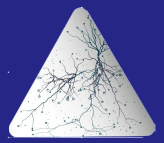
$$err(h_S^A) < \epsilon$$

The function $m_{\mathcal{C}}^{A}(\epsilon, \delta)$ is referred to as the sample complexity *of algorithm A.*



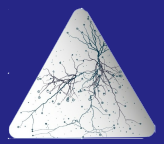
Realizability hypothesis

Something was silently assumed to be true.. and it's the fact that **there exists** a hypothesis $h \in H$ such that $err(h) = 0$, but is it always true?



Realizability hypothesis

Something was silently assumed to be true.. and it's the fact that **there exists** a hypothesis $h \in H$ such that $err(h) = 0$, but is it always true? **NO**!



PAC & APAC

This gives us 2 variants :

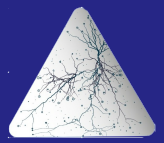
PAC learning: we admit the existence of such a hypothesis ($err(h^*) = 0$), thus we look for :

$$err(h_A^S) < \epsilon$$

Agnostic PAC learning: we can't tell if it exists or not, so we admit that there exists a hypothesis that minimize this error over the set of hypotheses $\mathcal{H}$, thus we look for:

$$err(h_A^S) < \min_{h^* \in \mathcal{H}} err(h^*) + \epsilon$$

Agnostic PAC learning can be considered a generalization of PAC learning, This means that if a learning algorithm is capable of solving the APAC problem, it is also capable of solving the PAC problem. (APAC $\Longrightarrow$ PAC)

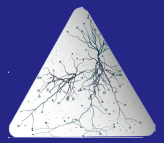


Empirical Loss

Recall: We've seen in the past sessions:

Empirical Loss (L_s): The "loss" is a quantified measure of how bad it is to get an error of a particular size/direction, which is affected by the negative consequences that accrue for inaccurate prediction. It is empirical because it's based on the data samples used in training.

General Loss (L_D): the True error related to the distribution of the domain set $\mathcal{X}$, it's the estimation of the empirical loss.



Common loss functions

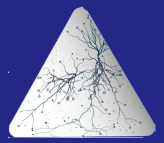
The general form of error:

Let l be a cost function, such that: $l: H \times Z \longrightarrow \mathbb{R}^+$ and $Z = X \times Y$

The general error of h: $L_{\mathcal{D}}(h) = \underset{z \sim \mathcal{D}}{\mathbf{E}}[l(h, z)]$

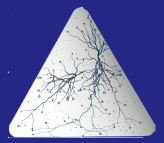
The empirical error of h: $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$

Classification	Regression
$l(h,z) = \begin{cases} 1 \; si \; h(x) \neq y \\ 0 \; si \; h(x) = y \end{cases}$ with: $z = (x,y) \in Z = X \times \{0,1\}$ This function is also valid for the multinomial classification.	$l(h,z) = (h(x) - y)^2$ with: $z = (x,y) \in Z = X \times \mathbb{R}^+$



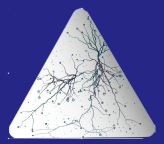
Empirical Risk Minimization

Since we don't have access to the data distribution, and we don't have access to the full domain set, the simplest approach is to work with the sample we've got, so the error evaluation is related to the sample, thus to "empirical" description.
The risk represents the error, and the objective is to minimize it, therefore the name **Empirical Risk Minimization**.



Question pour Champion!

what is the minimal sample size m we need so that the probability of the empirical loss exceeding the upper-bound ϵ is less than the confidence δ ?

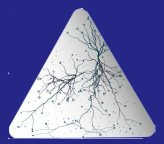


Finite Classes are ERM-learnable

Mathematicians have proven that we can use an ERM algorithm to learn in any finite hypothesis class, with sample complexity defined as:

PAC: $m(\epsilon, \delta) = \frac{1}{\epsilon} log(\frac{|\mathcal{H}|}{\delta})$

APAC: $m(\epsilon, \delta) = \frac{1}{\epsilon^2} log(\frac{|\mathcal{H}|}{\delta})$



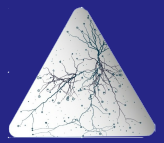
Finite Classes are ERM-learnable

Mathematicians have proven that we can use an ERM algorithm to learn in any finite hypothesis class, with sample complexity defined as:

PAC: $m(\epsilon, \delta) = \frac{1}{\epsilon} log(\frac{|\mathcal{H}|}{\delta})$

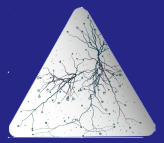
APAC: $m(\epsilon, \delta) = \frac{1}{\epsilon^2} log(\frac{|\mathcal{H}|}{\delta})$

But, again, we've implicitly supposed that such a complexity function exists, which is not always true!



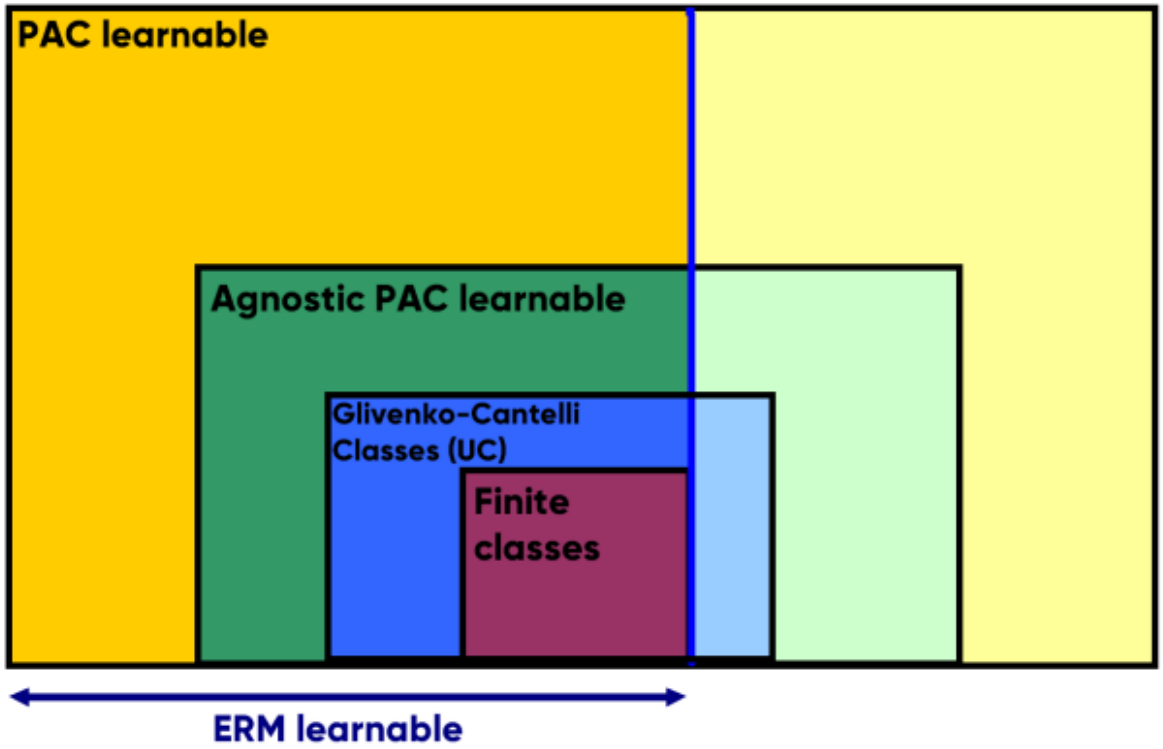
Uniform Convergent Learning

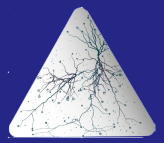
We consider the classes of hypothesis $\mathcal{H}$ for which this property is valid. We call them **Glivenko–Cantelli Classes**, and we have the following equivalences:
1) "A finite Hypothesis class has the uniform convergence property"
2) "if H has the UC property, then H is APAC-learnable"
3) "if H is finite and has the UC property, then H is PAC-learnable"
if $\mathcal{C}$ (or $\mathcal{H}$) is a Glivenko-Cantelli class, then with enough samples we can estimate uniformly the performance of all its hypothesis. In other words, empirical error and true error **converge uniformly** towards the same value over all the class's hypothesis.



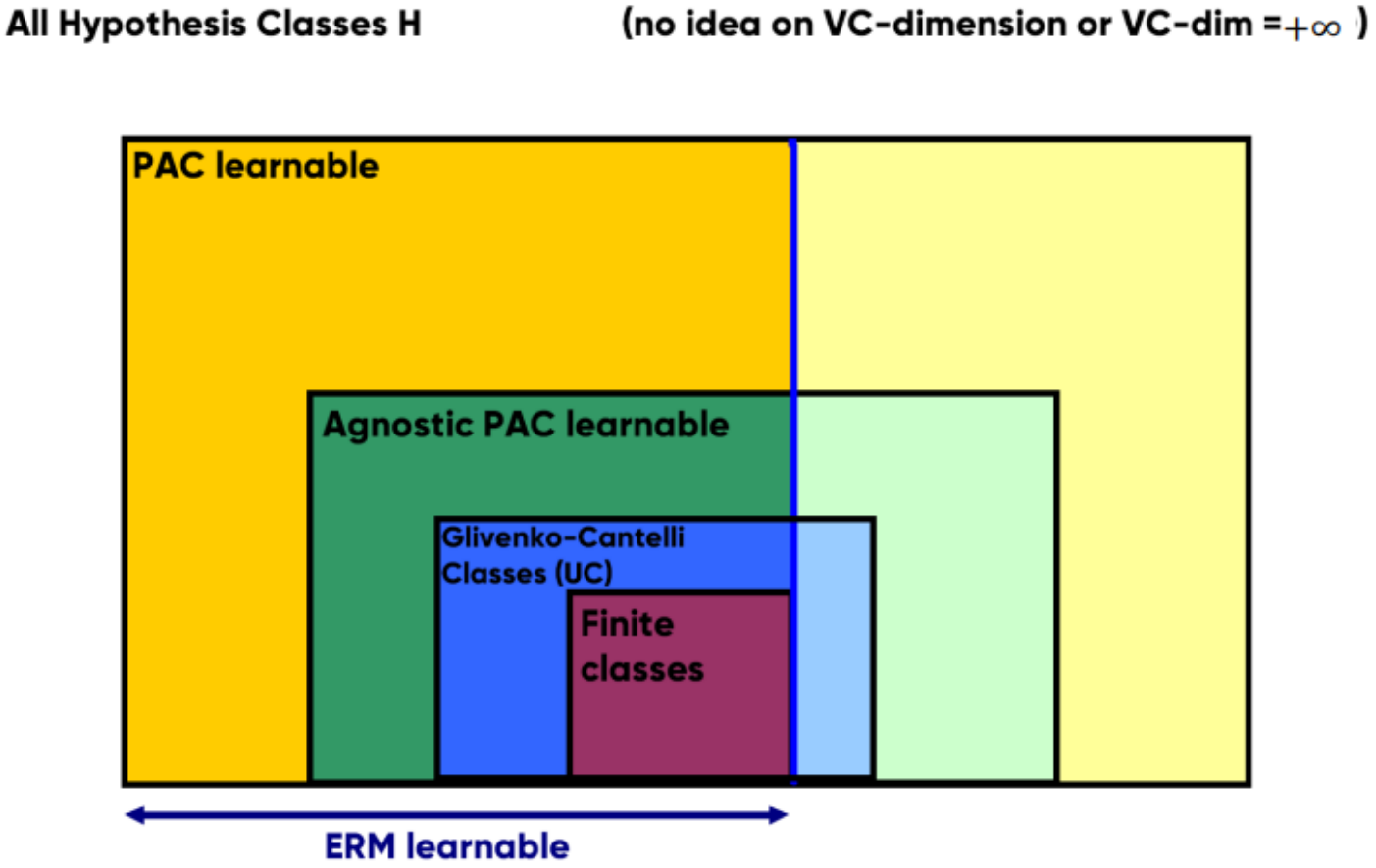
Overview

All Hypothesis Classes H (no idea on VC-dimension or VC-dim = $+\infty$)

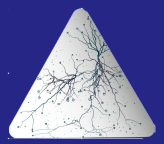




Overview



what if $|\mathcal{H}| = +\infty$?

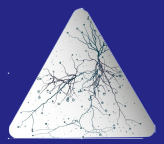


Growth Funtion $\Pi_{\mathcal{H}}$

We define the growth function as **the maximum number of ways $|S| = m$ points can be classified using $\mathcal{H}$.**

Theorem: Upper bound for growth function with finite hypothesis classes:

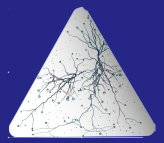
$$\Pi_{\mathcal{H}}(m) \leq 2^m$$



Shattering Sets

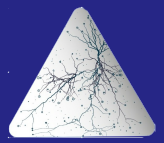
ay classe d'hypothèse 3ndha wa7ed propriété combinatoire smiytha VC dimension li indépendante 3la D o Y.
nmshiw b une approche constructive bash "nbniw" had la fonction
awalan, let H_s be the restriction of H to S shattered class : kanqolo anna A (subclass of X) is shattered by H_s ila kan b 2imkanna njem3o ay k elements f A $(0 < k \leq |A|)$ f 1 seul ensemble en utilisant un h dans Hs
d'ici découle la définition dyal VC-dimension:

$$VC\text{-}dim(\mathcal{H}) = \max\{|A| : \ A \ is \ shattered \ by \ \mathcal{H}\}$$



VC Dimension

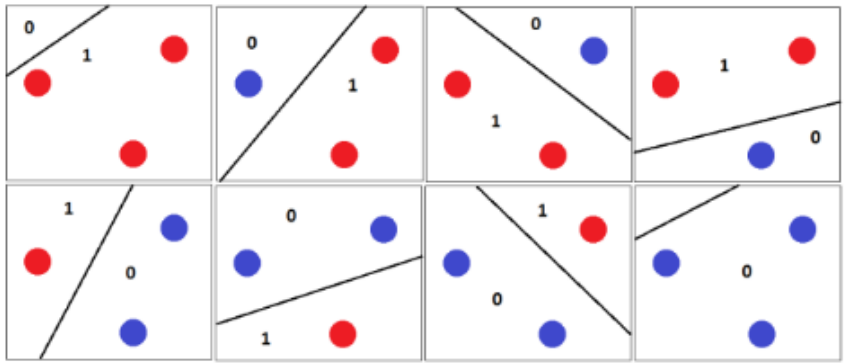
une autre formulation dyal shattering concpet:
ila kan 3ndna hypothesis class H o ensemble X, shno hiya la taille dyal le plus grand sous-ensemble A dans X li nqdro nqsmoh l 2 catégories en utilisant une hypothèse h dans H et **quelques soit labels dyal les éléments dyal A ?**

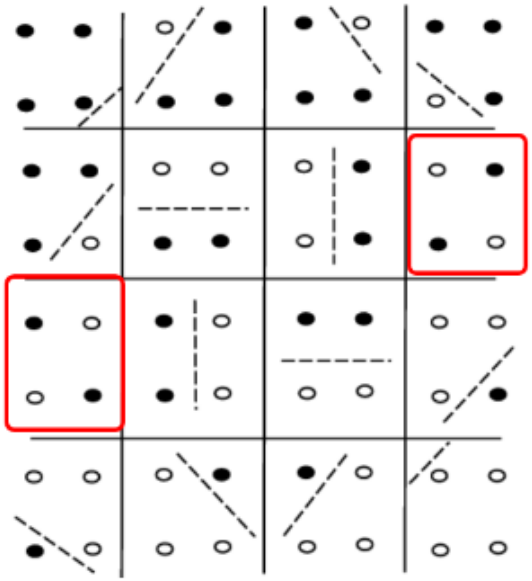


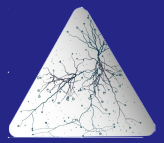
VC Dim: example

exemple 2:
soit H l'ensemble des lignes dans R^2, VC-dim(H) = 3

pour |A| = 3 ,pour |A| = 4







VC dimension

H shatters S if $\Pi_H(m) = 2^m$.

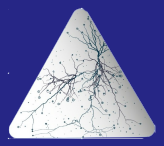
d'où:

$$VC(H) = \max\{m \mid \Pi_H(m) = 2^m\}$$

ou VC-dim(H) = +oo

PS1: if H is finite, then **VC-dim(H) <= log(|H|)**

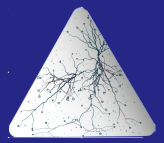
PS2: using Sauer-Shelah Lemma, if VC-dim(H) = d , then: $\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$



VC Dimension

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

(ps: R = Empirical Risk (erreur) et R^ = True Risk, et d = VC-dim(H))



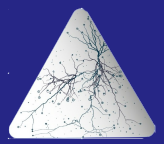
Fundamental Statistical Theorem of Learning

3.2.1 VC = ERM = Learnability

Theorem 3.5 (The Fundamental Theorem of Statistical Learning). *Let $\mathcal{C}$ be a concept class of functions from a domain χ to $\{-1, 1\}$, and let the loss function ℓ be the $0-1$ loss. Then the following are equivalent*

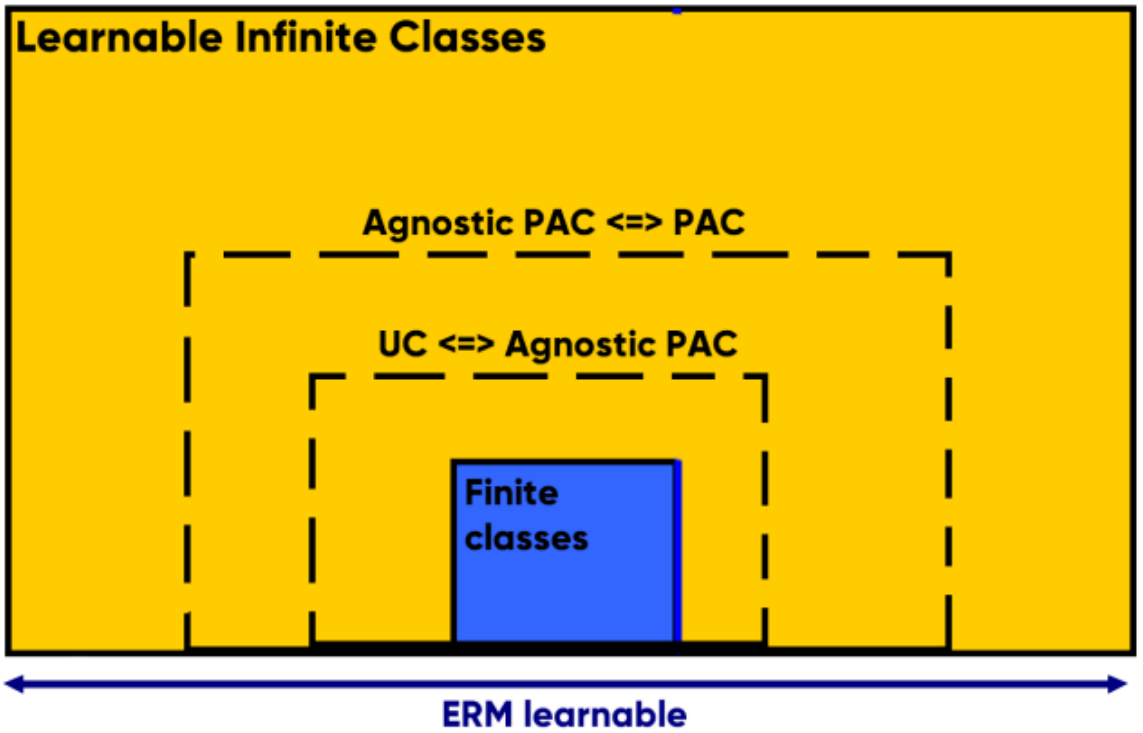
1. *$\mathcal{C}$ is (agnostic) PAC learnable.*
2. *$\mathcal{C}$ is (realizable) PAC learnable.*
3. *$\mathcal{C}$ has finite VC dimension.*
4. *$\mathcal{C}$ has the uniform convergence property*
5. *$\mathcal{C}$ is learnable by a $ERM_{\mathcal{C}}$ algorithm.*

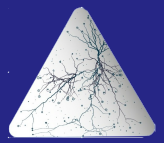
To summarize, the fundamental theorem states that that for the $0-1$ function the VC dimension completely characterizes the learnable classes, and as far as the PAC model goes, ERM algorithms are optimal.



Overview

All Hypothesis Classes H (when VC-dim < $+\infty$)





Further Notions

Bias Variance trade-off

Data Noise

Non Uniform Learning (NUL)

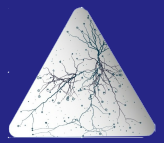
consistency learning

Ocam's Razor

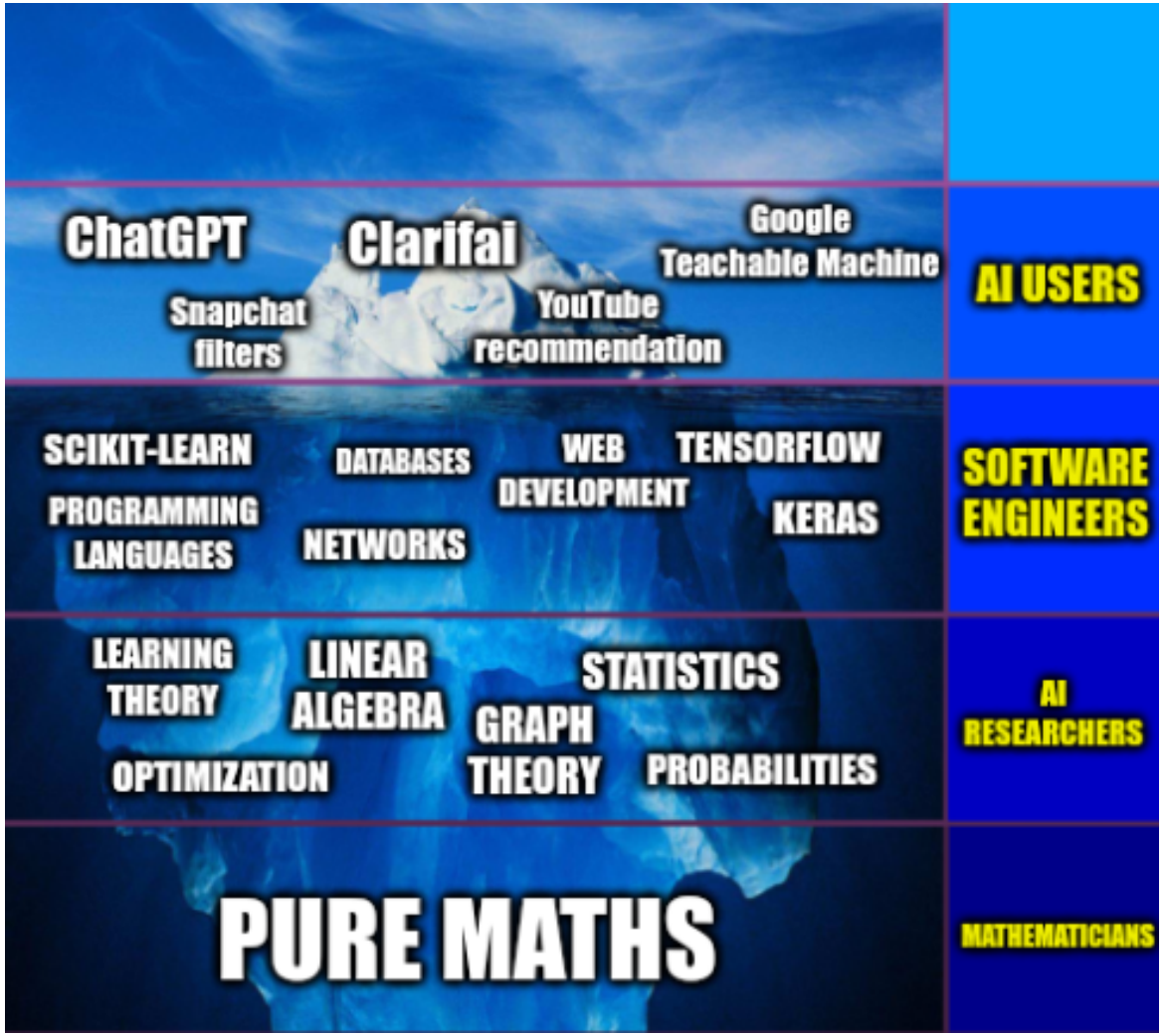
Minimal Description Length (MDL)

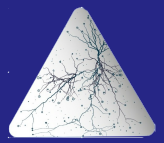
Structural Risk Minimization

...



Conclusion


ChatGPT
Clarifai
Google Teachable Machine
Snapchat filters
YouTube recommendation
AI USERS
SCIKIT-LEARN
DATABASES
WEB DEVELOPMENT
TENSORFLOW
PROGRAMMING LANGUAGES
NETWORKS
KERAS
SOFTWARE ENGINEERS
LEARNING THEORY
LINEAR ALGEBRA
STATISTICS
GRAPH THEORY
OPTIMIZATION
PROBABILITIES
AI RESEARCHERS
PURE MATHS
MATHEMATICIANS



The Theory of Learning

a special session

Hamza BA-MOHAMMED

Former EAIC Deputy President, Former Training Head

ENSIAS AI Club

November 25, 2023