# Machine Learning

## VC-Dimension

Fabio Vandin

December $11^{th}$, 2023

# PAC Learning

**Question**: which hypothesis classes $\mathcal{H}$ are PAC learnable?

**Up to now**: if $|\mathcal{H}| < +\infty \Rightarrow \mathcal{H}$ is PAC learnable.

What about $\mathcal{H}$: $|\mathcal{H}| = +\infty$? Not PAC learnable?

We focus on:

- *binary classification*:$\mathcal{Y} = \{0, 1\}$
- 0-1 loss

but similar results apply to other learning tasks and losses.

# Restrictions

## Definition (Restriction of $\mathcal{H}$ to $C$)

Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The restriction $\mathcal{H}_C$ of $\mathcal{H}$ to $C$ is:

$$\mathcal{H}_C = \{[h(c_1), \ldots, h(c_m)] : h \in \mathcal{H}\}$$

where we represent each function from $C$ to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

**Note**: $\mathcal{H}_C$ is the set of functions from $C$ to $\{0,1\}$ that can be derived from $\mathcal{H}$.

# VC-dimension and Shattering

## Definition (Shattering)

Given $C \subset \mathcal{X}$, $\mathcal{H}$ *shatters* $C$ if $\mathcal{H}_C$ contains all $2^{|C|}$ functions from $C$ to $\{0, 1\}$. se $|C| = m \Rightarrow 2^{|C|} = 2^m = |\mathcal{H}_C|$

## Definition (VC-dimension)

The *VC-dimension* $VCdim(\mathcal{H})$ of a hypothesis class $\mathcal{H}$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

**Notes:** ogni predizione possibile

- VC = Vapnik-Chervonenkis, that introduced it in 1971
- if $\mathcal{H}$ can shatter sets of arbitrarily large size then we say that $VCdim(\mathcal{H}) = +\infty$;
- if $|\mathcal{H}| < +\infty \Rightarrow VCdim(\mathcal{H}) \leq \log_2 |\mathcal{H}|$

$|C| = m \Rightarrow 2^m$ insiemi possibili $\Rightarrow 2^m \leq |\mathcal{H}|$

**Intuition**: the VC-dimension measures the *complexity* of $\mathcal{H}$ ($\approx$ how large a dataset that is perfectly classified using the functions in $\mathcal{H}$ can be)

# Example

$X$ ⇑

$H$ ⇐

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $h_1$ | 0 | 0 | **1** | 0 | 0 | 0 | **1** | 0 | 0 |
| $h_2$ | 0 | **1** | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| $h_3$ | **1** | 0 | 0 | 0 | **1** | **1** | 0 | 0 | 0 |
| $h_4$ | 0 | 0 | 0 | **1** | **1** | 0 | 0 | 0 | **1** |
| $h_5$ | 0 | 0 | **1** | 0 | 0 | 0 | 0 | **1** | 0 |
| $h_6$ | 0 | **1** | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| $h_7$ | **1** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| $h_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

VC dimension?

trovare $C \subset X$ + grande | $C$ frantumato da $H$ ⇒ tutte funzioni possibili ≡ tutti vettori possibili

$VCdim(H) \geq 1 \Rightarrow C = \{x_3\} \Rightarrow H_C = \{[0],[1]\}$

simile per $\geq 2$ ($\{x_5, x_6\}$) ⇒ non può essere 3: serve colonna con almeno $\neq 1$, non c'è

# Note

To show that $VCdim(\mathcal{H}) = d$ we need to show that:

1. $VCdim(\mathcal{H}) \geq d$
2. $VCdim(\mathcal{H}) \leq d$

that translates to

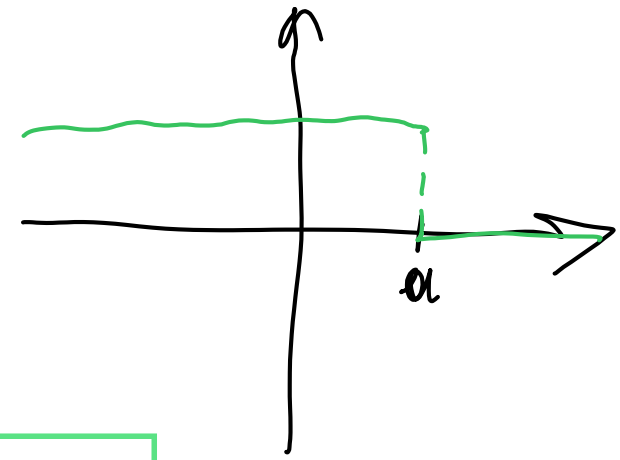
1. there exists a set $C$ of size $d$ which is shattered by $\mathcal{H}$
2. *every* set of size $d + 1$ is not shattered by $\mathcal{H}$

**Question**: why don't we need to consider sets of size $> d + 1$?

se potessi, avrei C di grandezza d+1 da poter frantumare, ma non ce l'ho

# Example: Threshold Functions

$$\mathcal{H} = \{h_a : a \in \mathbb{R}\}$$



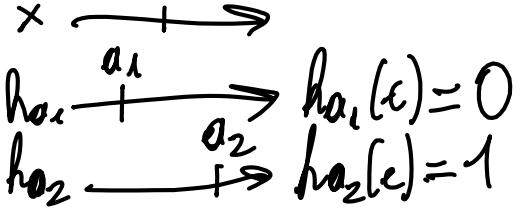

$|\mathcal{H}| = +\infty$

where $h_a : \mathbb{R} \to \{0, 1\}$ is

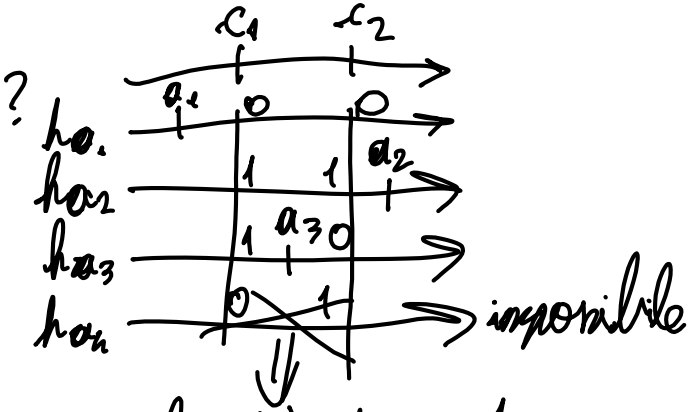
$$h_a(x) = \mathbb{1}[x < a] = \begin{cases} 1 & \text{if } x < a \\ 0 & \text{if } x \geq a \end{cases}$$

VC-dimension?

$VCdim(\mathcal{H}) = 1$? $\Rightarrow$

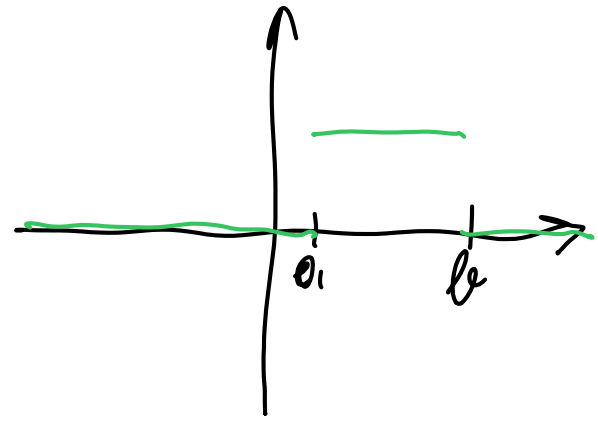



$VCdim(\mathcal{H}) = 2$?




$VCdim(\mathcal{H}) \leq 1 \Rightarrow = 1$

# Example: Intervals



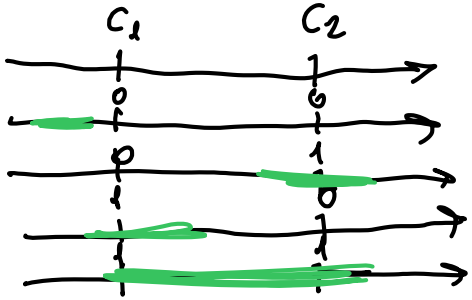

$$\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$$

where $h_{a,b} : \mathbb{R} \to \{0, 1\}$ is

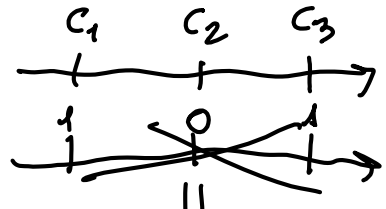
$$h_{a,b}(x) = \mathbb{1}[x \in (a, b)] = \begin{cases} 1 & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

VC-dimension?

$VCdim(\mathcal{H}) \geq 2$




$VCdim(\mathcal{H}) \leq 2$:




impossibile

$\Downarrow$

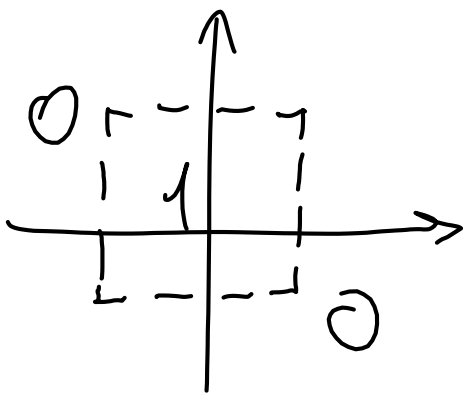
$VCdim(\mathcal{H}) = 2$

# Example: Axis Aligned Rectangles

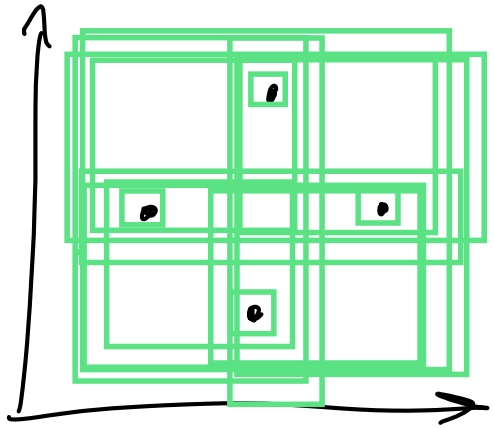
$$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1, a_2, b_1, b_2 \in \mathbb{R}, a_1 \leq a_2, b_1 \leq b_2\}$$

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 < a_2, b_1 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

VC-dimension?



$VCdim(\mathcal{H}) \geq 4$



$VCdim(\mathcal{H}) \leq 4$

il + alto

il + a sx

il + a dx

impossibile

$\Downarrow$

$VCdim(\mathcal{H}) = 4$

il + basso

# Example: Convex Sets

Model set $\mathcal{H}$ such that for $h_S \in \mathcal{H}$, $h_S : \mathbb{R}^2 \to \{0, 1\}$ with

$$h_S(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in S \\ 0 & \text{otherwise} \end{cases}$$

where $S$ is a convex subset of $\mathbb{R}^2$

VC-dimension?

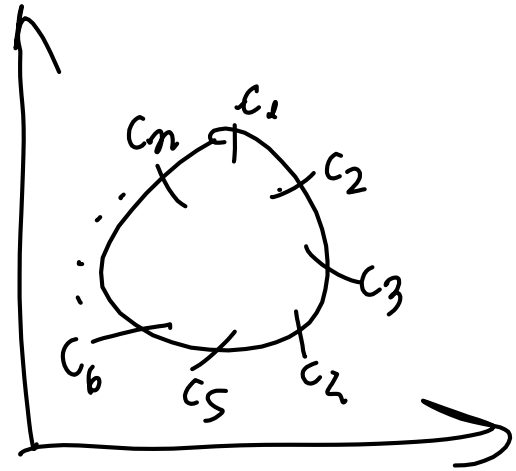
Prendiamo valore arbitrario di $n \in \mathbb{N}$ (dim. di insieme da frantumare)

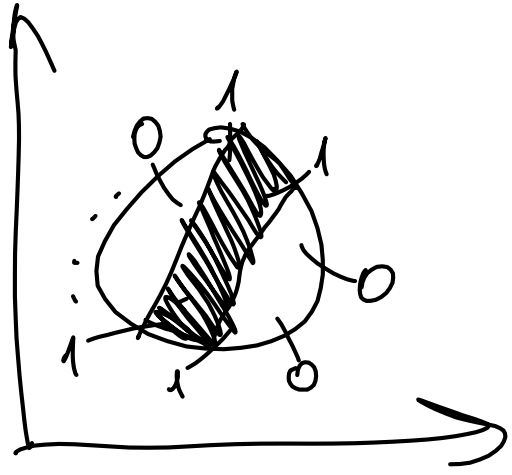
Poi, labeling arbitrario

$\Downarrow$

ipotesi corrispondente a insieme convesso

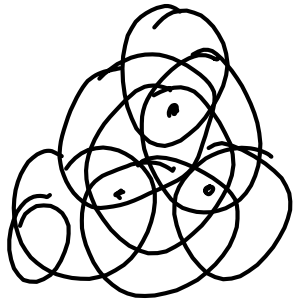
dato dai $(\vec{x}_i, y_i)$ con $y_i = 1$ dà labeling desiderato $\Rightarrow \mathcal{H}$ può frantumare set di $n$ punti $\forall n \Rightarrow VC.dim(\mathcal{H}) = +\infty$




soluzione: 3

i) VCdim $\geq 3$ $\Rightarrow$



posso coprire ogni funzione possibile

## Exercise

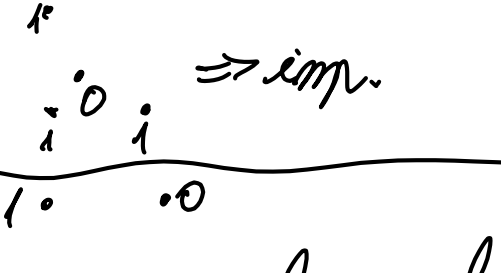
Consider the classification problem with $\mathcal{X} = \mathbb{R}^2, \mathbb{Y} = \{0, 1\}$.
Consider the hypothesis class $\mathcal{H} = \{h_{(\mathbf{c},a)}, \mathbf{c} \in \mathbb{R}^2, a \in \mathbb{R}\}$ with

$$h_{(\mathbf{c},a)}(\mathbf{x}) = \begin{cases} 1 & \text{if } ||\mathbf{x} - \mathbf{c}|| \leq a \\ 0 & \text{otherwise} \end{cases}$$

Find the VC-dimension of $\mathcal{H}$.

i) VCdim $\leq 3$ $\Rightarrow$ 4 istanze $\Rightarrow$ 3 casi:

- 3 istanze fanno triangolo e 4ª è dentro




- // // // // fuori

- stessa linea

# The Fundamental Theorems of Statistical Learning

**Theorem**

*Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and consider the 0-1 loss function. Assume that $VCdim(\mathcal{H}) = d < +\infty$. Then there are absolute constants $C_1, C_2$ such that*

- *$\mathcal{H}$ has the uniform convergence property with sample complexity*

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

- *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

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

Equivalently:

**Theorem**

*Let $\mathcal{H}$ be an hypothesis class with VC-dimension $VCdim(\mathcal{H}) < +\infty$. Then, with probability $\geq 1 - \delta$ (over $S \sim \mathcal{D}^m$) we have:*

$$\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \leq L_{\mathcal{S}}(h) + C\sqrt{\frac{VCdim(\mathcal{H}) + \log(1/\delta)}{2m}}$$

*where $C$ is a universal constant.*

**Note**: finding $h \in \mathcal{H}$ that minimizes the upper bound (above) to $L_{\mathcal{D}}(h)$ $\Rightarrow$ ERM rule

**Theorem**

*Let $\mathcal{H}$ be a class with $VCdim(\mathcal{H}) = +\infty$. Then $\mathcal{H}$ is not PAC learnable.*

**Notes:**

- the VC-dimension *characterizes* PAC learnable hypothesis classes

## Exercise

Let

$$\mathcal{H}_d = \{h_{\mathbf{w}}(\mathbf{x}) : h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(< \mathbf{w}, \mathbf{x} >)\}$$

where $\mathcal{X} = \mathbb{R}^d$.

Prove that $VCdim(\mathcal{H}_d) = d$.

# An Interesting Example...

**Note:** in previous examples the VC-dimension is equivalent to the number of parameters that define the model... but it is not always the case!

Function of one parameter: $f_\theta(x) = \sin^2\left[2^{8x} \arcsin\sqrt{\theta}\right]$

VC-dimension of $f_\theta(x)$ is infinite!

In fact, $f_\theta(x)$ can approximate any function $\mathbb{R} \to \mathbb{R}$ by changing the value of $\theta$!

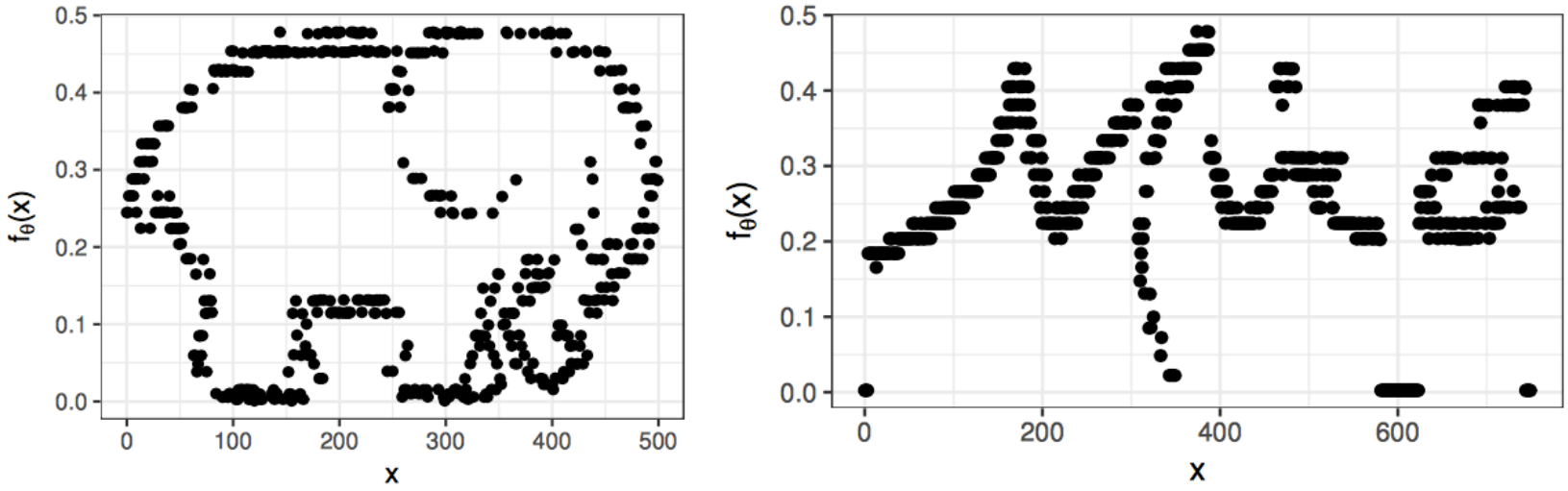



FIG. 1: A scatter plot of $f_\theta$ for $\theta = 0.2446847266734745458227540656\cdots$ plotted at integer $x$ values, showing that a single parameter can fit an elephant (left). The same model run with parameter $\theta = 0.00242654180550004019353876 20\cdots$ showing a fit of a scatter plot to Joan Miró's signature (right). Both use $r = 8$ and require hundreds to thousands of digits of precision in $\theta$.

["One parameter is always enough", Piantadosi, 2018]

# Bibliography

[UML] Chapter 6