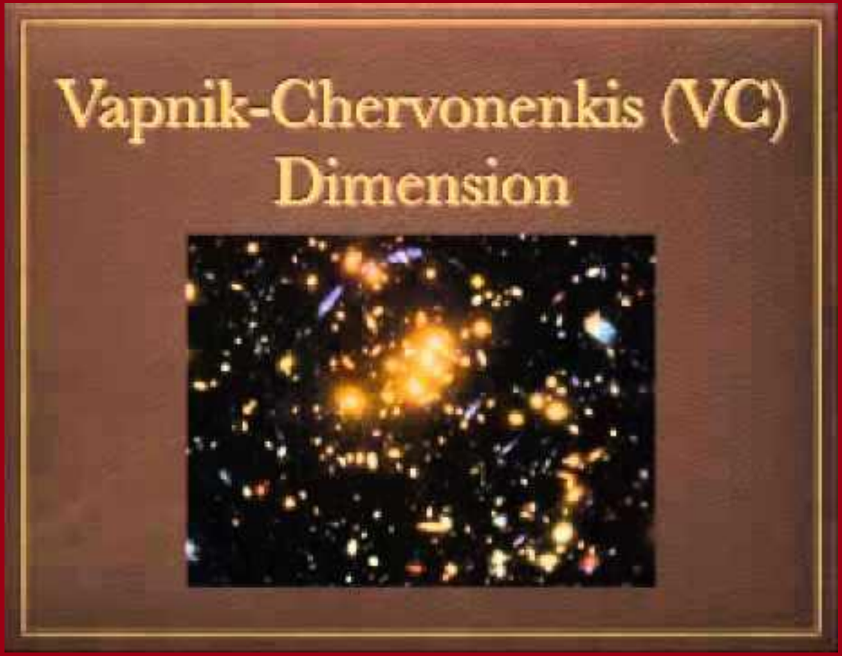
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA




# VC Dimension

Machine Learning 2022-23
UML Book Chapter 6
Slides: F. Chiariotti, P. Zanuttigh, F. Vandin

# Which hypothesis classes are PAC learnable ?

Simplification: focus on binary classification and 0-1 loss

1. Theorem (*uniform convergence*): finite classes are agnostic PAC learnable
2. Theorem (*corollary of NFL*): The set of all functions from an infinite domain set to {0,1} is not PAC learnable

- Up to now, if $|\mathcal{H}| < \infty \Rightarrow \mathcal{H}$ is PAC learnable (finite size classes are agnostic PAC learnable)
- What about infinite size classes ($|\mathcal{H}| = \infty$) ?

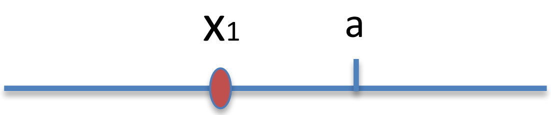
→ We'll demonstrate that the finite size is a sufficient but not necessary condition


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

$h_a : \mathbb{R} \rightarrow \{0,1\}$

$$h_A(x) = \begin{cases} 1 \; if \; x < a \\ 0 \; if \; x \geq a \end{cases}$$

*Example: threshold function* →it is PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) = \left\lceil \frac{\log\left(\frac{2}{\delta}\right)}{\epsilon} \right\rceil$



# Restriction of a Function

*Definition: Restriction of $\mathcal{H}$ to $\mathcal{C}$*

- Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$
- Let $C = \{c_1, \dots, c_m\} \subset \mathcal{X}$

The restriction $\mathcal{H}_c$ of $\mathcal{H}$ to $\mathcal{C}$ is the set of functions from $\mathcal{C}$ to *{0,1}* that can be derived from $\mathcal{H}$ :

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

Each entry: A vector of 0s and 1s of length $m$ with the output for each $c_i$

Notes:

- We can represent each function from $\mathcal{C}$ to *{0,1}* as a vector in $\{0,1\}^{|C|}$
- *No Free Lunch theorem*: the idea is to select a distribution concentrated on a set C (→restriction) on which the algorithm A fails

# Shattering

**Definition (Shattering)**

Given $C \subset X, \mathcal{H}$ shatters C if $\mathcal{H}_c$ contains all the $2^{|C|}$ functions from C to {0,1}

**Corollary (of No Free Lunch)**

Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to {0,1}. Let $m$ be a training set size. Assume that there exist a set $C \subset \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$. Then for any learning algorithm A there exist a distribution $D$ over $\mathcal{X}$ $x$ $\{0,1\}$ and a predictor $h \in \mathcal{H}$ such that $L_d(h) = 0$ but with probability at least 1/7 over the choice of S we have that $L_D(A(S)) \geq \frac{1}{8}$

*Demonstration (intuition): on set C all functions from C to {0,1} can be chosen and we fall back into the situation of the NFL corollary*

# VC Dimension (1)

**Definition (VC-dimension)**

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

*Note*: if $\mathcal{H}$ can shatter sets of arbitrarily large size then we say that $VCdim(\mathcal{H}) = +\infty$

# VC Dimension (2)

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

- *In the case of finite class hypotheses:*
  1. They are agnostic PAC learnable (already demonstrated)
  2. To shatter a set of size $|C|$ → at least $2^{|C|}$ functions (need all combinations)
  3. With $|\mathcal{H}|$ functions → the largest set that can be shattered has size $\log_2|\mathcal{H}|$
  4. To have $VCdim(\mathcal{H}) = d \Longrightarrow$ shatter a set of size d $\Longrightarrow$ $VCdim(\mathcal{H}) \leq \log_2|\mathcal{H}|$

- *If $\mathcal{H}$ has an infinite VC dimension: it is not PAC learnable*
  1. $VCdim(\mathcal{H}) = \infty \Longrightarrow \forall m$: ∃ a shattered set of size 2m (can shatter any size)
  2. Apply NFL corollary: $\exists D$ on which A does not work (for any possible A)
  3. $\exists D$ with probability $\geq \frac{1}{7}$ $L_D \geq \frac{1}{8} \Longrightarrow$ it is not PAC learnable (for ∀ A)

# Compute VC Dimension

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

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

1. $VCdim(\mathcal{H}) \geq d$ : there exists a set C of size *d* which is shattered by $\mathcal{H}$

2. $VCdim(\mathcal{H}) < (d + 1)$ : every set of size *d + 1* is not shattered by $\mathcal{H}$

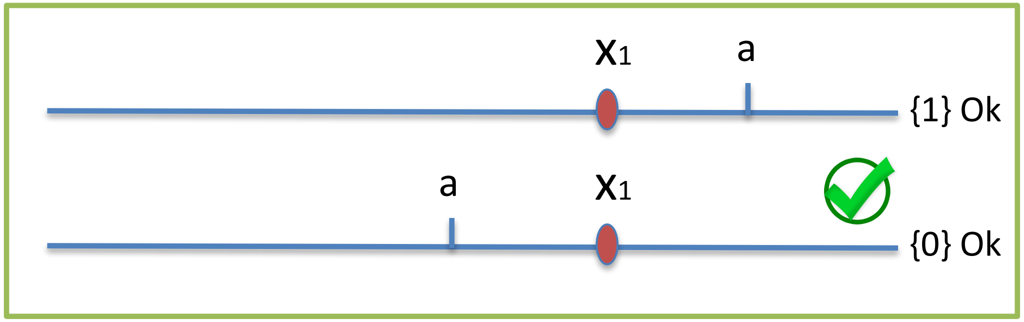
# Compute VC Dimension: Example (1)

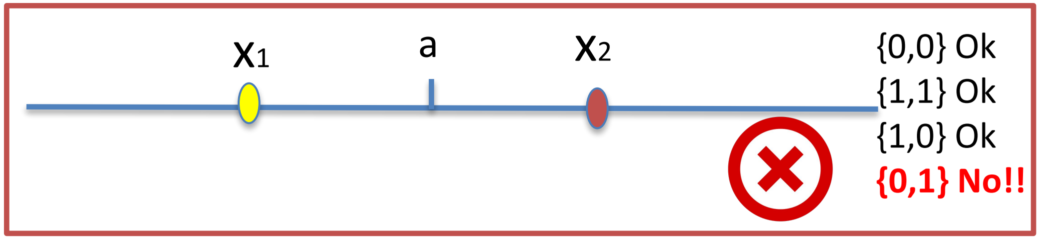
**Threshold function**

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

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

$$h_A(x) = \begin{cases} 1 \; if \; x < a \\ 0 \; if \; x \geq a \end{cases}$$




$VCdim(\mathcal{H}) \geq 1$

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

$\Downarrow$

$\boldsymbol{VCdim(\mathcal{H}) = 1}$

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

# Compute VC Dimension: Example (2)

**Interval**

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

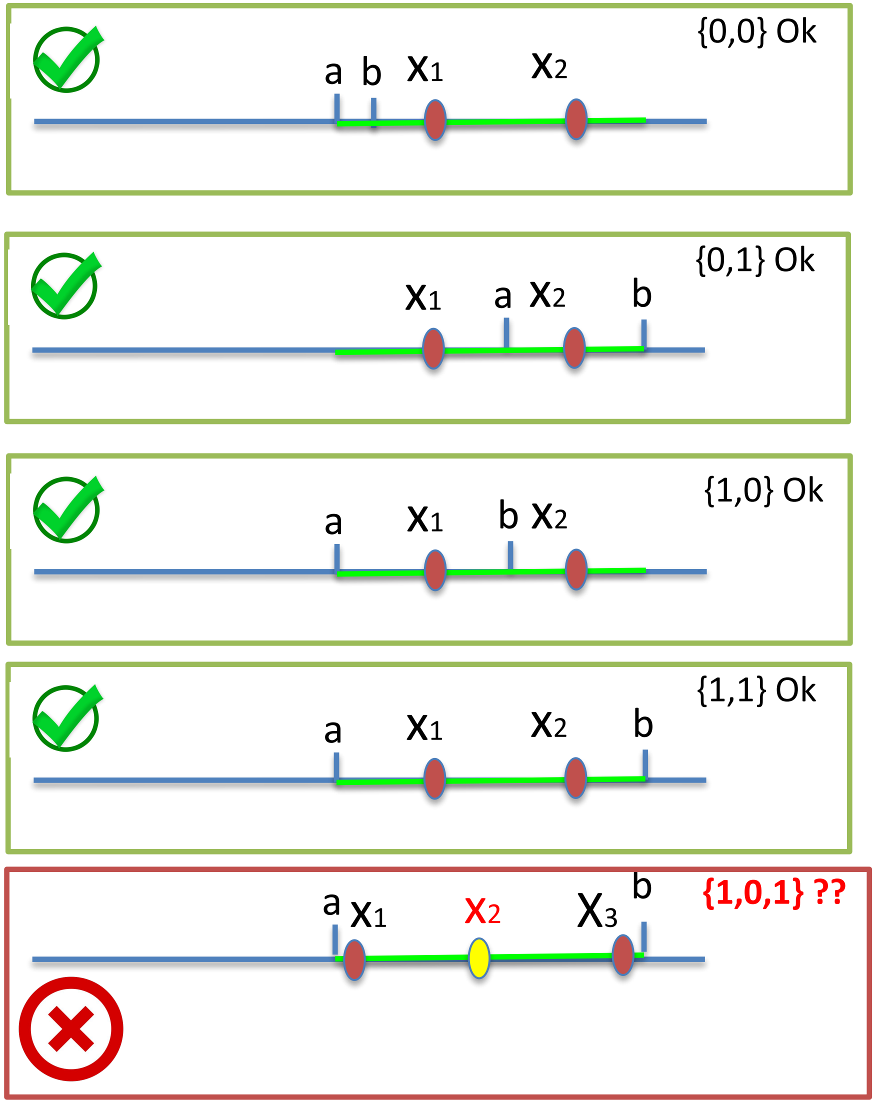
$h_{a,b}: \mathbb{R} \to \{0,1\}$ is:

$$h_{a,b}(x) = \begin{cases} 1\ if\ a < x < b \\ 0 \quad otherwise \end{cases}$$

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

$$VCdim(\mathcal{H}) < 3$$

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



DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

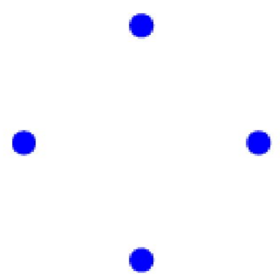
# Compute VC Dimension: Example (3)

Axis aligned rectangle

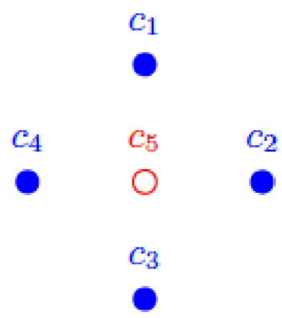
$\mathcal{H} = \{h_{a_1,a_2,b_1,b_2} : a_1, a_2, b_1, b_2 \in \mathbb{R}\ , a_1 \le a_2, b_1 \le b_2\ \}$

$h_{a_1,a_2,b_1,b_2} : \mathbb{R} \to \{0,1\}$ is:

$$h_{a_1,a_2,b_1,b_2}(x_1, x_2) = \begin{cases} 1 \ if \ a_1 \le x_1 \le a_2, b_1 \le \ x_2 \le b_2 \\ 0 \quad otherwise \end{cases}$$



The set can be shattered




{1,1,1,1,0} No!!

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

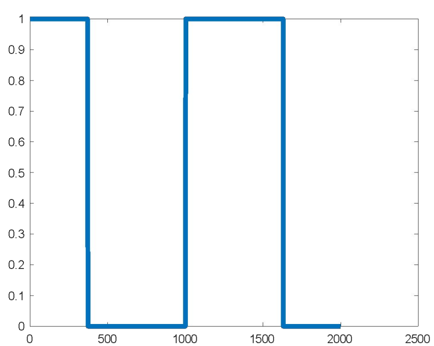
$VCdim(\mathcal{H}) < 5$

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

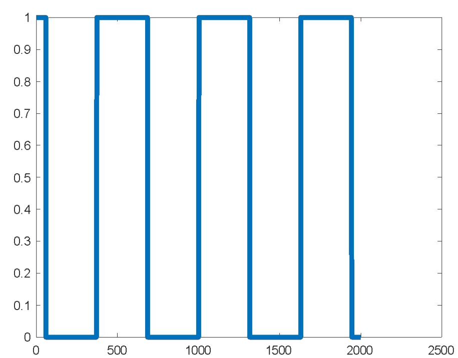
Case 5 points: define $c_1$ top point, $c_2$ rightmost, $c_3$ bottom, $c_4$ leftmost, $c_5$ the reamaining one. If different labeling just swaps the case that can not be obtained.

# Compute VC Dimension: Example (4)

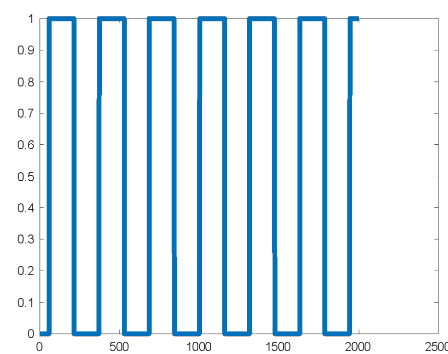
- *Recall*: for finite classes: $VCdim(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$ ...
- ... but VC dimension does not always correspond to the number of parameters !!
- Example $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$, $h_\theta : \mathcal{X} \rightarrow \{0,1\}$ $h_\theta = \lceil 0.5\sin(\theta x) \rceil$
  - It has infinite VC dimension !!




$\theta = 0.5$




$\theta = 1$




$\theta = 2$

# Fundamental Theorem of Statistical Learning

Let H be a hypothesis class $h:X \rightarrow [0,1]$, and let the loss function be the 0-1 loss.

The following statements are equivalent:

1. H has the **uniform convergence** property
2. Any ERM rule is a successful **agnostic PAC learner** for H
3. H is **agnostic PAC learnable**
4. H is **PAC learnable**
5. Any ERM rule is a **successful PAC learner** for H
6. H has a **finite VC dimension**

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

# Theorem of Statistical Learning: Notes on the demonstration

1. We have already seen that 1 → 2 → 3 (uniform convergence implies agnostic PAC learnable and ERM rule is PAC learner)
2. 3 → 4 is trivial (if realizable they are the same if not PAC condition does not apply)
3. 2 → 5 also trivial (ERM rule, if realizable same target)
4. 4 → 6 and 5 → 6 follow from corollary of No-Free-Lunch (by contradiction, infinite VC is not PAC learnable)
5. The key part is how to close the loop (6 → 1, from finite VC dimension to uniform convergence)

*The proof 6 → 1 (not part of the course) can be divided in two main parts:*

- *If VCdim($\mathcal{H}$) = d, then even though $|\mathcal{H}|$ might be infinite, when restricting $\mathcal{H}$ to a finite set C, its "effective size" $|\mathcal{H}_C|$, is only $O(|C|^d)$. That is, $|\mathcal{H}_C|$ grows polynomially rather than exponentially with $|C|$ (Sauer's lemma)*
- *Recall that finite hypothesis classes enjoy the uniform convergence property. This result can be generalized by showing that uniform convergence holds whenever the hypothesis class has a "small effective size" (i.e., classes for which $|\mathcal{H}_C|$ grows polynomially with $|C|$)*

# Theorem of Statistical Learning: Quantitative Version

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

1. $\mathcal{H}$ has the uniform convergence property with sample complexity

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

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity

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

3. $\mathcal{H}$ is PAC learnable with sample complexity

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

$$\log\left(\frac{2|\mathcal{H}|}{\delta}\right) = \log 2|\mathcal{H}| + \log\frac{1}{\delta}$$

Not part of the course, just notice how the number of samples depend on $VCdim(\mathcal{H}) = d$

**Recall:**
*(notice the role of $|\mathcal{H}|$ and of the VC dimension d)*

**Proposition**

Let $\mathcal{H}$ be a finite hypothesis class, let $Z$ be a domain, and let $\ell : \mathcal{H} \times Z \rightarrow [0, 1]$ be a loss function. Then:

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

$$m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil$$