Machine Learning

VC-Dimension

Fabio Vandin

December 15^{th}, 2023

Restrictions

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

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 $\mathcal{H}$.

$$1 \leq |\mathcal{H}_C| \leq 2^m$$

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 $\{0, 1\}$.

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:

- 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 $|\mathcal{H}| < +\infty \Rightarrow VCdim(\mathcal{H}) \leq \log_2 |\mathcal{H}|$

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

$h_1(\vec{x}_4)=0$

$\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_9\}$

$\mathcal{H} = \{h_1, h_2, \dots, h_8\}$

	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

$h_4(\vec{x}_9)=1$

We need to find the "largest" set $C \subseteq \mathcal{X}$ s.t. C is shattered by $\mathcal{H}$

VC dimension?

Is the VC-dim$(\mathcal{H}) \geq 1$? $C = \{x_3\} \Rightarrow \mathcal{H}_C = \{[0],[1]\}$ YES

Is the VC-dim$(\mathcal{H}) \geq 2$? $C = \{x_5, x_6\} \Rightarrow \mathcal{H}_C = \{[0,0],[0,1],[1,1],[1,0]\}$ Yes

Is the VC-dim$(\mathcal{H}) \geq 3$? No, because we need at least a column with ≥ 4 1's $\Rightarrow$ no set of size 3 can be shattered by $\mathcal{H}$

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

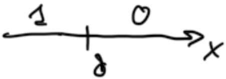
1. there exists a set C of size d which is shattered by $\mathcal{H}$
2. *every* set of size $d + 1$ is not shattered by $\mathcal{H}$

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

Example: Threshold Functions

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

h_a :



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

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

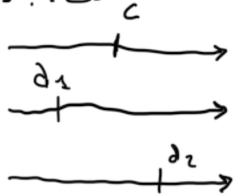
$$h_a(x) = \mathbb{1}[x < a] = \begin{cases} 1 & \text{if } x < a \\ 0 & \text{if } x \geq a \end{cases}$$

VC-dim$(\mathcal{H}) \geq 1$? YES

instance

h_{a_1}

h_{a_2}



$\Rightarrow h_{a_1}(c) = 0$

$\Rightarrow h_{a_2}(c) = 1$

Example: Threshold Functions

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

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?

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

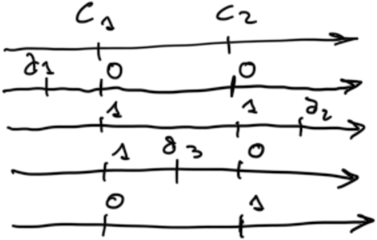
$\Downarrow$

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

h_{a_1}

h_{a_2}

h_{a_3}



$(c_1 < c_2)$

$a_1 < c_1$

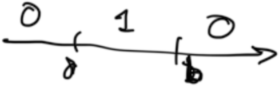
$a_2 > c_2$

$c_1 < a_3 < c_2$

CANNOT BE OBTAINED

Example: Intervals

$h_{a,b}$

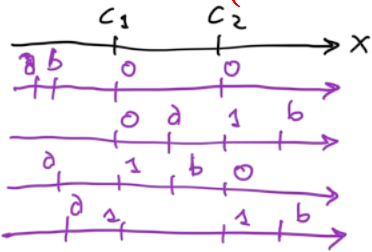


$$\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$$

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

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

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



$c_1 < c_2$

Example: Intervals

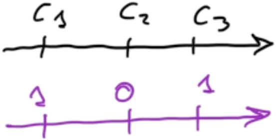
$$\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$$

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?

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



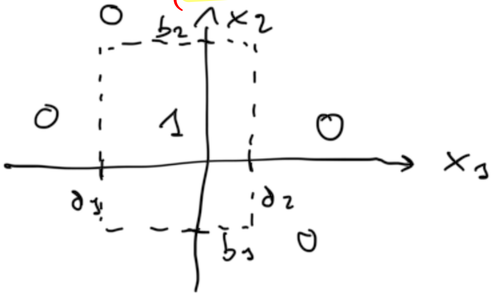
cannot be obtained with $\mathcal{H}$

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

Example: Axis Aligned Rectangles

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

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

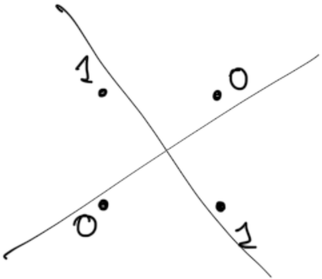


Example: Axis Aligned Rectangles

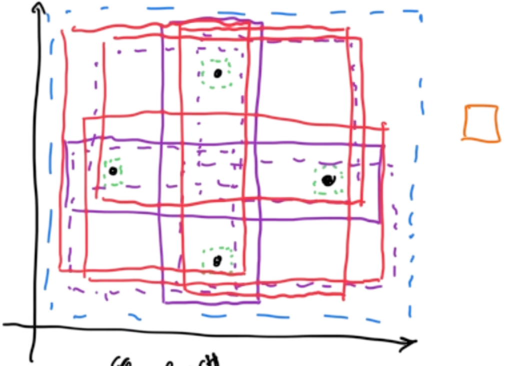
$$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1, a_2, b_1, b_2 \in \mathbb{R}, a_1 \le a_2, b_1 \le b_2\}$$

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

VC-dimension?



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



$\mathrm{VCdim} \leq 4$:

cannot be obtained

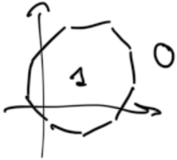
$\rightarrow \mathrm{VCdim}(\mathcal{H}) = 4$

Example: Convex Sets

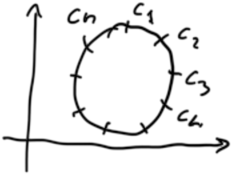
Model set $\mathcal{H}$ such that for $h_S \in \mathcal{H}$, $h_S : \mathbb{R}^2 \to \{0, 1\}$ with

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

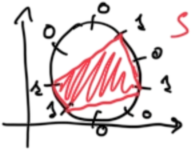
where S is a convex subset of $\mathbb{R}^2$



Consider an arbitrary value of $n \in \mathbb{N}^+$
(n = size of the set to be shattered)



Consider an arbitrary labeling of $c_1, c_2, \dots, c_n$:



the hypothesis corresponding to the convex set with vertices given by pairs $(\vec{c}_i, y_i)$ with $y_i = 1$ gives the desired labeling

$\Rightarrow \mathcal{H}$ can shatter a set of n points for any arbitrarily large n

$\Rightarrow \text{vcdim}(\mathcal{H}) = +\infty$

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 $\mathcal{H}$.

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:

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

size of the training set

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

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

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 θ!

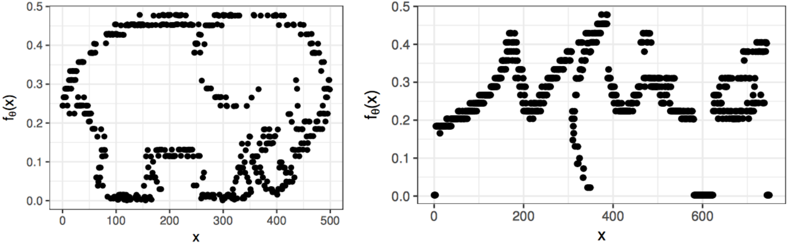


FIG. 1: A scatter plot of f_θ for $\theta = 0.2446847266734745458227540656\cdots$ plotted at integer x values, showing that a single parameter can fit an elephant (left). The same model run with parameter $\theta = 0.0024265418055000401935387620\cdots$ showing a fit of a scatter plot to Joan Miró's signature (right). Both use $r = 8$ and require hundreds to thousands of digits of precision in θ.

[“One parameter is always enough”, Piantadosi, 2018]

Machine Learning

Clustering

Fabio Vandin December 15^{th}, 2023

Unsupervised Learning

In unsupervised learning, the training dataset is $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$
$\Rightarrow$ no target values!

We are interested in finding some interesting *structure* in the data, or, equivalently, to organize it in some meaningful way.

We are going to see the most common unsupervised learning approaches: *clustering*

We are going to focus on the most commonly used techniques:

- k-means
- linkage-based clustering,

There are also other general techniques: dimensionality reduction, association analysis,...

Example

- Data: features (e.g. product bought, demographic info, etc.) for a large number of customers
- Goal: **customers segmentation** = identify subgroups of homogeneous customers
- useful for: advertizing, product development, ...

Example (2)

Data:

- rows = genes ($\approx 20 \times 10^3$)
- columns = samples, cancer patients ($\approx 10^3 - 10^4$)
- values = expression of a gene in a patient ($\in \mathbb{R}$)

Goal: find similar cancer samples

- cluster columns (samples) to find similar subgroups of patients (e.g., *disease subtypes*)

Goal: find genes with similar gene expression profiles

- cluster rows (genes) to deduce function of unknown genes from experimentally known genes with similar profiles

Other Applications

- **Information Retrieval**: clustering is used to *find* topics/categories of documents that are not explicitly given
- **Image Processing**: used for several tasks/applications, including: identification of different types of tissues in PET scans; identification of areas of similar land use in satellite pictures;...
- **Analysis of Social Networks**: detection of communities
- ...