Lecture 7: The VC dimension

Presentation · May 2019

1 author:

Alaa Tharwat
Fachhochschule Bielefeld

Lecture 7: The VC dimension

Alaa Tharwat

Lecture 7: The VC dimension

- Review of Lecture 6
- Mathematical proof ($d_{VC} = d + 1$)

Lecture 7: The VC dimension

- Review of Lecture 6
- Mathematical proof ($d_{VC} = d + 1$)

$$B(N,k) \leq B(N-1,k) + B(N-1,k-1)$$

$$\leq 1 + \sum_{i=1}^{k-1} \binom{N-1+1}{i} = \sum_{i=0}^{k-1} \binom{N}{i}$$

$$m_H(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

		# of rows	x_1	x_2	....	x_{N-1}	x_N
	S_1	α	+1	+1	....	+1	+1
			-1	+1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	-1	-1
			-1	+1	....	-1	+1
S_2	S_2^+	β	+1	-1	....	+1	+1
			-1	-1	....	+1	+1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	+1
			-1	-1	....	-1	+1
	S_2^-	β	+1	-1	....	+1	-1
			-1	-1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	-1
			-1	-1	....	-1	-1

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2m_H(N)e^{-2\epsilon^2 N}$$

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4m_H(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

This is called, the Vapnik-Chervonenkis inequality

N-1,k-1 N-1,k

N \ k	1	2	3	4	5	6	...
1	1	2	2	2	2	2	...
2	1	3	4	4	4	4	...
3	1	4	7	8	8	8	...
4	1	5	11	...	...	...	...
5	1	6	...				
6	1	7	...				
⋮	⋮	⋮	⋮				

N,k

Lecture 7: The VC dimension

- Review of Lecture 6
- Mathematical proof ($d_{VC} = d + 1$)

- The VC dimension is denoted by $d_{VC}(H)$, and it is the largest value of N or the most points that H can shatter, $m_H(N) = 2^N$
- If $N \leq d_{VC}(H) \Rightarrow H$ can shatter N points
- If $k > d_{VC}(H) \Rightarrow k$ is a break point for H

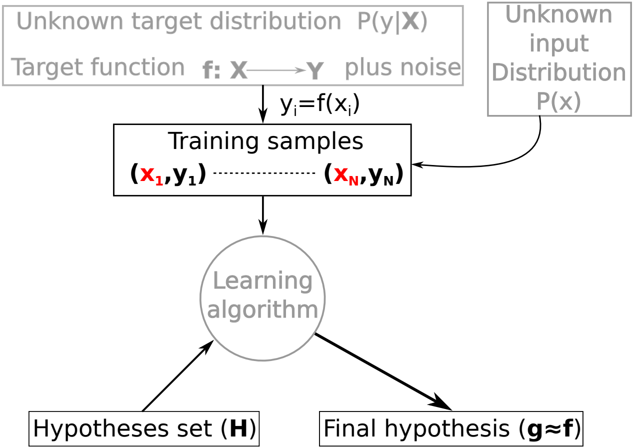
$$m_H(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

In terms of VC dimension:

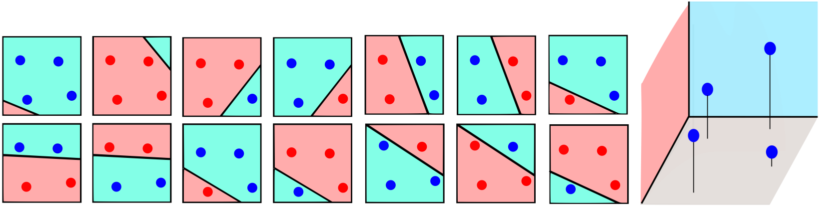
$$m_H(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$$

Hence, the maximum power is $N^{d_{VC}}$

- $d_{VC}(H)$ is finite
- VC-dimension is independent of the learning algorithm
- VC-dimension is independent of the input distribution
- VC-dimension is independent of the target function



- In positive rays example: $d_{VC} = k - 1 = 1$
- In positive intervals example: $d_{VC} = 2$
- In 2D perceptrons example: $d_{VC} = 3$



We can say, $d_{VC} = d + 1$ (the proof next slides)

To prove that $d_{VC} = d + 1$, we prove two different directions:

- $d_{VC} \geq d + 1$
- $d_{VC} \leq d + 1$

- First direction: $\boxed{d_{VC} \geq d + 1}$
  - Given a set of $N = d + 1$ points in R^d to be shattered by the perceptron

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_{d+1}^T \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix}$$

- In X, we select the samples to be shattered and X is then invertible. The y for each sample can be ± 1.
- What $\mathbf{w}$ that satisfies $sign(\mathbf{w}X) = y \rightarrow X\mathbf{w} = y \rightarrow \mathbf{w} = X^{-1}y$. So, it can shatter $d + 1$ points.
- Hence $d_{VC} \geq d + 1$ and we cannot shatter any set of $d + 2$ points.

To prove that $d_{VC} = d + 1$, we prove two different directions:

- $d_{VC} \geq d + 1$
- $d_{VC} \leq d + 1$

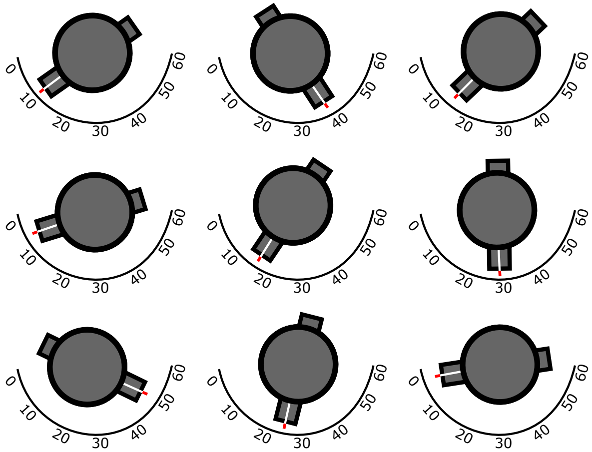
- Second direction: $\boxed{d_{VC} \leq d + 1}$
  - We need to prove that we cannot shatter any set with $d + 2$ points
  - Given $d + 2$ points, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{d+1}, \mathbf{x}_{d+2}$ (more points than the dimensions)
  - Hence, there is at least one sample which linearly dependent to the others (linear combination of the rest), $\mathbf{x}_j = \sum_{i \neq j} a_i \mathbf{x}_i$ where not all a's are zeros
  - Consider the following dichotomy: $\mathbf{x}_i$'s with non-zero a_i get $y_i = \text{sign}(a_i)$, and $\mathbf{x}_j$ gets $y_i = -1$ (No perceptron can implement this dichotomy, why?)
    - Answer: $\mathbf{x}_j = \sum_{i \neq j} a_i \mathbf{x}_i \xrightarrow{\times \mathbf{w}} \rightarrow \mathbf{w}^T \mathbf{x}_j = \sum_{i \neq j} a_i \mathbf{w}^T y_i$, and if $y_i = \text{sign}(\mathbf{w}^T \mathbf{x}_i) = \text{sign}(a_i) \Rightarrow a_i \mathbf{w}^T \mathbf{x}_i > 0$. Thus, $\mathbf{w}^T \mathbf{x}_j = \sum_{i \neq j} a_i \mathbf{w}^T x_i > 0$; therefore, $y_i = \text{sign}(\mathbf{w}^T \mathbf{x}_j) = +1$

- $d_{VC} \geq d + 1$
- $d_{VC} \leq d + 1$

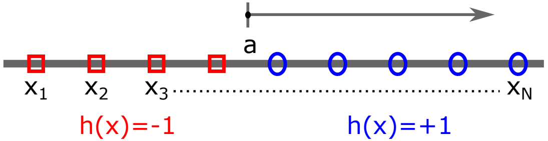
$$\boxed{\Rightarrow d_{VC} = d + 1}$$

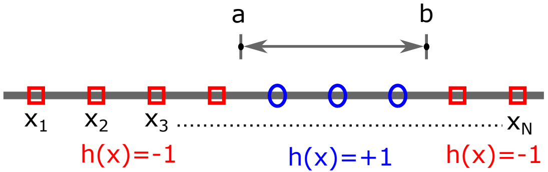
- In a 2D perceptron, $d + 1$ is the number of parameters $w_0, w_1, \ldots, w_d$

- Parameters create *degrees of freedom*
- $d_{VC} \equiv$ binary degrees of freedom



- In positive rays example: $d_{VC} = k - 1 = 1$, and there is only one parameter
- In positive intervals example: $d_{VC} = 2$, and there are two parameters





- Parameters may not contribute the degrees for freedom
- d_{VC} measures the effective number of parameters

From Lecture 4:

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

From Lecture 6:

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4m_H(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

- Let $\delta = 4m_H(2N)e^{-\frac{1}{8}\epsilon^2 N}$

$$E_{out} \leq E_{in} + \epsilon$$

$$E_{out} \leq E_{in} + \sqrt{\frac{ln\frac{2M}{\delta}}{2N}}$$

From Lecture 4:

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

From Lecture 6:

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4m_H(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

- Let $\delta = 4m_H(2N)e^{-\frac{1}{8}\epsilon^2 N}$
- $\epsilon = \sqrt{\frac{8ln\frac{4m_H(2N)}{\delta}}{N}} \Rightarrow \Omega(N, H, \delta)$

$$E_{out} \leq E_{in} + \Omega(N, H, \delta)$$

- We can consider this term like that $4m_H(2N)e^{-\frac{1}{8}\epsilon^2 N} \Rightarrow N^d e^{-N}$
- $N^d e^{-N}$ decreases (goes to zero) when N is very large
- How does N affects d?
- Rule of thumb: $N \geq 10 d_{VC}$ (the number of samples to reach to the comfort zone of the VC-inequality)