Hoeffding inequality

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

VC inequality

$$P\left(\left|E_{in}(g) - E_{out}(g)\right| > \epsilon \right) \leq 4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N}$$

Quick review

Hypothesis space: $\mathcal{H}$

Growth-function: $m_{\mathcal{H}}(N)$ (counts dichotomies)

Break point: k is a break point for $\mathcal{H}$ if there is no dataset of size k for which $\mathcal{H}$ generates all 2^k dichotomies

$m_{\mathcal{H}}(N)$ is polynomial if there is a break-point

The bound $4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N}$ in the VC inequality tends to zero as N increases (The negative exponential starts to dominate the polinomial at some point)

VC dimension

VC dimension $d_{\text{VC}}(\mathcal{H})$:

The largest number of points that can be shattered by $\mathcal{H}$
(The largest value of N for which $m_{\mathcal{H}}(N) = 2^N$)

Break point:

k is a break point for $\mathcal{H}$ if there is no dataset of size k shattered by $\mathcal{H}$

If k is a break point for $\mathcal{H}$, then $d_{\text{VC}}(\mathcal{H}) < k$

$d_{\text{VC}}(\mathcal{H}) + 1$ is a *break-point* for $\mathcal{H}$

The growth function

In terms of a break point k:

$$m_{\mathcal{H}}(N) \;\le\; \sum_{i=0}^{k-1} \binom{N}{i}$$

In terms of the VC dimension d_{VC}:

$$m_{\mathcal{H}}(N) \;\le\; \underbrace{\sum_{i=0}^{d_{\mathrm{VC}}} \binom{N}{i}}_{\text{maximum power is } N^{d_{\mathrm{VC}}}}$$

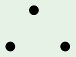
$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{\mathrm{VC}}} \binom{N}{i} \leq N^{d_{\mathrm{VC}}} + 1$$

Examples

- $\mathcal{H}$ is positive rays:

$$d_{\mathrm{VC}} = 1$$

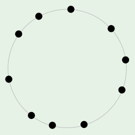
- $\mathcal{H}$ is 2D perceptrons:



$$d_{\mathrm{VC}} = 3$$

- $\mathcal{H}$ is convex sets:

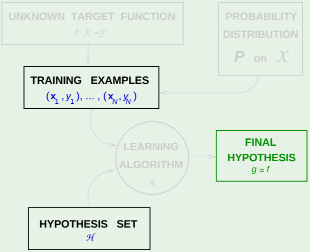
$$d_{\mathrm{VC}} = \infty$$



VC dimension and learning

$d_{\text{VC}}(\mathcal{H})$ is finite $\implies$ $g \in \mathcal{H}$ will generalize

- Independent of the learning algorithm
- Independent of the input distribution
- Independent of the target function



The VC inequality holds for

- any target function
- any input distribution
- any learning algorithm

It is a "worst case bound"

Example: VC dimension of the perceptron

Let d be the input data dimension ($\mathbf{x} = (x_1, x_2, \dots, x_d)$)

For perceptrons, $d_{\textbf{vc}} = d + 1$

To prove it, it is enough to show that

(a) $d_{\text{vc}} \geq d + 1$, and

(b) $d_{\text{vc}} \leq d + 1$

What do we need to do to prove (a) $d_{\text{VC}} \geq d + 1$?

What do we need to do to prove (a) $d_{\text{VC}} \geq d + 1$?

A. We need to show that there is a set of $d + 1$ points that can be shattered by the perceptron

How? Carefully choose $d + 1$ points, assign arbitrary labels in $\{-1, +1\}$ for each of them, and then show that there is a hypothesis that agrees with the labels

Here is one direction

A set of $N = d+1$ points in $\mathbb{R}^d$ shattered by the perceptron:

$$\mathrm{X} = \begin{bmatrix} -\ \mathbf{x}_1^{\intercal}\ - \\ -\ \mathbf{x}_2^{\intercal}\ - \\ -\ \mathbf{x}_3^{\intercal}\ - \\ \vdots \\ -\mathbf{x}_{d+1}^{\intercal}- \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix}$$

X is invertible

Can we shatter this data set?

For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, can we find a vector $\mathbf{w}$ satisfying

$$\text{sign}(\mathrm{X}\mathbf{w}) = \mathbf{y}$$

Easy! Just make $\mathrm{X}\mathbf{w} = \mathbf{y}$

which means $\mathbf{w} = \mathrm{X}^{-1}\mathbf{y}$

What do we need to do to prove (b) $d_{\text{VC}} \leq d + 1$?

What do we need to do to prove (b) $d_{\text{VC}} \leq d + 1$?

A. We need to show that no set of $d + 2$ points can be shattered by the perceptron

How? Take any set of $d + 2$ points and show that it is always possible to build a dichotomy that can not be generated by any of the hypotheses

Take any $d+2$ points

For any $d+2$ points,

$$\mathbf{x}_1, \cdots, \mathbf{x}_{d+1}, \mathbf{x}_{d+2}$$

More points than dimensions $\implies$ we must have

$$\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{x}_i$$

where not all the a_i's are zeros

So?

$$\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{x}_i$$

Consider the following dichotomy:

$\mathbf{x}_i$'s with non-zero a_i get $\quad y_i = \text{sign}(a_i)$

and $\mathbf{x}_j$ gets $\quad y_j = -1$

No perceptron can implement such dichotomy!

Why?

$$\mathbf{x}_j \;=\; \sum_{i\neq j} a_i\ \mathbf{x}_i \quad\Longrightarrow\quad \mathbf{w}^{\intercal}\mathbf{x}_j \;=\; \sum_{i\neq j} a_i\ \mathbf{w}^{\intercal}\mathbf{x}_i$$

If $y_i = \mathrm{sign}(\mathbf{w}^{\intercal}\mathbf{x}_i) = \mathrm{sign}(a_i)$, then $a_i\ \mathbf{w}^{\intercal}\mathbf{x}_i \;>\; 0$

This forces

$$\mathbf{w}^{\intercal}\mathbf{x}_j = \sum_{i\neq j} a_i\ \mathbf{w}^{\intercal}\mathbf{x}_i \;>\; 0$$

Therefore, $y_j = \mathrm{sign}(\mathbf{w}^{\intercal}\mathbf{x}_j) = +1$

Putting it together

We proved $d_{\mathrm{VC}} \leq d+1$ and $d_{\mathrm{VC}} \geq d+1$

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

What is $d+1$ in the perceptron?

It is the number of parameters $w_0, w_1, \cdots, w_d$

Discussions

- Interpretation of VC dimension
  - what it signifies
  - is there a practical use ?

- Some comments on the VC bound

1. Degrees of freedom

Parameters create degrees of freedom

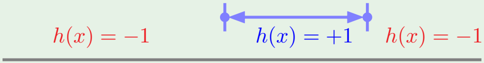
# of parameters: **analog** degrees of freedom

d_{VC}: equivalent '**binary**' degrees of freedom

The usual suspects

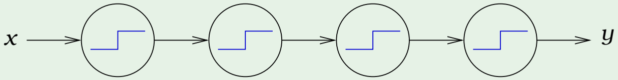
Positive rays ($d_{\text{VC}} = 1$):

Positive intervals ($d_{\text{VC}} = 2$):



Not just parameters

Parameters may not contribute degrees of freedom:



d_{VC} measures the **effective** number of parameters

$$P\left(\left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right) \leq 4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N}$$

If d_{VC} is finite, learning generalises

How many examples do we need ?

Let us examine the behavior of a rough approximation for the bound:

$$N^{d_{\text{VC}}} e^{-N}$$ (Recall that $m_{\mathcal{H}}(N) \leq N^{d_{\text{VC}}} + 1$)

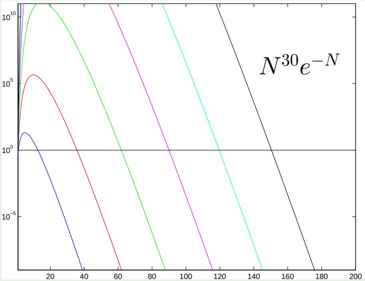
$N^{d}e^{-N}$

Fix $N^{d}e^{-N}$ = small value

How does N change with d?

Rule of thumb:

$$N \;\geq\; 10\, d_{\mathrm{VC}}$$



Given ϵ, we have the bound (δ):

$$P\Big(\big|E_{in}(g) - E_{out}(g)\big| > \epsilon \Big) \leq \underbrace{4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

Given δ, we can compute ϵ:

$$\delta = 4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N} \Longrightarrow \epsilon = \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}$$

$$P\Big(\big|E_{in}(g) - E_{out}(g)\big| > \epsilon \Big) \leq \delta \Longleftrightarrow P\Big(\big|E_{in}(g) - E_{out}(g)\big| \leq \epsilon \Big) > 1 - \delta$$

With probability at leas $1 - \delta$ we have

$$\big|E_{in}(g) - E_{out}(g)\big| \leq \epsilon$$

Probably approximately correct (PAC)

Rearranging things

Start from the VC inequality:

$$\mathbb{P}\left[|E_{\text{out}} - E_{\text{in}}| > \epsilon\right] \ \leq \ \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

Get ϵ in terms of δ:

$$\delta = 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \ \Longrightarrow \ \epsilon = \underbrace{\sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}}_{\Omega}$$

With probability $\geq 1 - \delta$, $\qquad |E_{\text{out}} - E_{\text{in}}| \leq \Omega(N, \mathcal{H}, \delta)$

Generalization bound

With probability $\geq 1 - \delta$, $\quad E_{\text{out}} - E_{\text{in}} \leq \Omega$

$$\Longrightarrow$$

With probability $\geq 1 - \delta$,

$$E_{\text{out}} \leq E_{\text{in}} + \Omega$$

Summary

1. Dichotomies are the key for the definition of VC dimension

2. The VC dimension replaces M (size of $\mathcal{H}$) in the Hoeffding inequality bound

$$P\Big(\big|E_{in} - E_{out}\big| > \epsilon \Big) \leq 4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N} \qquad (m_{\mathcal{H}}(2N) \leq (2N)^{d_{\mathrm{VC}}} + 1)$$

3. VC dimension is related to the expressiveness of $\mathcal{H}$

4. $E_{out} \leq E_{in} + \underbrace{\sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}}_{\Omega}$

d_{VC}	E_{in}	Ω
small	large	small
$\downarrow$	$\uparrow$	$\downarrow$
large	small	large