MAC 0460 / 5832
Introduction to Machine Learning

10 – Is learning feasible ?

• VC dimension •

IME/USP (17/05/2021)

Hoeffding inequality

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

VC inequality

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

Growth function

Quick review

Hypothesis space: $\mathcal{H}$

Growth-function: $m_{\mathcal{H}}(N)$ (counts max number of 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 polynomial at some point)

VC dimension

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

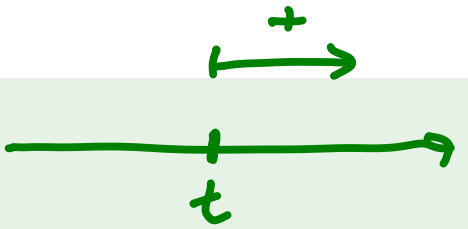
> *The largest number of points that can be shattered by* $\mathcal{H}$
> *(The largest value of* N *for which* $m_{\mathcal{H}}(N) = 2^N$*)*

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 $d_{VC}(\mathcal{H}) < k$

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



Examples

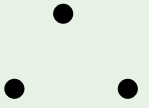
- $\mathcal{H}$ is positive rays:

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

K = 2

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

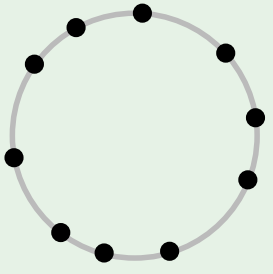
$$d_{\text{VC}} = 3$$



K = 4 is break-point.

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

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



Example: VC dimension of the perceptron

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

For perceptrons, $d_{\text{VC}} = d + 1$

To prove it, it is enough to show that

- (a) $d_{\text{VC}} \geq d + 1$, and
- (b) $d_{\text{VC}} \leq d + 1$

2D
$K=4$ is break-point
$K=3 \Rightarrow 2^3$ partitions
$d_{VC} = 3$ $(d=2)$

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

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:

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

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

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

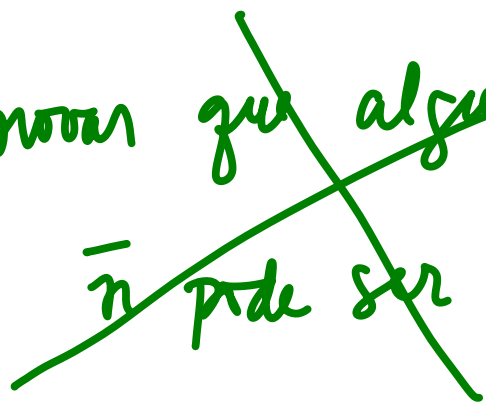
Easy! Just make $\mathrm{X}\mathbf{w} = \mathbf{y}$

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

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

1) Mostrar que $K = d + 2$ é break-point ⇐

2) provar que algum conjunto com $d+2$ pode ser shattered.



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 $y_i = \text{sign}(a_i)$

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

No perceptron can implement such dichotomy!

Why?

$$\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{x}_i \implies \mathbf{w}^\mathsf{T}\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{w}^\mathsf{T}\mathbf{x}_i$$

If $y_i = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_i) = \text{sign}(a_i)$, then $a_i \, \mathbf{w}^\mathsf{T}\mathbf{x}_i > 0$

This forces

$$\mathbf{w}^\mathsf{T}\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{w}^\mathsf{T}\mathbf{x}_i > 0$$

Therefore, $y_j = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_j) = +1$

Putting it together

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

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

What is $d+1$ in the perceptron?

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

The growth function

In terms of a break point k:

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

In terms of the VC dimension d_{VC}:

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

$d_{VC} < k$

$k = d_{VC} + 1$ is a break-point

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

(Proof by induction; Problem 2.5 of the book)

VC inequality

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

$$m_{\mathcal{H}}(2N) \leq (2N)^{d_{\text{VC}}} + 1$$

VC bound

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

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

0.05

If $P(|a-b| > \epsilon) \leq \delta$, then with probability $1-\delta$ we have $|a-b| \leq \epsilon$, i.e., $b - a \leq \epsilon \leq a - b$

Letting $a = E_{in}$ and $b = E_{out}$, with probability $1-\delta$ we have:

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

Ω

Rearranging things

Start from the VC inequality:

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

bad event

$\delta = 0.05$

Get ϵ in terms of δ:

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

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

Generalization bound

With probability $\geq 1 - \delta$,

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

$$\Longrightarrow$$

With probability $\geq 1 - \delta$,

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

Remarks

- **Generalization error**:
  - in this course it refers to $|E_{in}(h) - E_{out}(h)|$
  - However, often it is used to name $E_{out}(h)$

- **VC generalization bound**: with probability $1 - \delta$ ($\delta > 0$)

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

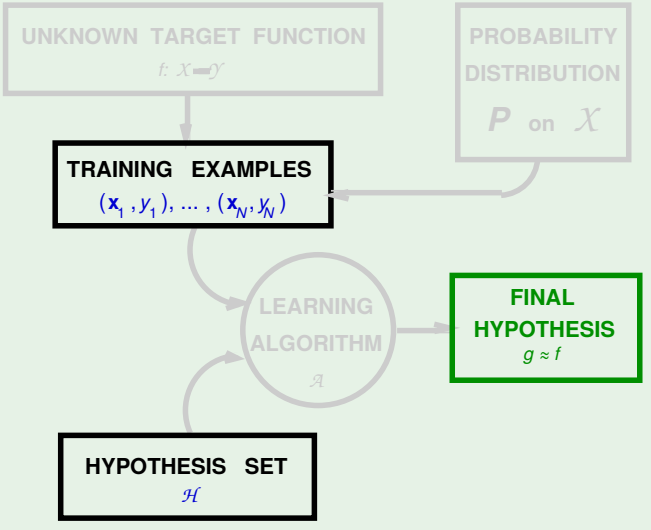
Some discussions

- Intuitive meaning of VC dimension
- Sample complexity

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



1. Degrees of freedom

Parameters create degrees of freedom

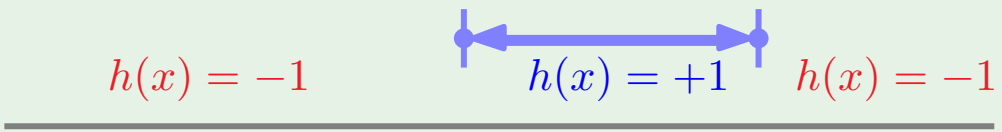
# of parameters: **analog** degrees of freedom

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

The usual suspects

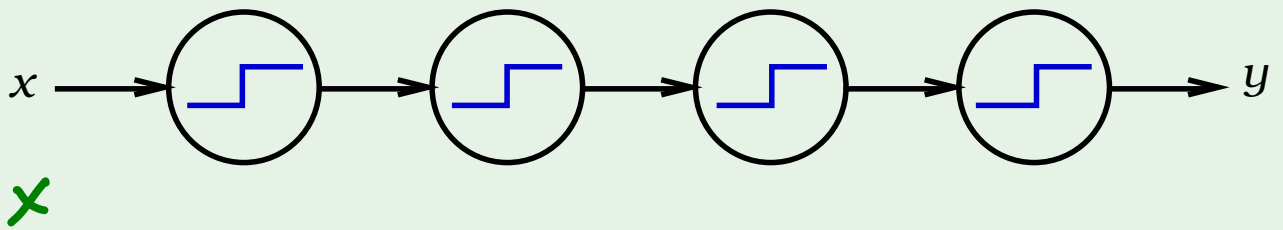
Positive rays ($d_{\text{VC}} = 1$):

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



Not just parameters

Parameters may not contribute degrees of freedom:



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

Sample complexity

If d_{VC} is finite, learning generalizes

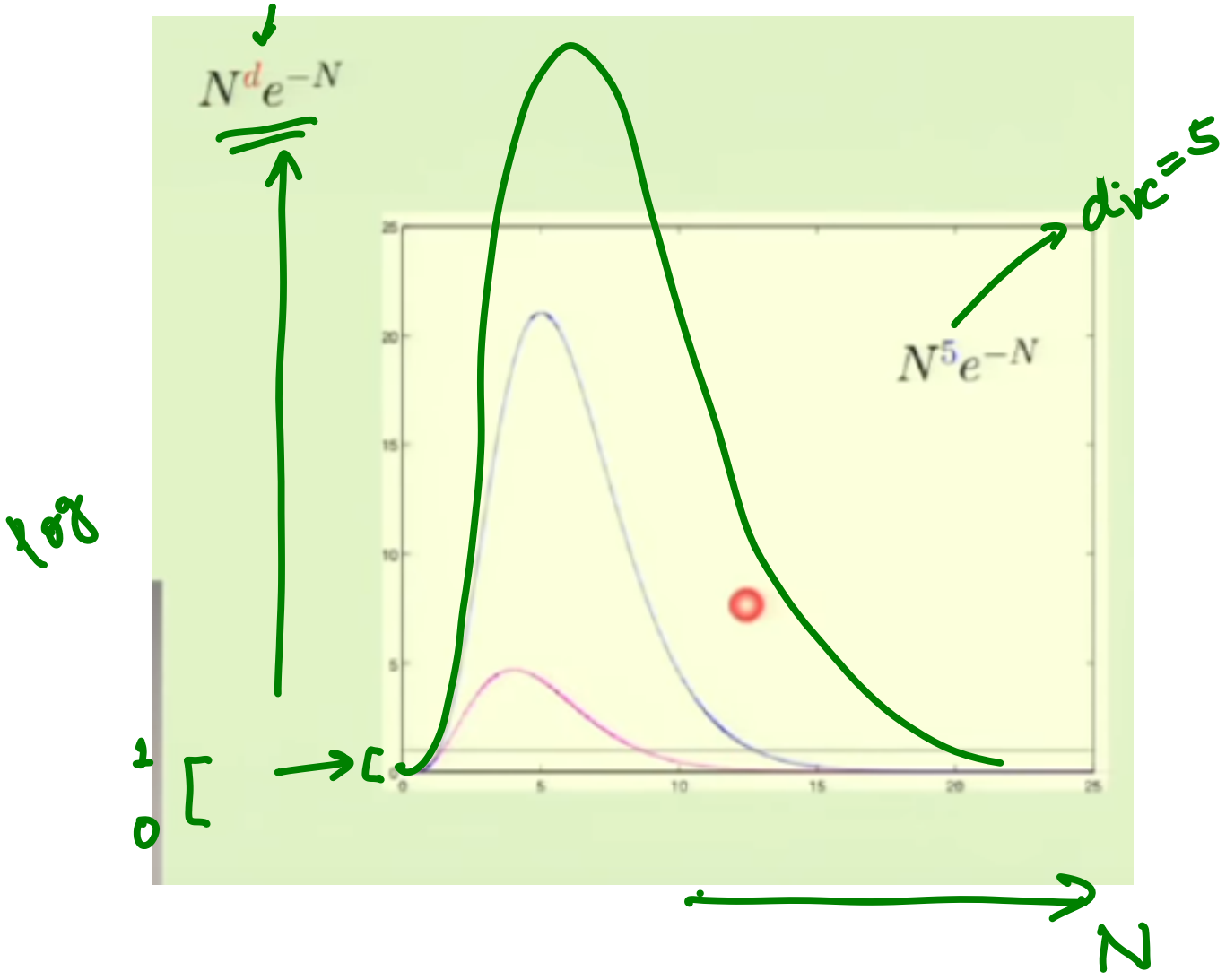
How many examples do we need ?

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

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

$N^{d_{VC}} e^{-N}$

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

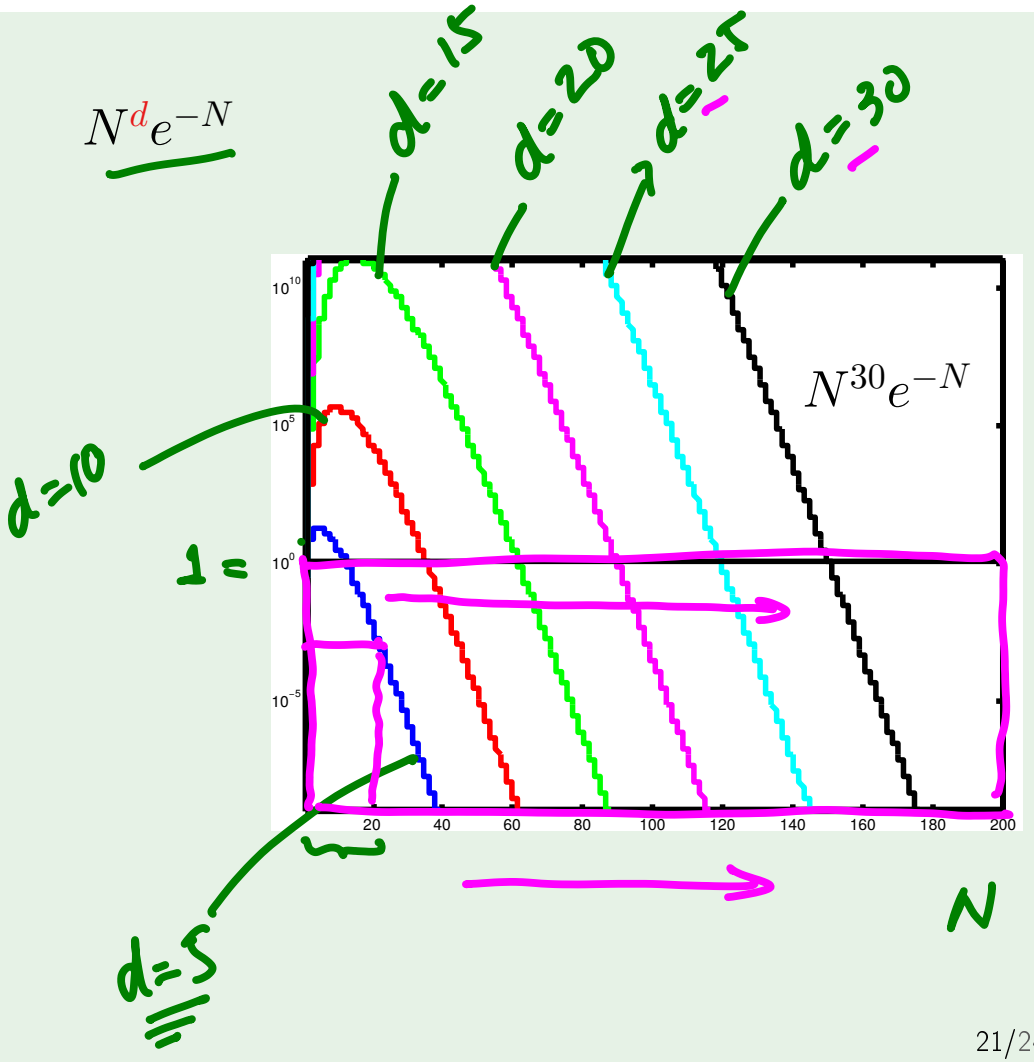

$N^d e^{-N}$
$N^5 e^{-N}$
div=5
log
N

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

How does N change with d?

Rule of thumb:

$$N \geq 10\, d_{\text{VC}}$$



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

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

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