

# Machine Learning Foundations

## (機器學習基石)



### Lecture 7: The VC Dimension

Hsuan-Tien Lin (林軒田)

htlin@csie.ntu.edu.tw

Department of Computer Science  
& Information Engineering

National Taiwan University  
(國立台灣大學資訊工程系)



# Roadmap

- 1 When Can Machines Learn?
- 2 **Why** Can Machines Learn?

## Lecture 6: Theory of Generalization

$E_{\text{out}} \approx E_{\text{in}}$  possible  
if  $m_{\mathcal{H}}(N)$  **breaks somewhere** and  $N$  **large enough**

## Lecture 7: The VC Dimension

- Definition of VC Dimension
- VC Dimension of Perceptrons
- Physical Intuition of VC Dimension
- Interpreting VC Dimension

- 3 How Can Machines Learn?
- 4 How Can Machines Learn Better?

# Recap: More on Growth Function

$$m_{\mathcal{H}}(N) \text{ of break point } k \leq B(N, k) = \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{highest term } N^{k-1}}$$

$B(N, k)$		$k$				
		1	2	3	4	5
$N$	1	1	2	2	2	2
	2	1	3	4	4	4
	3	1	4	7	8	8
	4	1	5	11	15	16
	5	1	6	16	26	31
	6	1	7	22	42	57

$N^{k-1}$	$k$				
	1	2	3	4	5
1	1	1	1	1	1
2	1	2	4	8	16
3	1	3	9	27	81
4	1	4	16	64	256
5	1	5	25	125	625
6	1	6	36	216	1296

**provably** & loosely, for  $N \geq 2, k \geq 3$ ,

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

# Recap: More on Vapnik-Chervonenkis (VC) Bound

For any  $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$  and ‘statistical’ large  $\mathcal{D}$ , for  $N \geq 2, k \geq 3$

$$\begin{aligned}
 & \mathbb{P}_{\mathcal{D}} \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \\
 & \leq \mathbb{P}_{\mathcal{D}} \left[ \exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon \right] \\
 & \leq 4m_{\mathcal{H}}(2N) \exp \left( -\frac{1}{8} \epsilon^2 N \right) \\
 & \stackrel{\text{if } k \text{ exists}}{\leq} 4(2N)^{k-1} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
 \end{aligned}$$

if ①  $m_{\mathcal{H}}(N)$  breaks at  $k$  (good  $\mathcal{H}$ )

②  $N$  large enough (good  $\mathcal{D}$ )

$\Rightarrow$  probably generalized ‘ $E_{\text{out}} \approx E_{\text{in}}$ ’, and

if ③  $\mathcal{A}$  picks a  $g$  with small  $E_{\text{in}}$  (good  $\mathcal{A}$ )

$\Rightarrow$  probably learned! (:-) good luck)

# VC Dimension

the formal name of **maximum non-break point**

## Definition

VC dimension of  $\mathcal{H}$ , denoted  $d_{VC}(\mathcal{H})$  is

**largest**  $N$  for which  $m_{\mathcal{H}}(N) = 2^N$

- the **most** inputs  $\mathcal{H}$  that can shatter
- $d_{VC} = \text{'minimum } k' - 1$

$N \leq d_{VC} \implies \mathcal{H}$  can shatter some  $N$  inputs

$k > d_{VC} \implies k$  is a break point for  $\mathcal{H}$

if  $N \geq 2, d_{VC} \geq 2, m_{\mathcal{H}}(N) \leq N^{d_{VC}}$

# The Four VC Dimensions

- positive rays:

$$d_{VC} = 1$$



$$m_{\mathcal{H}}(N) = N + 1$$

- positive intervals:

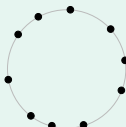
$$d_{VC} = 2$$



$$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

- convex sets:

$$d_{VC} = \infty$$



$$m_{\mathcal{H}}(N) = 2^N$$

- 2D perceptrons:

$$d_{VC} = 3$$



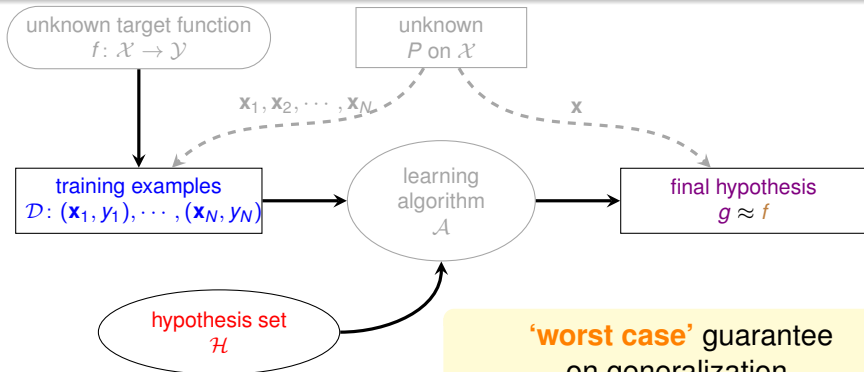
$$m_{\mathcal{H}}(N) \leq N^3 \text{ for } N \geq 2$$

good: **finite**  $d_{VC}$

# VC Dimension and Learning

**finite  $d_{\text{VC}} \implies g$  'will' generalize ( $E_{\text{out}}(g) \approx E_{\text{in}}(g)$ )**

- regardless of learning algorithm  $\mathcal{A}$
- regardless of input distribution  $P$
- regardless of target function  $f$



**'worst case' guarantee**  
on generalization

# Fun Time

If there is a set of  $N$  inputs that cannot be shattered by  $\mathcal{H}$ . Based only on this information, what can we conclude about  $d_{\text{VC}}(\mathcal{H})$ ?

- ①  $d_{\text{VC}}(\mathcal{H}) > N$
- ②  $d_{\text{VC}}(\mathcal{H}) = N$
- ③  $d_{\text{VC}}(\mathcal{H}) < N$
- ④ no conclusion can be made

Reference Answer: ④

It is possible that there is another set of  $N$  inputs that can be shattered, which means  $d_{\text{VC}} \geq N$ . It is also possible that no set of  $N$  input can be shattered, which means  $d_{\text{VC}} < N$ . Neither cases can be ruled out by one non-shattering set.



## 2D PLA Revisited

linearly separable  $\mathcal{D}$ with  $\mathbf{x}_n \sim P$  and  $y_n = f(\mathbf{x}_n)$ 

PLA can converge

 $\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \dots$  by  $d_{\text{VC}} = 3$  $T$  large $N$  large



$$E_{\text{in}}(g) = 0$$

$$E_{\text{out}}(g) \approx E_{\text{in}}(g)$$

$$E_{\text{out}}(g) \approx 0 \text{ :-)}$$

general PLA for  $\mathbf{x}$  with more than 2 features?

# VC Dimension of Perceptrons

- 1D perceptron (pos/neg rays):  $d_{VC} = 2$
- 2D perceptrons:  $d_{VC} = 3$ 
  - $d_{VC} \geq 3$ : 
  - $d_{VC} \leq 3$ : 
- $d$ -D perceptrons:  $d_{VC} \stackrel{?}{=} d + 1$

two steps:

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

## Extra Fun Time

What statement below shows that  $d_{VC} \geq d + 1$ ?

- ① There are some  $d + 1$  inputs we can shatter.
- ② We can shatter any set of  $d + 1$  inputs.
- ③ There are some  $d + 2$  inputs we cannot shatter.
- ④ We cannot shatter any set of  $d + 2$  inputs.

Reference Answer: ①


$d_{VC}$  is the maximum that  $m_{\mathcal{H}}(N) = 2^N$ , and  $m_{\mathcal{H}}(N)$  is the most number of dichotomies of  $N$  inputs. So if we can find  $2^{d+1}$  dichotomies on *some*  $d + 1$  inputs,  $m_{\mathcal{H}}(d + 1) = 2^{d+1}$  and hence  $d_{VC} \geq d + 1$ .

$$d_{\text{VC}} \geq d + 1$$

There are **some**  $d + 1$  **inputs** we can shatter.

- some 'trivial' inputs:

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

- visually in 2D: 

note: **X invertible!**

## Can We Shatter X?

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

to shatter ...

for any  $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_{d+1} \end{bmatrix}$ , find  $\mathbf{w}$  such that

$$\text{sign}(\mathbf{X}\mathbf{w}) = \mathbf{y} \iff (\mathbf{X}\mathbf{w}) = \mathbf{y} \overset{\text{X invertible!}}{\iff} \mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$$

‘special’  $X$  can be shattered  $\implies d_{VC} \geq d + 1$

## Extra Fun Time

What statement below shows that  $d_{VC} \leq d + 1$ ?

- ① There are some  $d + 1$  inputs we can shatter.
- ② We can shatter any set of  $d + 1$  inputs.
- ③ There are some  $d + 2$  inputs we cannot shatter.
- ④ We cannot shatter any set of  $d + 2$  inputs.

Reference Answer: ④

$d_{VC}$  is the maximum that  $m_{\mathcal{H}}(N) = 2^N$ , and  $m_{\mathcal{H}}(N)$  is the most number of dichotomies of  $N$  inputs. So if we cannot find  $2^{d+2}$  dichotomies on *any*  $d + 2$  inputs (i.e. break point),  $m_{\mathcal{H}}(d + 2) < 2^{d+2}$  and hence  $d_{VC} < d + 2$ . That is,  $d_{VC} \leq d + 1$ .

$$d_{VC} \leq d + 1 \quad (1/2)$$

## A 2D Special Case

$$\begin{matrix} \bullet & \bullet \\ \bullet & \bullet \end{matrix} \quad X = \begin{bmatrix} -\mathbf{x}_1^T - \\ -\mathbf{x}_2^T - \\ -\mathbf{x}_3^T - \\ -\mathbf{x}_4^T - \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

○ ?  
× ○

? cannot be ×

$$\mathbf{w}^T \mathbf{x}_4 = \underbrace{\mathbf{w}^T \mathbf{x}_2}_{\circ} + \underbrace{\mathbf{w}^T \mathbf{x}_3}_{\circ} - \underbrace{\mathbf{w}^T \mathbf{x}_1}_{\times} > 0$$

linear dependence **restricts dichotomy**

$$d_{VC} \leq d + 1 \quad (1/2)$$

## $d$ -D General Case

$$X = \begin{bmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ & \vdots & \\ - & \mathbf{x}_{d+1}^T & - \\ - & \mathbf{x}_{d+2}^T & - \end{bmatrix}$$

more rows than columns:

linear dependence (some  $a_i$  non-zero)

$$\mathbf{x}_{d+2} = a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \dots + a_{d+1} \mathbf{x}_{d+1}$$

- can you generate  $(\text{sign}(a_1), \text{sign}(a_2), \dots, \text{sign}(a_{d+1}), \times)$ ? if so, what  $\mathbf{w}$ ?

$$\begin{aligned} \mathbf{w}^T \mathbf{x}_{d+2} &= a_1 \underbrace{\mathbf{w}^T \mathbf{x}_1}_0 + a_2 \underbrace{\mathbf{w}^T \mathbf{x}_2}_{\times} + \dots + a_{d+1} \underbrace{\mathbf{w}^T \mathbf{x}_{d+1}}_{\times} \\ &> 0 \text{ (contradiction!)} \end{aligned}$$

'general'  $X$  no-shatter  $\implies d_{VC} \leq d + 1$



# Fun Time

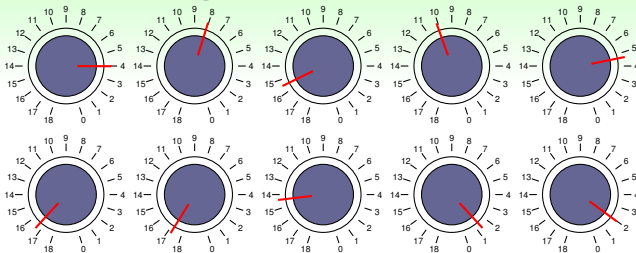
Based on the proof above, what is  $d_{VC}$  of 1126-D perceptrons?

- ① 1024
- ② 1126
- ③ 1127
- ④ 6211

Reference Answer: ③

Well, **too much fun for this section! :-)**

# Degrees of Freedom

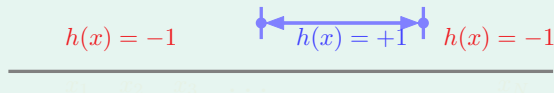


(modified from the work of Hugues Vermeiren on <http://www.texample.net>)

- hypothesis parameters  $\mathbf{w} = (w_0, w_1, \dots, w_d)$ :  
**creates degrees of freedom**
- hypothesis quantity  $M = |\mathcal{H}|$ :  
'analog' degrees of freedom
- hypothesis 'power'  $d_{\text{VC}} = d + 1$ :  
**effective 'binary' degrees of freedom**

$d_{\text{VC}}(\mathcal{H})$ : **powerfulness** of  $\mathcal{H}$

## Two Old Friends

Positive Rays ( $d_{VC} = 1$ )free parameters:  $a$ Positive Intervals ( $d_{VC} = 2$ )free parameters:  $\ell, r$ 

practical rule of thumb:

 $d_{VC} \approx \# \text{free parameters}$  (but not always)

$M$  and  $d_{VC}$ 

copied from Lecture 5 :-)

- ① can we make sure that  $E_{out}(g)$  is close enough to  $E_{in}(g)$ ?
- ② can we make  $E_{in}(g)$  small enough?

small  $M$ 

- ① Yes!,  
 $\mathbb{P}[\mathbf{BAD}] \leq 2 \cdot M \cdot \exp(\dots)$
- ② No!, too few choices

large  $M$ 

- ① No!,  
 $\mathbb{P}[\mathbf{BAD}] \leq 2 \cdot M \cdot \exp(\dots)$
- ② Yes!, many choices

small  $d_{VC}$ 

- ① Yes!,  $\mathbb{P}[\mathbf{BAD}] \leq$   
 $4 \cdot (2N)^{d_{VC}} \cdot \exp(\dots)$
- ② No!, too limited power

large  $d_{VC}$ 

- ① No!,  $\mathbb{P}[\mathbf{BAD}] \leq$   
 $4 \cdot (2N)^{d_{VC}} \cdot \exp(\dots)$
- ② Yes!, lots of power

using the right  $d_{VC}$  (or  $\mathcal{H}$ ) is important

# Fun Time

Origin-crossing Hyperplanes are essentially perceptrons with  $w_0$  fixed at 0. Make a guess about the  $d_{VC}$  of origin-crossing hyperplanes in  $\mathbb{R}^d$ .

- ① 1
- ②  $d$
- ③  $d + 1$
- ④  $\infty$

Reference Answer: ②

The proof is almost the same as proving the  $d_{VC}$  for usual perceptrons, but it is the **intuition** ( $d_{VC} \approx \# \text{free parameters}$ ) that you shall use to answer this quiz.

# VC Bound Rephrase: Penalty for Model Complexity

For any  $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$  and 'statistical' large  $\mathcal{D}$ , for  $N \geq 2, d_{VC} \geq 2$

$$\mathbb{P}_{\mathcal{D}} \left[ \underbrace{|E_{\text{in}}(g) - E_{\text{out}}(g)|}_{\text{BAD}} > \epsilon \right] \leq \underbrace{4(2N)^{d_{VC}} \exp\left(-\frac{1}{8}\epsilon^2 N\right)}_{\delta}$$

## Rephrase

..., with probability  $\geq 1 - \delta$ , **GOOD**:  $|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon$

$$\text{set } \delta = 4(2N)^{d_{VC}} \exp\left(-\frac{1}{8}\epsilon^2 N\right)$$

$$\frac{\delta}{4(2N)^{d_{VC}}} = \exp\left(-\frac{1}{8}\epsilon^2 N\right)$$

$$\ln\left(\frac{4(2N)^{d_{VC}}}{\delta}\right) = \frac{1}{8}\epsilon^2 N$$

$$\sqrt{\frac{8}{N} \ln\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)} = \epsilon$$

# VC Bound Rephrase: Penalty for Model Complexity

For any  $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$  and 'statistical' large  $\mathcal{D}$ , for  $N \geq 2, d_{vc} \geq 2$

$$\mathbb{P}_{\mathcal{D}} \left[ \underbrace{|E_{in}(g) - E_{out}(g)|}_{\text{BAD}} > \epsilon \right] \leq \underbrace{4(2N)^{d_{vc}} \exp\left(-\frac{1}{8}\epsilon^2 N\right)}_{\delta}$$

## Rephrase

..., with probability  $\geq 1 - \delta$ , **GOOD!**

$$\text{gen. error } |E_{in}(g) - E_{out}(g)| \leq \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{vc}}}{\delta} \right)}$$

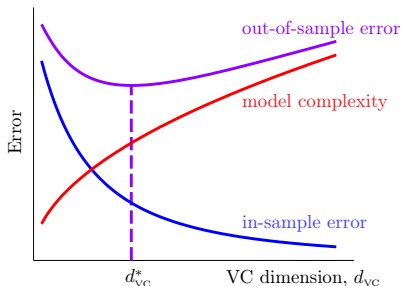
$$E_{in}(g) - \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{vc}}}{\delta} \right)} \leq E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{vc}}}{\delta} \right)}$$

$$\underbrace{\sqrt{\dots}}_{\Omega(N, \mathcal{H}, \delta)} : \text{penalty for model complexity}$$

# THE VC Message

with a high probability,

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \underbrace{\sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)}}_{\Omega(N, \mathcal{H}, \delta)}$$



- $d_{\text{VC}} \uparrow$ :  $E_{\text{in}} \downarrow$  but  $\Omega \uparrow$
- $d_{\text{VC}} \downarrow$ :  $\Omega \downarrow$  but  $E_{\text{in}} \uparrow$
- best  $d_{\text{VC}}^*$  in the middle

powerful  $\mathcal{H}$  not always good!



# VC Bound Rephrase: Sample Complexity

For any  $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$  and ‘statistical’ large  $\mathcal{D}$ , for  $N \geq 2, d_{VC} \geq 2$

$$\mathbb{P}_{\mathcal{D}} \left[ \underbrace{|E_{\text{in}}(g) - E_{\text{out}}(g)|}_{\text{BAD}} > \epsilon \right] \leq \underbrace{4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)}_{\delta}$$

given specs  $\epsilon = 0.1, \delta = 0.1, d_{VC} = 3$ , want  $4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \leq \delta$

$N$	bound
100	$2.82 \times 10^7$
1,000	$9.17 \times 10^9$
10,000	$1.19 \times 10^8$
100,000	$1.65 \times 10^{-38}$
29,300	$9.99 \times 10^{-2}$

sample complexity:

need  $N \approx 10,000 d_{VC}$  in theory

practical rule of thumb:

$N \approx 10 d_{VC}$  often enough!

# Looseness of VC Bound

$$\mathbb{P}_{\mathcal{D}} \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \leq 4(2N)^{d_{\text{vc}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$

theory:  $N \approx 10,000 d_{\text{vc}}$ ; practice:  $N \approx 10 d_{\text{vc}}$

## Why?

- Hoeffding for unknown  $E_{\text{out}}$  **any distribution, any target**
- $m_{\mathcal{H}}(N)$  instead of  $|\mathcal{H}(\mathbf{x}_1, \dots, \mathbf{x}_N)|$  **'any' data**
- $N^{d_{\text{vc}}}$  instead of  $m_{\mathcal{H}}(N)$  **'any'  $\mathcal{H}$  of same  $d_{\text{vc}}$**
- union bound on worst cases **any choice made by  $\mathcal{A}$**

— **but hardly better, and 'similarly loose for all models'**

**philosophical message** of VC bound  
important for improving ML

# Fun Time

Consider the VC Bound below. How can we decrease the probability of getting **BAD** data?

$$\mathbb{P}_{\mathcal{D}} \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \leq 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$

- ❶ decrease model complexity  $d_{\text{VC}}$
- ❷ increase data size  $N$  a lot
- ❸ increase generalization error tolerance  $\epsilon$
- ❹ all of the above

Reference Answer: ❹

**Congratulations on being  
Master of VC bound! :-)**

# Summary

- 1 When Can Machines Learn?
- 2 **Why** Can Machines Learn?

## Lecture 6: Theory of Generalization

## Lecture 7: The VC Dimension

- Definition of VC Dimension

**maximum non-break point**

- VC Dimension of Perceptrons

$$d_{VC}(\mathcal{H}) = d + 1$$

- Physical Intuition of VC Dimension

$$d_{VC} \approx \# \text{free parameters}$$

- Interpreting VC Dimension

**loosely: model complexity & sample complexity**

- **next: more than noiseless binary classification?**

- 3 How Can Machines Learn?
- 4 How Can Machines Learn Better?