

Machine Learning

(機器學習)

Lecture 4: Theory of Generalization

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 4: Theory of Generalization

- Effective Number of Lines
- Effective Number of Hypotheses
- Break Point
- Definition of VC Dimension
- VC Dimension of Perceptrons
- Physical Intuition of VC Dimension
- Interpreting VC Dimension

Is $M = \infty$ Feasible?

- input $x \in [-1, +1] \subset \mathbb{R}^1$, uniform iid
- target $f(x) = \text{sign}(x)$, taking $\text{sign}(0) = +1$
- hypothesis set: $h_a(x) = \text{sign}(x - a)$ for $a \in [-1, 1]$
infinitely many a
- algorithm: $g = h_{a^*}$ with $a^* = \min_{y_n = +1} x_n$,
assuming at least one $y_n = 1$
- for any $\epsilon < 0.5$, $E_{\text{out}}(g) > \epsilon$ iff every $y_n = +1$ satisfies $x_n > 2\epsilon$

$$\mathbb{P} \left[\underbrace{|E_{\text{in}}(g) - E_{\text{out}}(g)|}_0 > \epsilon \right] \leq \left(\frac{2 - 2\epsilon}{2} \right)^N$$

BAD data can happen rarely
even for **infinitely many hypotheses**

Where Did M Come From?

$$\mathbb{P} \left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \leq 2 \cdot M \cdot \exp \left(-2\epsilon^2 N \right)$$

- **BAD events** \mathcal{B}_m : $|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon$
- to give \mathcal{A} freedom of choice: bound $\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \dots \mathcal{B}_M]$
- worst case: all \mathcal{B}_m non-overlapping

$$\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \dots \mathcal{B}_M] \underbrace{\leq}_{\text{union bound}} \mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \dots + \mathbb{P}[\mathcal{B}_M]$$

where did **union bound fail**
to consider for $M = \infty$?

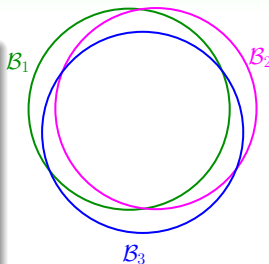
Where Did Union Bound Fail?

union bound $\mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \dots + \mathbb{P}[\mathcal{B}_M]$

- **BAD events** \mathcal{B}_m : $|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon$

overlapping for similar hypotheses $h_1 \approx h_2$
(e.g. if $a_1 \approx a_2$ in previous example)

- why? ① $E_{\text{out}}(h_1) \approx E_{\text{out}}(h_2)$
② for most \mathcal{D} , $E_{\text{in}}(h_1) = E_{\text{in}}(h_2)$
- union bound **over-estimating**

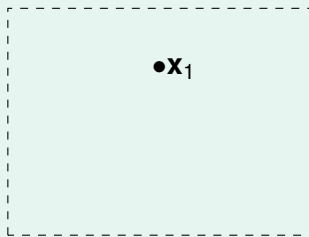


to account for overlap,
can we group similar hypotheses by **kind**?

How Many Lines Are There? (1/2)

$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

- how many lines? ∞
- how many **kinds of** lines if viewed from one input vector \mathbf{x}_1 ?

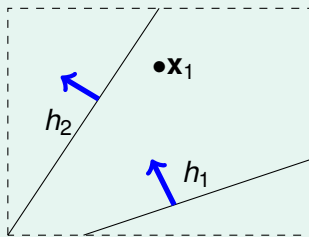


2 kinds: $h_1\text{-like}(\mathbf{x}_1) = \circ$ or $h_2\text{-like}(\mathbf{x}_1) = \times$

How Many Lines Are There? (1/2)

$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

- how many lines? ∞
- how many **kinds of** lines if viewed from one input vector \mathbf{x}_1 ?

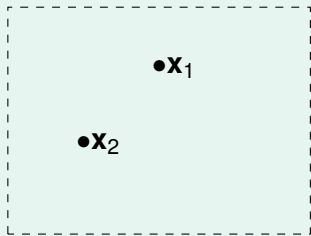


2 kinds: h_1 -like(\mathbf{x}_1) = \circ or h_2 -like(\mathbf{x}_1) = \times

How Many Lines Are There? (2/2)

$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

- how many **kinds of** lines if viewed from two inputs $\mathbf{x}_1, \mathbf{x}_2$?



4:

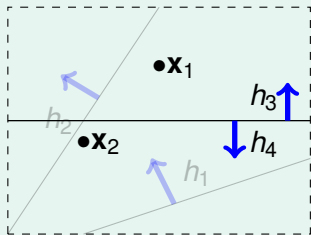


one input: 2; two inputs: 4; **three inputs?**

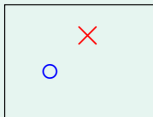
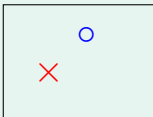
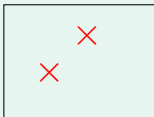
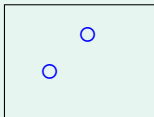
How Many Lines Are There? (2/2)

$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

- how many **kinds of** lines if viewed from two inputs $\mathbf{x}_1, \mathbf{x}_2$?



4:

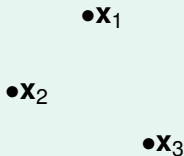


one input: 2; two inputs: 4; **three inputs?**

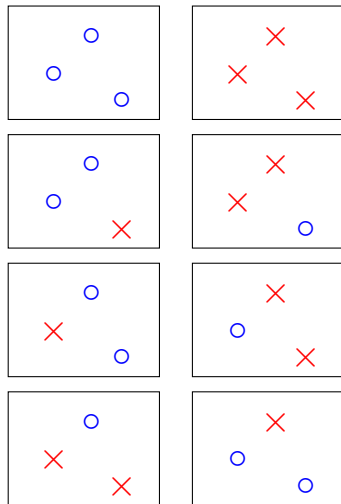
How Many Kinds of Lines for Three Inputs? (1/2)

$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

for three inputs $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$



8:



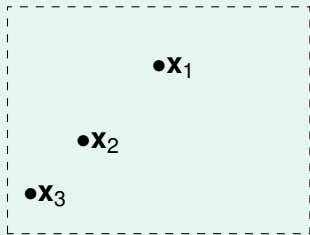
always 8 for three inputs?

How Many Kinds of Lines for Three Inputs? (2/2)

$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

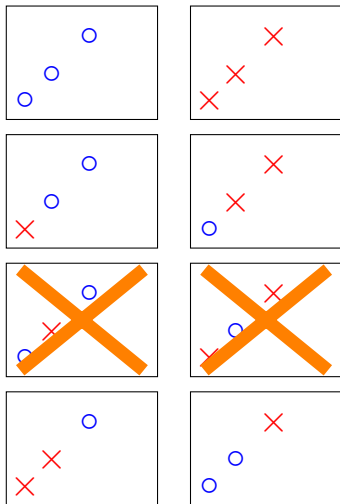
for **another** three inputs

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$



‘fewer than 8’ when degenerate
(e.g. collinear or same inputs)

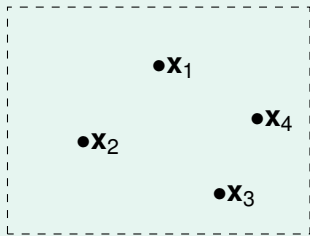
6:



How Many Kinds of Lines for Four Inputs?

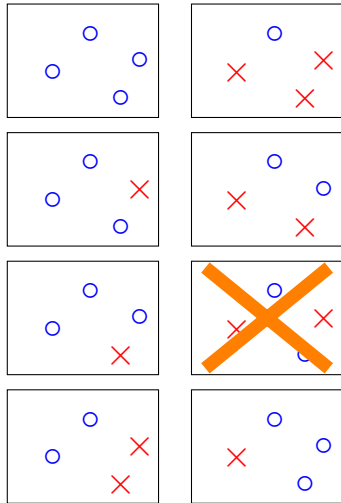
$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

for four inputs $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$



for any four inputs
at most 14

14: $2 \times$



Effective Number of Lines

maximum kinds of lines with respect to N inputs $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$
 \iff **effective number of lines**

- must be $\leq 2^N$ (why?)
- finite 'grouping' of infinitely-many lines $\in \mathcal{H}$
- wish:

$$\begin{aligned} & \mathbb{P} \left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \\ & \leq 2 \cdot \text{effective}(N) \cdot \exp \left(-2\epsilon^2 N \right) \end{aligned}$$

lines in 2D

N	effective(N)
1	2
2	4
3	8
4	14 $< 2^N$

- if ① effective(N) can replace M and
 ② effective(N) $\ll 2^N$

learning possible with infinite lines :-)

Questions?

Dichotomies: Mini-hypotheses

$$\mathcal{H} = \{\text{hypothesis } h: \mathcal{X} \rightarrow \{\times, \circ\}\}$$

- call

$$h(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = (h(\mathbf{x}_1), h(\mathbf{x}_2), \dots, h(\mathbf{x}_N)) \in \{\times, \circ\}^N$$

a **dichotomy**: hypothesis ‘limited’ to the eyes of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$

- $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$:

all dichotomies ‘implemented’ by \mathcal{H} on $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$

	hypotheses \mathcal{H}	dichotomies $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$
e.g.	all lines in \mathbb{R}^2	$\{\circ\circ\circ\circ, \circ\circ\circ\times, \circ\circ\times\times, \dots\}$
size	possibly infinite	upper bounded by 2^N

$|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)|$: candidate for **replacing M**

Growth Function

- $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)|$: depend on inputs $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$
- growth function:
remove dependence by **taking max of all possible** $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)|$$

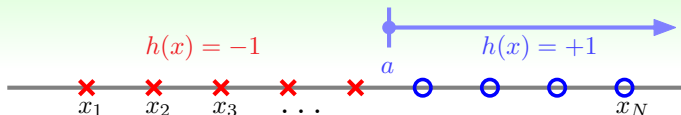
- finite, upper-bounded by 2^N

lines in 2D

N	$m_{\mathcal{H}}(N)$
1	2
2	4
3	$\max(\dots, 6, 8)$ $= 8$
4	$14 < 2^N$

how to 'calculate' the growth function?

Growth Function for Positive Rays



- $\mathcal{X} = \mathbb{R}$ (one dimensional)
- \mathcal{H} contains h , where **each** $h(x) = \text{sign}(x - a)$ **for threshold** a
- 'positive half' of 1D perceptrons

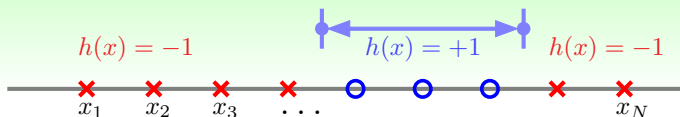
one dichotomy for $a \in$ each spot (x_n, x_{n+1}) :

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

$$(N + 1) \ll 2^N \text{ when } N \text{ large!}$$

x_1	x_2	x_3	x_4
○	○	○	○
×	○	○	○
×	×	○	○
×	×	×	○
×	×	×	×

Growth Function for Positive Intervals



- $\mathcal{X} = \mathbb{R}$ (one dimensional)
- \mathcal{H} contains h , where **each** $h(x) = +1$ **iff** $x \in [\ell, r)$, **-1 otherwise**

one dichotomy for each 'interval kind'

$$\begin{aligned}
 m_{\mathcal{H}}(N) &= \underbrace{\binom{N+1}{2}}_{\text{interval ends in } N+1 \text{ spots}} + \underbrace{1}_{\text{all } \times} \\
 &= \frac{1}{2}N^2 + \frac{1}{2}N + 1
 \end{aligned}$$

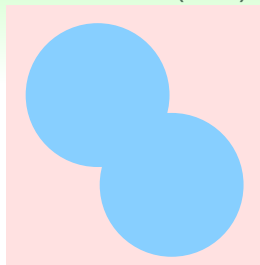
x_1	x_2	x_3	x_4
o	x	x	x
o	o	x	x
o	o	o	x
o	o	o	o
x	o	x	x
x	o	o	x
x	o	o	o
x	x	o	x
x	x	o	o
x	x	x	o
x	x	x	x

$$\left(\frac{1}{2}N^2 + \frac{1}{2}N + 1\right) \ll 2^N \text{ when } N \text{ large!}$$

Growth Function for Convex Sets (1/2)



convex region in blue



non-convex region

- $\mathcal{X} = \mathbb{R}^2$ (two dimensional)
- \mathcal{H} contains h , where $h(\mathbf{x}) = +1$ iff \mathbf{x} in a convex region, -1 otherwise

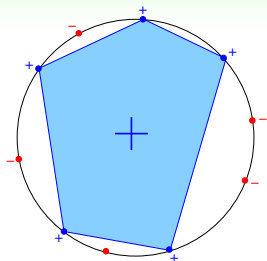
what is $m_{\mathcal{H}}(N)$?

Growth Function for Convex Sets (2/2)

- one possible set of N inputs:
 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ on a big circle
- **every dichotomy can be implemented** by \mathcal{H} using a convex region slightly extended from **contour of positive inputs**

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

- call those N inputs **'shattered'** by \mathcal{H}



shatter some N inputs by $\mathcal{H} \Leftrightarrow m_{\mathcal{H}}(N) = 2^N$;
not shatter any N inputs by $\mathcal{H} \Leftrightarrow m_{\mathcal{H}}(N) < 2^N$

The Four Growth Functions

- positive rays:

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

- positive intervals:

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

- convex sets:

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

- 2D perceptrons:

$$m_{\mathcal{H}}(N) < 2^N \text{ in some cases}$$

what if $m_{\mathcal{H}}(N)$ replaces M ?

$$\mathbb{P} \left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \stackrel{?}{\leq} 2 \cdot m_{\mathcal{H}}(N) \cdot \exp \left(-2\epsilon^2 N \right)$$

polynomial: good; exponential: bad

for 2D or general perceptrons,

$m_{\mathcal{H}}(N)$ **polynomial?**

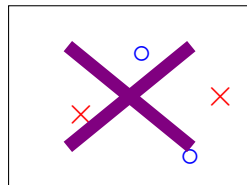
Break Point of \mathcal{H}

what do we know about 2D perceptrons now?

three inputs: 'exists' shatter;
four inputs, 'for all' no shatter

if no k inputs can be shattered by \mathcal{H} ,
call k a **break point** for \mathcal{H}

- $m_{\mathcal{H}}(k) < 2^k$
- $k + 1, k + 2, k + 3, \dots$ also break points!
- will study **minimum break point** k



2D perceptrons: **minimum break point at 4**

The Four Minimum Break Points

- positive rays: $m_{\mathcal{H}}(N) = N + 1 = O(N)$
minimum break point at 2
- positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 = O(N^2)$
minimum break point at 3
- convex sets: $m_{\mathcal{H}}(N) = 2^N$
no break point
- 2D perceptrons: $m_{\mathcal{H}}(N) < 2^N$ in some cases
minimum break point at 4

theorem from combinatorics
(not going to prove in class):

- no break point: $m_{\mathcal{H}}(N) = 2^N$ (sure!)
- minimum break point k :
 $m_{\mathcal{H}}(N) = O(N^{k-1})$

Questions?

BAD Bound for General \mathcal{H}

want:

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2 m_{\mathcal{H}}(N) \cdot \exp\left(-2 \epsilon^2 N\right)$$

actually, when N large enough,

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2 \cdot 2 m_{\mathcal{H}}(2N) \cdot \exp\left(-2 \cdot \frac{1}{16} \epsilon^2 N\right)$$

called Vapnik-Chervonenkis (VC) Bound

Interpretation of 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})
 \implies probably generalized ' $E_{\text{out}} \approx E_{\text{in}}$ ', and
 if ③ \mathcal{A} picks a g with small E_{in} (good \mathcal{A})
 \implies probably learned! (:-) good luck

VC Dimension

the formal name of **maximum non**-break point d_{VC}
= (minimum break point $k - 1$)

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)

$N \leq d_{VC} \implies \mathcal{H}$ 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:

$$d_{VC} = 3$$



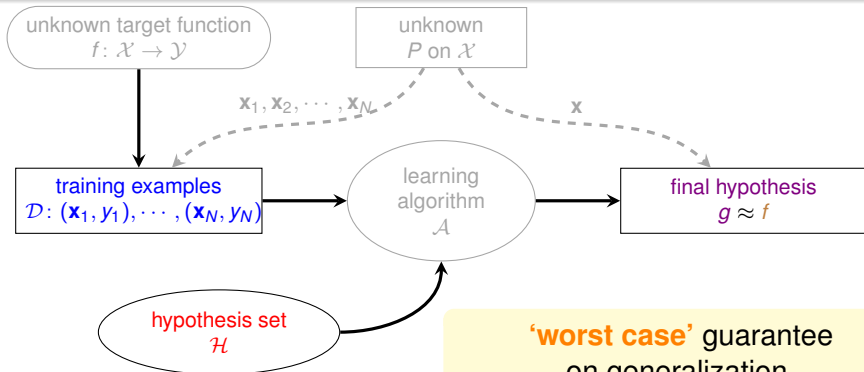
$$m_{\mathcal{H}}(N) \leq N^3 \text{ for } N \geq 2$$

good: **finite** d_{VC}

VC Dimension and Learning

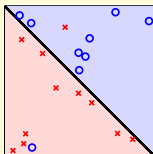
finite $d_{\text{VC}} \implies g$ 'will' generalize ($E_{\text{out}}(g) \approx E_{\text{in}}(g)$)

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



'worst case' guarantee
on generalization

From Noiseless VC to Noisy VC



real-world learning problems are often **noisy**

age	23 years
gender	female
annual salary	NTD 1,000,000
year in residence	1 year
year in job	0.5 year
current debt	200,000

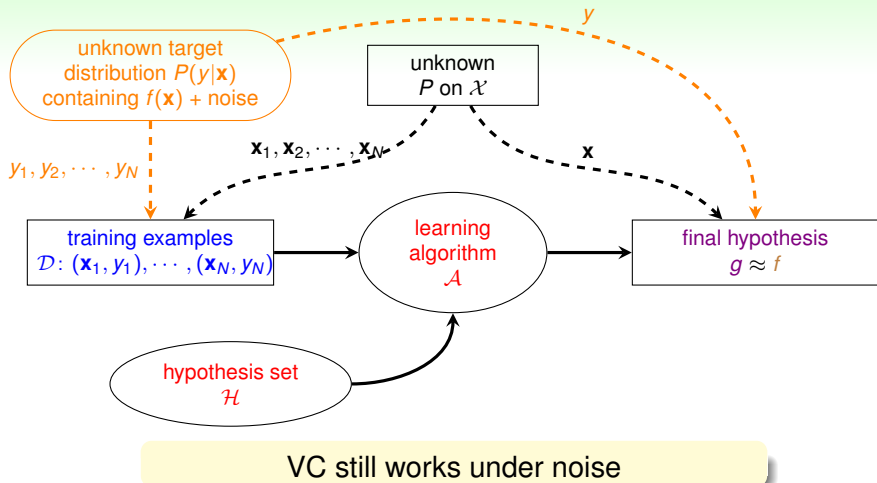
credit? {no(-1), yes(+1)}

but more!

- **noise in \mathbf{x}** (covered by $P(\mathbf{x})$): inaccurate customer information?
- **noise in y** (covered by $P(y|\mathbf{x})$): good customer, 'misabeled' as bad?

does VC bound work under **noise**? Yes!

The New Learning Flow



Questions?

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

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

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**

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

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)

$$\mathbf{x}_{d+2} = \textcolor{blue}{a}_1 \mathbf{x}_1 + \textcolor{red}{a}_2 \mathbf{x}_2 + \dots + \textcolor{red}{a}_{d+1} \mathbf{x}_{d+1}$$

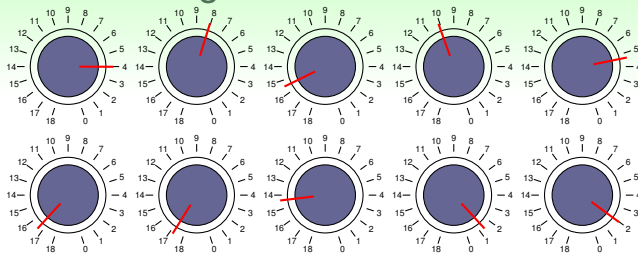
- can you generate $(\textcolor{blue}{\text{sign}}(a_1), \textcolor{red}{\text{sign}}(a_2), \dots, \textcolor{red}{\text{sign}}(a_{d+1}), \times)$? if so, what \mathbf{w} ?

$$\begin{aligned} \mathbf{w}^T \mathbf{x}_{d+2} &= \textcolor{blue}{a}_1 \underbrace{\mathbf{w}^T \mathbf{x}_1}_{\circ} + \textcolor{red}{a}_2 \underbrace{\mathbf{w}^T \mathbf{x}_2}_{\times} + \dots + \textcolor{red}{a}_{d+1} \underbrace{\mathbf{w}^T \mathbf{x}_{d+1}}_{\times} \\ &> \textcolor{violet}{0}(\text{contradiction!}) \end{aligned}$$

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

Questions?

Degrees of Freedom

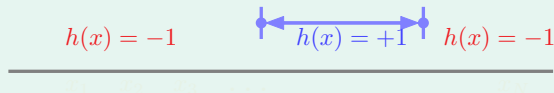


(modified from the work of Hugues Vermeiren on <http://www.texample.net>)

- 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' $d_{VC} = d + 1$:
effective 'binary' degrees of freedom

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

Two Old Friends

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

practical rule of thumb:

$d_{VC} \approx \text{\#free parameters}$ (but not always, e.g.,
mystery about deep learning models)

M and d_{VC}

copied from Lecture 3 :-)

- 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

Questions?

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 ~~$N \geq 2$~~ , $d_{VC} \geq 2$

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

Rephrase

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

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

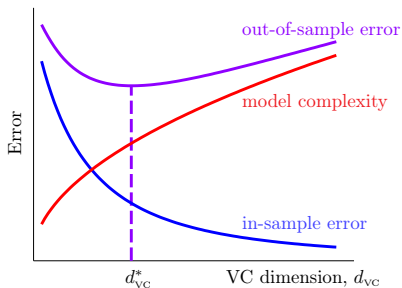
$$E_{\text{in}}(g) - \sqrt{\frac{8}{N} \ln \left(\frac{4(2N)^{d_{VC}}}{\delta} \right)} \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \left(\frac{4(2N)^{d_{VC}}}{\delta} \right)}$$

$\underbrace{\sqrt{\dots}}_{\Omega(N, \mathcal{H}, \delta)}$: 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_{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 ~~$N \geq 2$~~ , $d_{VC} \geq 2$

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

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	2.82×10^7
1,000	9.17×10^9
10,000	1.19×10^8
100,000	1.65×10^{-38}
29,300	9.99×10^{-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_{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_{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

Questions?

Summary

1 When Can Machines Learn?

Lecture 3: Feasibility of Learning

2 Why Can Machines Learn?

Lecture 4: Theory of Generalization

- Effective Number of Lines
at most 14 through the eye of 4 inputs
- Effective Number of Hypotheses
at most $m_{\mathcal{H}}(N)$ through the eye of N inputs
- Break Point
when $m_{\mathcal{H}}(N)$ becomes 'non-exponential'
- Definition of VC Dimension
maximum non-break point
- VC Dimension of Perceptrons
 $d_{\text{VC}}(\mathcal{H}) = d + 1$
- Physical Intuition of VC Dimension
 $d_{\text{VC}} \approx \# \text{free parameters}$
- Interpreting VC Dimension
loosely: model complexity & sample complexity

- **next: beyond VC theory, please :-)**