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 $a \in [-1, 1]$
  infintely many a
- algorithm: $g = h_{a^*}$ with $a^* = \min_{y_n=+1} x_n$,
  assuming at least one $y_n = 1$

- for $\epsilon < 0.5$, $E_{\text{out}}(g) > \epsilon$ if every $y_n = +1$ satisfies $x_n > 2\epsilon$

$$\mathbb{P}\left[\left|\underbrace{E_{\text{in}}(g)}_{0} - E_{\text{out}}(g)\right| > \epsilon\right] \leq \left(\frac{2 - 2\epsilon}{2}\right)^N$$

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

Where Did M Come From?

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

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

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

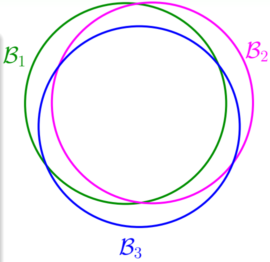
where did **union bound fail**
to consider for $M = \infty$?

Where Did Union Bound Fail?

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

- **$\mathcal{B}$AD 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**

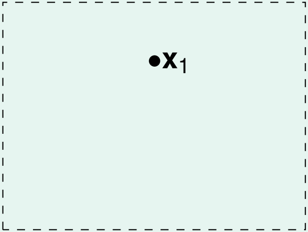


to account for overlap,
can we group similar hypotheses by **kind**?

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 $\mathbf{x}_1$?

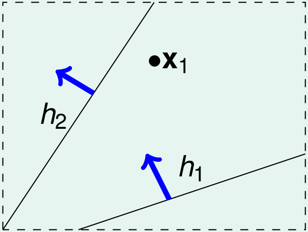


2 kinds: h_1-like$(\mathbf{x}_1) = \circ$ or h_2-like$(\mathbf{x}_1) = \times$

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 $\mathbf{x}_1$?

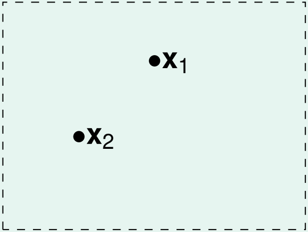


2 kinds: h_1-like$(\mathbf{x}_1) = \circ$ or h_2-like$(\mathbf{x}_1) = \times$

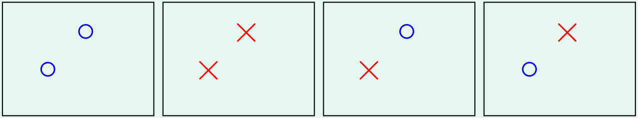
How Many Lines Are There? (2/2)

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

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



4:

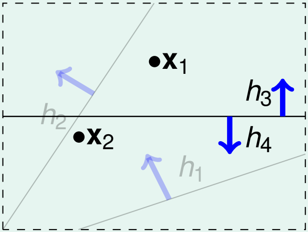


one input: 2; two inputs: 4; **three inputs?**

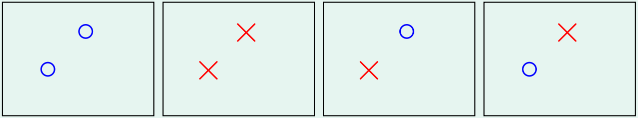
How Many Lines Are There? (2/2)

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

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



4:

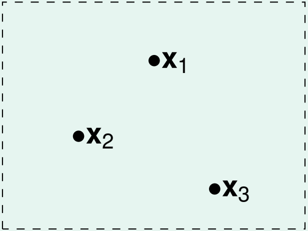


one input: 2; two inputs: 4; **three inputs?**

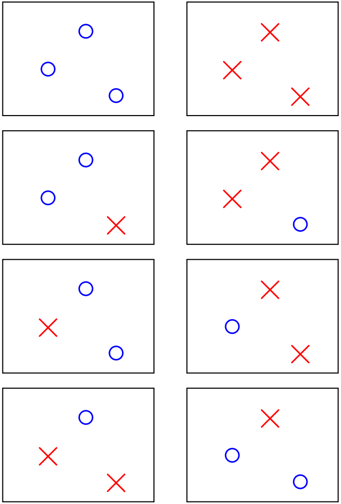
How Many Kinds of Lines for Three Inputs? (1/2)

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

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



8:

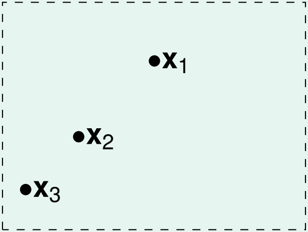


always 8 **for three inputs**?

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

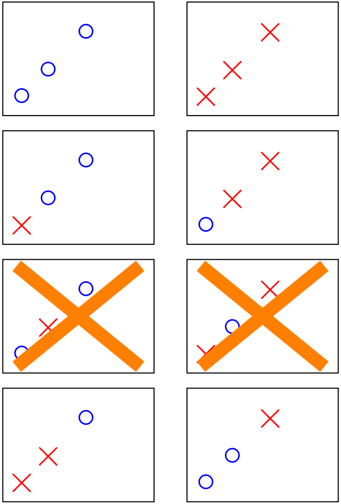
$$\mathcal{H} = \left\{\text{all lines in } \mathbb{R}^2\right\}$$

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



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

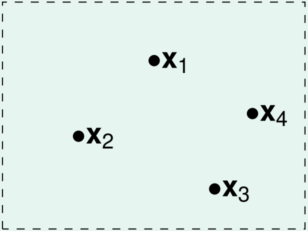
6:



How Many Kinds of Lines for Four Inputs?

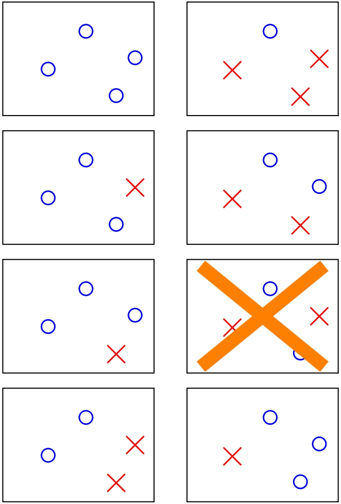
$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

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



for any four inputs
at most 14

14: 2×



Effective Number of Lines

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

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

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

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

$$\mathcal{H} = \{\text{hypothesis } h \colon \mathcal{X} \to \{\times, \circ\}\}$$

- call

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

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

- $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$:
  all dichotomies 'implemented' by $\mathcal{H}$ on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

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

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

Growth Function

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

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

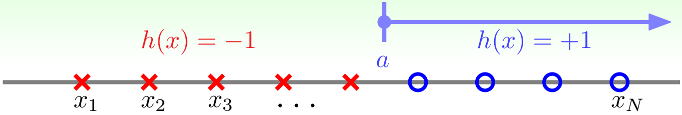
- finite, upper-bounded by 2^N

lines in 2D

N	$m_{\mathcal{H}}(N)$
1	2
2	4
3	$\max(\ldots, 6, 8) = 8$
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

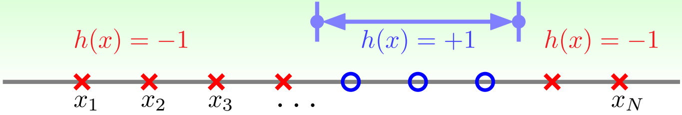
one dichotomy for $a \in$ each spot (x_n, x_{n+1}):

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

$(N+1) \ll 2^N$ when N 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'

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

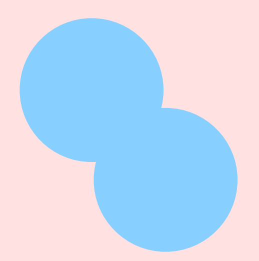
$$= \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

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

x_1	x_2	x_3	x_4
◦	×	×	×
◦	◦	×	×
◦	◦	◦	×
◦	◦	◦	◦
×	◦	×	×
×	◦	◦	×
×	◦	◦	◦
×	×	◦	×
×	×	◦	◦
×	×	×	◦
×	×	×	×

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**

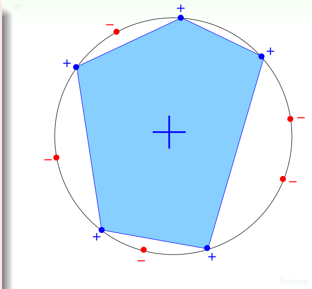
what is $m_{\mathcal{H}}(N)$?

Growth Function for Convex Sets (2/2)

- one possible set of N inputs: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ 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 $\mathcal{H}$**



$m_{\mathcal{H}}(N) = 2^N \iff$
exists N inputs that can be shattered

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

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

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

polynomial: good; **exponential: bad**

for 2D or general perceptrons,
$m_{\mathcal{H}}(N)$ **polynomial**?

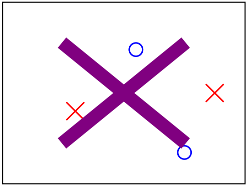
Break Point of $\mathcal{H}$

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 $\mathcal{H}$

- $m_{\mathcal{H}}(k) < 2^k$
- $k + 1$, $k + 2$, $k + 3$, ... 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:

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

actually, when N large enough,

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

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 ~~$N \geq 2$,~~ $k \geq 3$

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

if (1) $m_{\mathcal{H}}(N)$ breaks at k (**good** $\mathcal{H}$)
(2) N large enough (**good** $\mathcal{D}$)
$\Longrightarrow$ probably generalized '$E_{\text{out}} \approx E_{\text{in}}$', and
if (3) $\mathcal{A}$ picks a g with small E_{in} (**good** $\mathcal{A}$)
$\Longrightarrow$ 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 $d_{\mathrm{VC}}(\mathcal{H})$ is

largest N for which $m_{\mathcal{H}}(N) = 2^N$
(the **most** inputs $\mathcal{H}$ that can shatter)

$N \le d_{\mathrm{VC}} \implies$ $\mathcal{H}$ can shatter some N inputs
$k > d_{\mathrm{VC}} \implies$ k is a break point for $\mathcal{H}$

if $N \ge 2, d_{\mathrm{VC}} \ge 2$, $m_{\mathcal{H}}(N) \le N^{d_{\mathrm{VC}}}$

The Four VC Dimensions

- positive rays: $m_{\mathcal{H}}(N) = N + 1$

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

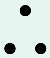
- positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

  $d_{\text{VC}} = 2$ 

- convex sets: $m_{\mathcal{H}}(N) = 2^N$

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

- 2D perceptrons: $m_{\mathcal{H}}(N) \le N^3$ for $N \ge 2$

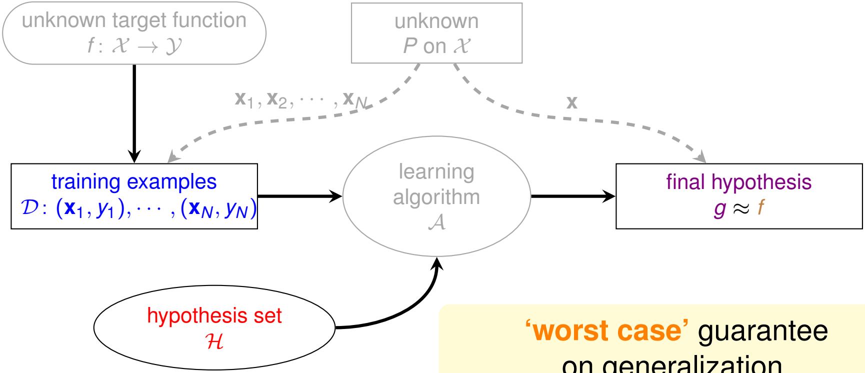
  $d_{\text{VC}} = 3$ 

good: **finite** d_{VC}

VC Dimension and Learning

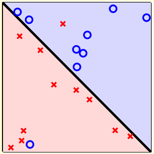
finite $d_{\mathrm{VC}} \Longrightarrow g$ 'will' generalize ($E_{\mathrm{out}}(g) \approx E_{\mathrm{in}}(g)$)

- 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? $\{\text{no}(-1), \text{yes}(+1)\}$

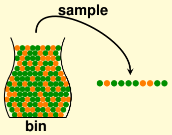
but more!

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

does VC bound work under **noise**?

Probabilistic Marbles

one key of VC bound: **marbles**!



'deterministic' marbles

- marble $\mathbf{x} \sim P(\mathbf{x})$
- deterministic color $[\![f(\mathbf{x}) \neq h(\mathbf{x})]\!]$

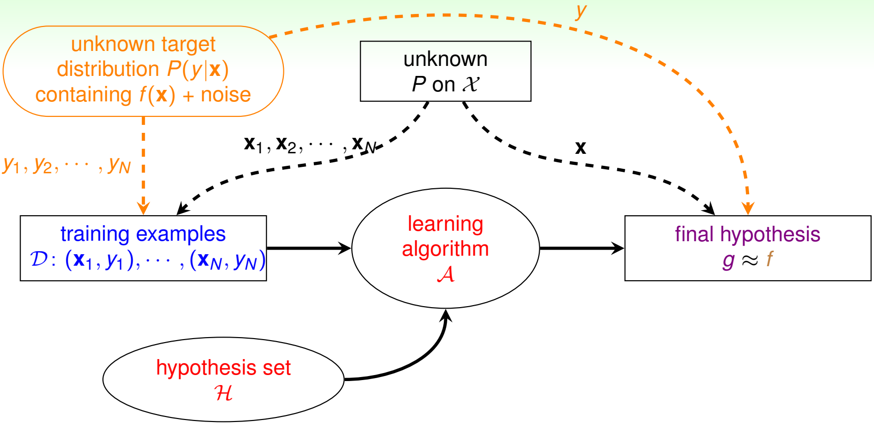
'probabilistic' (noisy) marbles

- marble $\mathbf{x} \sim P(\mathbf{x})$
- probabilistic color $[\![y \neq h(\mathbf{x})]\!]$ with $y \sim P(y|\mathbf{x})$

same nature: can estimate $\mathbb{P}$[orange] if $\overset{i.i.d.}{\sim}$

VC holds for $\underbrace{\mathbf{x} \overset{i.i.d.}{\sim} P(\mathbf{x}), y \overset{i.i.d.}{\sim} P(y|\mathbf{x})}_{(\mathbf{x},y) \overset{i.i.d.}{\sim} P(\mathbf{x},y)}$

The New Learning Flow



VC still works under noise

Questions?

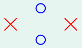
2D PLA Revisited

$$\begin{array}{ccc}
\text{linearly separable } \mathcal{D} & & \text{with } \mathbf{x}_n \sim P \text{ and } y_n = f(\mathbf{x}_n) \\
\downarrow & & \downarrow \\
\text{PLA can converge} & & \mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \ldots \ \boxed{\text{by } d_{\text{VC}} = 3} \\
T \text{ large} \downarrow & & \downarrow N \text{ large} \\
E_{\text{in}}(g) = 0 & & E_{\text{out}}(g) \approx E_{\text{in}}(g)
\end{array}$$

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

general PLA for **x** with more than 2 features?

VC Dimension of Perceptrons

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

two steps:

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

Extra Fun Time

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

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

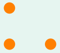
d_{VC} is the maximum that $m_{\mathcal{H}}(N) = 2^N$, and $m_{\mathcal{H}}(N)$ is the most number of dichotomies of N inputs. So if we can find 2^{d+1} dichotomies on *some* $d + 1$ inputs, $m_{\mathcal{H}}(d + 1) = 2^{d+1}$ and hence $d_{\text{VC}} \geq d + 1$.

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

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

- some 'trivial' inputs:

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

- visually in 2D: 

note: X **invertible!**

Can We Shatter X?

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

to shatter ...

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

$$\text{sign}\left(\mathrm{X}\mathbf{w}\right) = \mathbf{y} \quad \Longleftarrow \quad \left(\mathrm{X}\mathbf{w}\right) = \mathbf{y} \quad \overset{\mathrm{X}\text{ invertible!}}{\Longleftrightarrow} \quad \mathbf{w} = \mathrm{X}^{-1}\mathbf{y}$$

'special' X can be shattered $\Longrightarrow d_{\mathrm{VC}} \geq d + 1$

Extra Fun Time

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

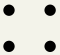
1. There are some $d + 1$ inputs we can shatter.
2. We can shatter any set of $d + 1$ inputs.
3. There are some $d + 2$ inputs we cannot shatter.
4. We cannot shatter any set of $d + 2$ inputs.

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 $d_{\mathrm{VC}} < d + 2$. That is, $d_{\mathrm{VC}} \le d + 1$.

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

A 2D Special Case



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

$$\begin{matrix} \circ & \textbf{?} \\ \times & \circ \end{matrix}$$

? cannot be $\times$

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

linear dependence **restricts dichotomy**

$d_{\text{VC}} \le d + 1$ (2/2)

d-D General Case

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

more rows than columns:

linear dependence (some a_i non-zero)

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

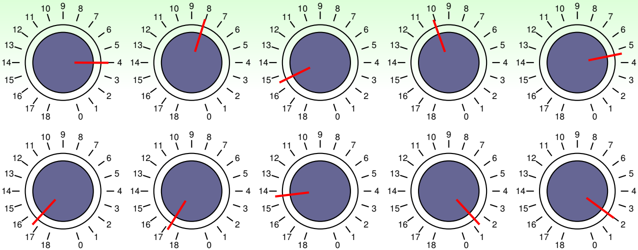
- can you generate $(\text{sign}(a_1), \text{sign}(a_2), \ldots, \text{sign}(a_{d+1}), \times)$? if so, what $\mathbf{w}$?

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

'general' X no-shatter $\Longrightarrow d_{\text{VC}} \le d + 1$

Questions?

Degrees of Freedom



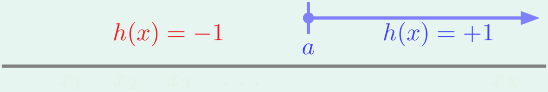
(modified from the work of Hugues Vermeiren on `http://www.texample.net`)

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

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

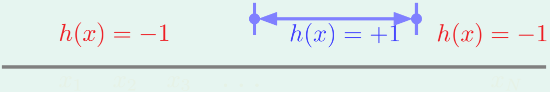
Two Old Friends

Positive Rays ($d_{\mathrm{VC}} = 1$)



free parameters: a

Positive Intervals ($d_{\mathrm{VC}} = 2$)



free parameters: ℓ, r

practical rule of thumb:

$d_{\mathrm{VC}} \approx$ #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 $E_{\mathrm{out}}(g)$ is close enough to $E_{\mathrm{in}}(g)$?
2. can we make $E_{\mathrm{in}}(g)$ small enough?

small M

1. Yes!, $\mathbb{P}[\textbf{BAD}] \le 2 \cdot M \cdot \exp(\ldots)$
2. No!, too few choices

large M

1. No!, $\mathbb{P}[\textbf{BAD}] \le 2 \cdot M \cdot \exp(\ldots)$
2. Yes!, many choices

small d_{VC}

1. Yes!, $\mathbb{P}[\textbf{BAD}] \le 4 \cdot (2N)^{d_{\mathrm{VC}}} \cdot \exp(\ldots)$
2. No!, too limited power

large d_{VC}

1. No!, $\mathbb{P}[\textbf{BAD}] \le 4 \cdot (2N)^{d_{\mathrm{VC}}} \cdot \exp(\ldots)$
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 ~~$N \geq 2$,~~ $d_{\mathrm{VC}} \geq 2$

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

Rephrase

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

$$\begin{aligned}
\text{set} \quad \delta &= 4(2N)^{d_{\mathrm{VC}}} \exp\left(-\tfrac{1}{8}\epsilon^2 N\right) \\
\frac{\delta}{4(2N)^{d_{\mathrm{VC}}}} &= \exp\left(-\tfrac{1}{8}\epsilon^2 N\right) \\
\ln\left(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta}\right) &= \tfrac{1}{8}\epsilon^2 N \\
\sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta}\right)} &= \epsilon
\end{aligned}$$

VC Bound Rephrase: Penalty for Model Complexity

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

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

Rephrase

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

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

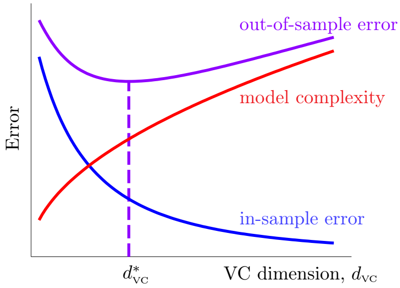
$$E_{\mathrm{in}}(g) - \sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta}\right)} \leq E_{\mathrm{out}}(g) \leq E_{\mathrm{in}}(g) + \sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta}\right)}$$

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

THE VC Message

with a high probability,

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



- $d_{\text{VC}} \uparrow$: $E_{\text{in}} \downarrow$ 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 ~~$N \geq 2$,~~ $d_{\text{VC}} \geq 2$

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

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

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

practical rule of thumb:

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

Looseness of VC Bound

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

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 $|\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)|$ — **'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
- Effective Number of Hypotheses
- Break Point
- Definition of VC Dimension
- VC Dimension of Perceptrons
- Physical Intuition of VC Dimension
- Interpreting VC Dimension

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