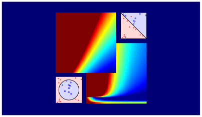
Machine Learning Foundations
(機器學習基石)



Lecture 5: Training versus Testing

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?

Lecture 4: Feasibility of Learning

learning is **PAC-**possible
if enough **statistical data** and **finite $|\mathcal{H}|$**

2. **Why** Can Machines Learn?

Lecture 5: Training versus Testing

- Recap and Preview
- Effective Number of Lines
- Effective Number of Hypotheses
- Break Point

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

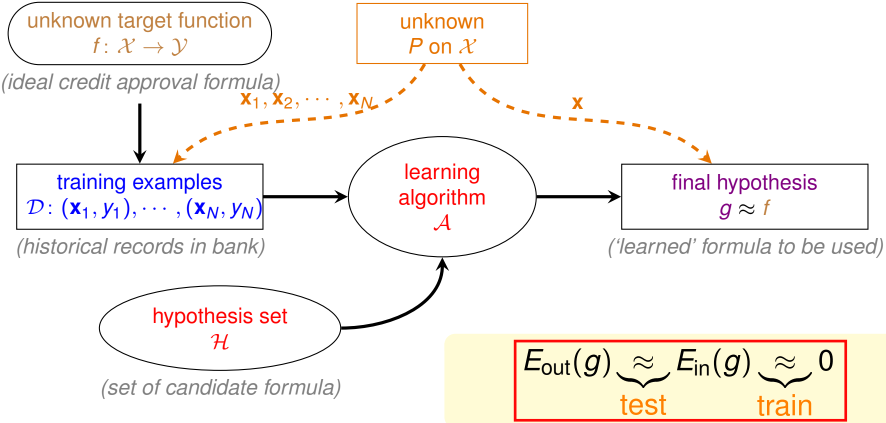
Recap: the ‘Statistical’ Learning Flow

if $|\mathcal{H}| = M$ finite, N large enough,

for whatever g picked by $\mathcal{A}$, $E_{\text{out}}(g) \approx E_{\text{in}}(g)$

if $\mathcal{A}$ finds one g with $E_{\text{in}}(g) \approx 0$,

PAC guarantee for $E_{\text{out}}(g) \approx 0 \Longrightarrow$ **learning possible :-)**


unknown target function
$f\colon \mathcal{X} \to \mathcal{Y}$
(ideal credit approval formula)
unknown
P on $\mathcal{X}$
$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$
$\mathbf{x}$
training examples
$\mathcal{D}\colon (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$
(historical records in bank)
learning algorithm
$\mathcal{A}$
final hypothesis
$g \approx f$
(‘learned’ formula to be used)
hypothesis set
$\mathcal{H}$
(set of candidate formula)

$$\boxed{E_{\text{out}}(g) \underbrace{\approx}_{\text{test}} E_{\text{in}}(g) \underbrace{\approx}_{\text{train}} 0}$$

Two Central Questions

for $\underbrace{\text{batch \& supervised binary classification}}_{\text{lecture 3}}$, $\underbrace{g \approx f}_{\text{lecture 1}} \Longleftrightarrow E_{\text{out}}(g) \approx 0$

achieved through $\underbrace{E_{\text{out}}(g) \approx E_{\text{in}}(g)}_{\text{lecture 4}}$ and $\underbrace{E_{\text{in}}(g) \approx 0}_{\text{lecture 2}}$

learning split to two central questions:

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

what role does $\underbrace{M}_{|\mathcal{H}|}$ play for the two questions?

Trade-off on M

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

small M

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

large M

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

using the right M (or $\mathcal{H}$) is important
$M = \infty$ **doomed?**

ex: PLA

Preview

Known

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

Todo

- establish **a finite quantity** that replaces M

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

- justify the feasibility of learning for infinite M
- study $m_{\mathcal{H}}$ to understand its trade-off for 'right' $\mathcal{H}$, just like M

mysterious PLA to be fully resolved
after 3 more lectures :-)

Fun Time

Data size: how large do we need?

One way to use the inequality

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

is to pick a tolerable difference ϵ as well as a tolerable **BAD** probability δ, and then gather data with size (N) large enough to achieve those tolerance criteria. Let $\epsilon = 0.1$, $\delta = 0.05$, and $M = 100$. What is the data size needed?

1. 215
2. 415
3. 615
4. 815

Fun Time

Data size: how large do we need?

One way to use the inequality

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

is to pick a tolerable difference ϵ as well as a tolerable **BAD** probability δ, and then gather data with size (N) large enough to achieve those tolerance criteria. Let $\epsilon = 0.1$, $\delta = 0.05$, and $M = 100$. What is the data size needed?

Max P

① 215 ② 415 ③ 615 ④ 815

Reference Answer: ②

We can simply express N as a function of those 'known' variables. Then, the needed $N = \frac{1}{2\epsilon^2} \ln \frac{2M}{\delta}$. = 414.7

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

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

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

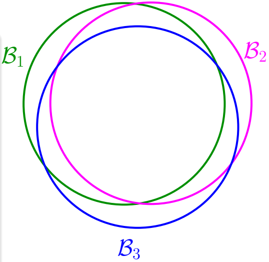
Where Did Union Bound Fail?

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 $h_1 \approx h_2$
- 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**

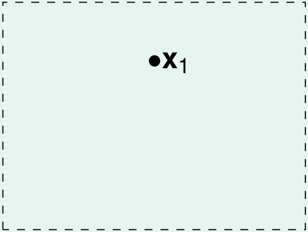


to account for overlap,
can we group similar hypotheses by **kind**?

How Many Lines Are There? (1/2)

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

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



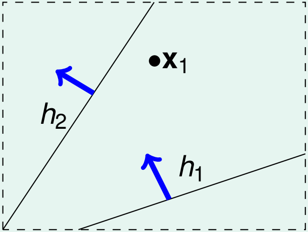
Use classification results to define the kinds of lines.

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 $\mathbf{x}_1$?

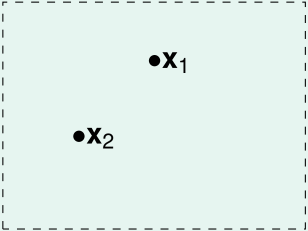


2 kinds: h_1-like$(\mathbf{x}_1) = \circ$ or h_2-like$(\mathbf{x}_1) = \times$

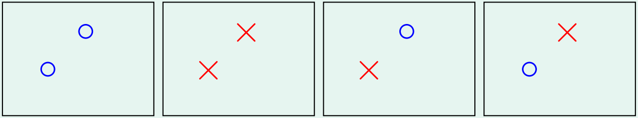
How Many Lines Are There? (2/2)

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

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



4:

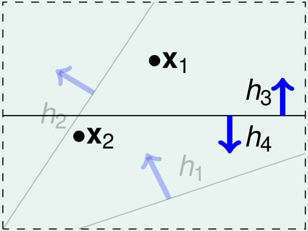


one input: 2; two inputs: 4; **three inputs?**

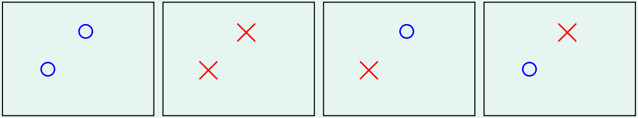
How Many Lines Are There? (2/2)

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

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



4:

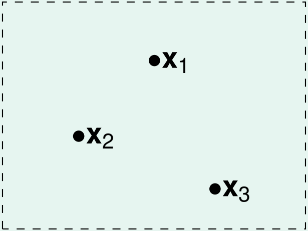


one input: 2; two inputs: 4; **three inputs?**

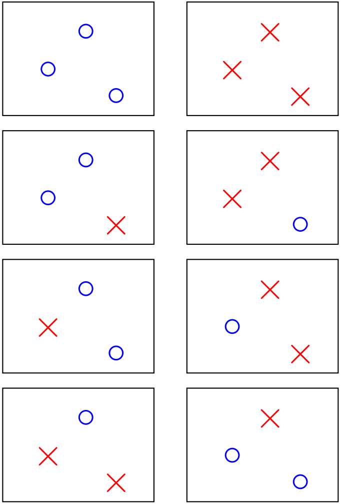
How Many Kinds of Lines for Three Inputs? (1/2)

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

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



8:

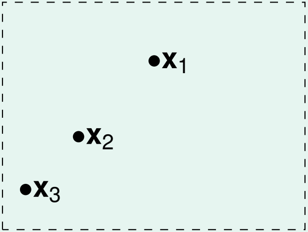


always 8 **for three inputs**?

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

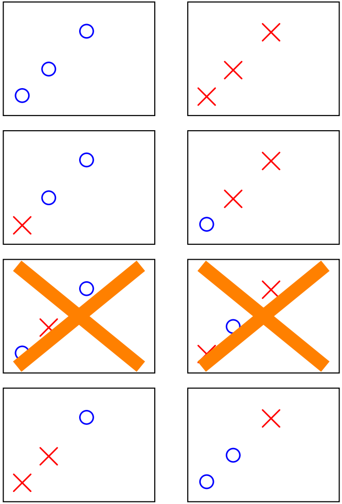
$$\mathcal{H} = \left\{\text{all lines in } \mathbb{R}^2\right\}$$

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



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

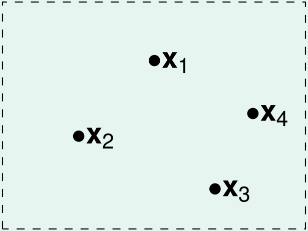
8 ↓ 6:



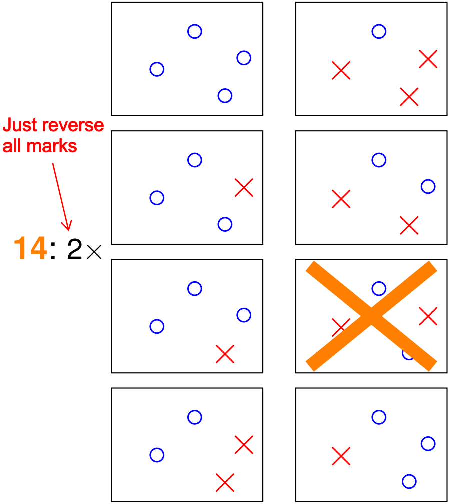
How Many Kinds of Lines for Four Inputs?

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

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



for any four inputs
at most 14



Effective Number of Lines

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

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

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

lines in 2D at most

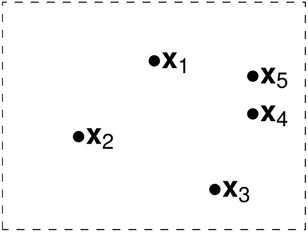
N	effective(N)
1	2
2	4
3	8
4	14 $< 2^N$

if $\textcircled{1}$ effective(N) can replace M and
$\textcircled{2}$ effective(N) $\ll 2^N$
learning possible with infinite lines :-)

Fun Time

What is the effective number of lines for five inputs $\in \mathbb{R}^2$?

1. 14
2. 16
3. 22
4. 32



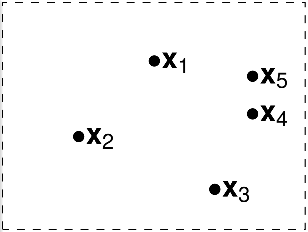
Fun Time

What is the effective number of lines for five inputs $\in \mathbb{R}^2$?

1. 14
2. 16
3. 22
4. 32

Reference Answer: ③

If you put the inputs roughly around a circle, you can then pick any consecutive inputs to be on one side of the line, and the other inputs to be on the other side. The procedure leads to effectively 22 kinds of lines, which is **much smaller than $2^5 = 32$**. You shall find it difficult to generate more kinds by varying the inputs, and we will give a formal proof in future lectures.



Dichotomies: Mini-hypotheses

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

- call

  binary classification results for each x

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

  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 $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$

  not good in the future

- growth function:

  remove dependence by **taking** max **of all possible** $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$

  The reasons behind the naming:

  1. It is a number that is smaller than M.
  2. It is related to $\mathcal{H}$.

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

- finite, upper-bounded by 2^N

lines in 2D

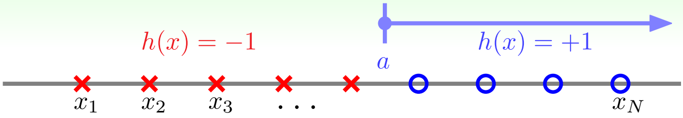
N	$m_{\mathcal{H}}(N)$
1	2
2	4
3	$\max(\ldots, 6, 8)$ $= 8$
4	$14 < 2^N$

"effective number of lines" in 2D perceptron ≡ growth function (We'll discuss 2D perceptron's growth function in the future.)

how to 'calculate' the growth function?

Let's see some easier examples.

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 We didn't discuss the negative rays here.

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

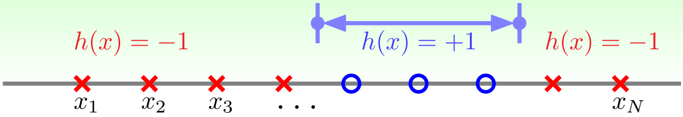
$$m_{\mathcal{H}}(N) = N + 1$$

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

If we only have 4 points:
(≡ 5 intervals)

	x_1	x_2	x_3	x_4
$a < x_1$:	○	○	○	○
$x_1 < a < x_2$:	×	○	○	○
	×	×	○	○
	×	×	×	○
	×	×	×	×

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'

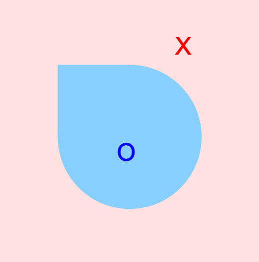
$$m_{\mathcal{H}}(N) = \underbrace{\binom{N+1}{2}}_{\text{interval ends in } N+1 \text{ spots}} + \underbrace{1}_{\text{all } \times}$$

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

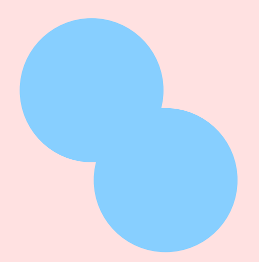
x_1	x_2	x_3	x_4
○	×	×	×
○	○	×	×
○	○	○	×
○	○	○	○
×	○	×	×
×	○	○	×
×	○	○	○
×	×	○	×
×	×	○	○
×	×	×	○
×	×	×	×

$\left(\frac{1}{2}N^2 + \frac{1}{2}N + 1\right) \ll 2^N$ when N 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 x in a convex region,** -1 **otherwise**

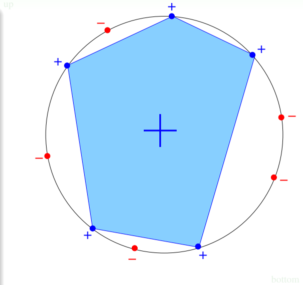
what is $m_{\mathcal{H}}(N)$?

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

$$\boxed{m_{\mathcal{H}}(N) = 2^N}$$ **= upper bound**

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



$$\boxed{\begin{array}{c} m_{\mathcal{H}}(N) = 2^N \iff \\ \textbf{exists } N \text{ inputs that can be shattered} \end{array}}$$

Fun Time

Consider positive **and negative** rays as $\mathcal{H}$, which is equivalent to the perceptron hypothesis set in 1D. The hypothesis set is often called **'decision stump'** to describe the shape of its hypotheses. What is the growth function $m_{\mathcal{H}}(N)$?

1. N
2. $N + 1$
3. $2N$
4. 2^N

Fun Time

Consider positive **and negative** rays as $\mathcal{H}$, which is equivalent to the perceptron hypothesis set in 1D. The hypothesis set is often called **'decision stump'** to describe the shape of its hypotheses. What is the growth function $m_{\mathcal{H}}(N)$?

1. N
2. $N+1$
3. $2N$
4. 2^N

Reference Answer: ③

Two dichotomies when threshold in each of the $N-1$ 'internal' spots; two dichotomies for the all-$\circ$ and all-$\times$ cases.

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?

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

polynomial: good; **exponential: bad**

↗ polynomially + ↘ exponentially = ↘ ⇒ good
↗ exponentially + ↘ exponentially = ? ⇒ bad

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

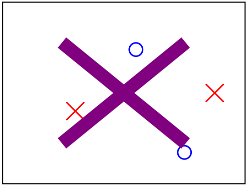
Break Point of $\mathcal{H}$

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: **break point at** 4

The Four Break Points

- positive rays: o x — break point at 2 — $m_{\mathcal{H}}(N) = N + 1 = O(N)$

  = O(N²⁻¹)

- positive intervals: o x o — break point at 3 — $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 = O(N^2)$

  = O(N³⁻¹)

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

  no break point

- 2D perceptrons: o x / x o — break point at 4 — $m_{\mathcal{H}}(N) < 2^N$ **in some cases**

  ? = O(N⁴⁻¹) = O(N³)

conjecture: Maybe the break point is related to growth function? (It's just an assumption!)

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

excited? wait for next lecture :-)

Need some prove

Spoiler alert: it's actually ≤.

Fun Time

Consider positive **and negative** rays as $\mathcal{H}$, which is equivalent to the perceptron hypothesis set in 1D. As discussed in an earlier quiz question, the growth function $m_{\mathcal{H}}(N) = 2N$. What is the minimum break point for $\mathcal{H}$?

1. 1
2. 2
3. 3
4. 4

Fun Time

Consider positive **and negative** rays as $\mathcal{H}$, which is equivalent to the perceptron hypothesis set in 1D. As discussed in an earlier quiz question, the growth function $m_{\mathcal{H}}(N) = 2N$. What is the minimum break point for $\mathcal{H}$?

① 1 ② 2 ③ 3 ④ 4

Reference Answer: ③

At $k = 3$, $m_{\mathcal{H}}(k) = 6$ while $2^k = 8$.

If we use the previous assumption, we will get the result of k = 2.
($\because$ $O(2N) = O(N) = O(N^{2-1})$)

$$m_{\mathcal{H}}(N) \le O(N^{k-1})$$
$$\Rightarrow \quad 2N \le O(N^{3-1}))$$

Summary

1. When Can Machines Learn?

   Lecture 4: Feasibility of Learning

2. **Why** Can Machines Learn?

   Lecture 5: Training versus Testing

   - Recap and Preview

     two questions: $E_{\text{out}}(g) \approx E_{\text{in}}(g)$, and $E_{\text{in}}(g) \approx 0$
   - 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'

   - **next: $m_{\mathcal{H}}(N) = poly(N)$?**

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