Training versus Testing

Starfly
starfly3119@gmail.com

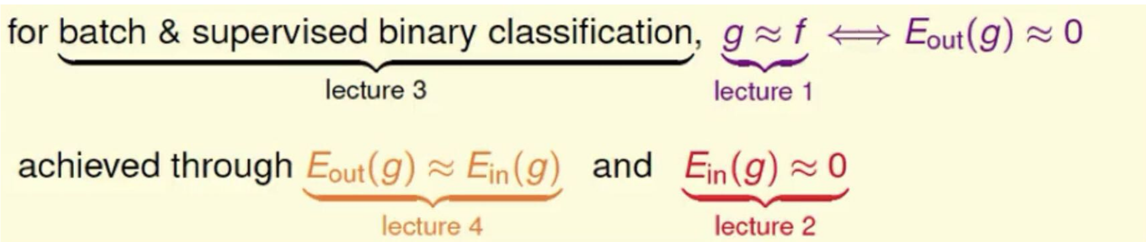
Beihang University — February 16, 2020

Introduction

In this lecture, we want to modify M in the Hoeffding's Non-equation. In the first section, we find that **learning is split to two central questions**; in the second section, we find that **for four inputs, we only have 14 different kinds of lines**; in the third section, we have **at most** $m_H(N)$ **through the eye of** N **inputs**; in the fourth section, we give the definition of **break point** and we want to find that **when** $m_H(N)$ **becomes 'non-exponential'**

1 Recap and Preview

1.1 Recap



Learning split to two central questions:

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

1.2 Trade-off on M

Question 1

What role does $|H|$ play for the two 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?

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

Using the right M (or H) is important, $M = \infty$ doomed?

1.3 Preview

Known

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

Todo

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

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

- 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 :-)

2 Effective Number of Lines

2.1 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

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

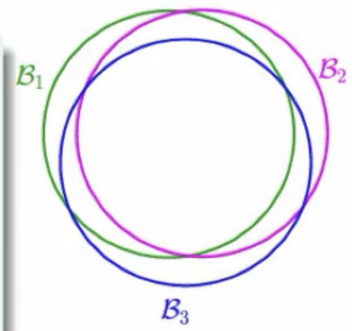
Where did uniform bound fail to consider for $M = \infty$?

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

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

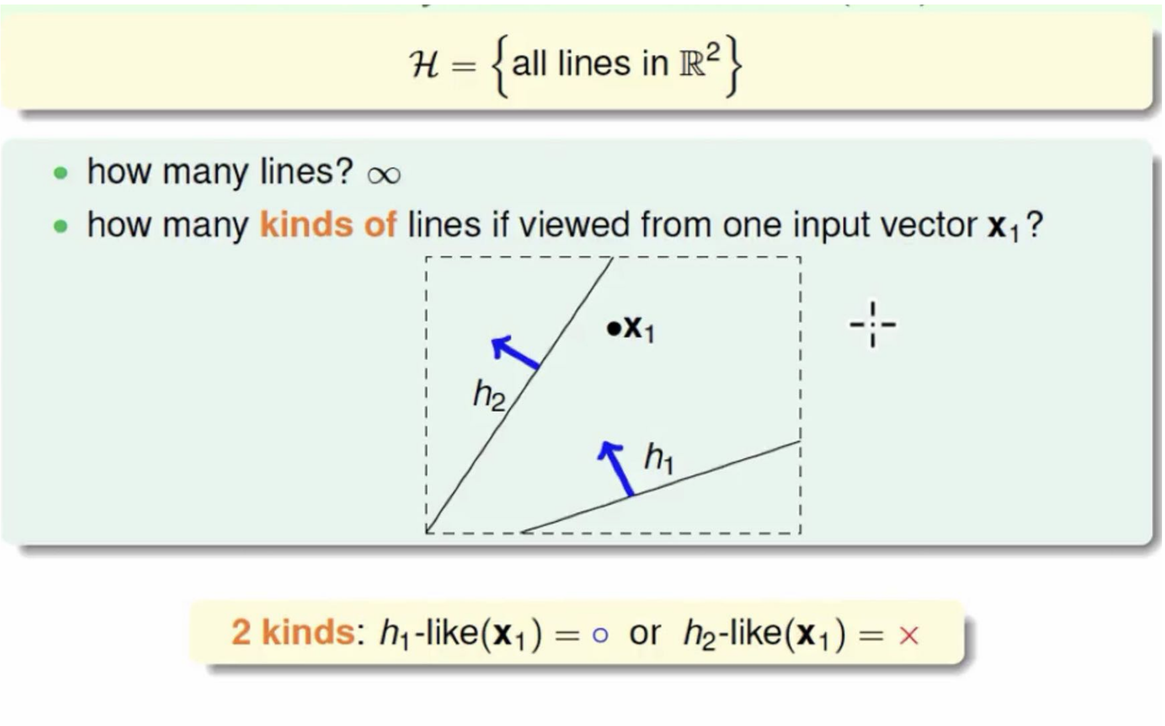
  overlapping for similar hypotheses $h_1 \approx h_2$
- why? 1 $E_{\text{out}}(h_1) \approx E_{\text{out}}(h_2)$

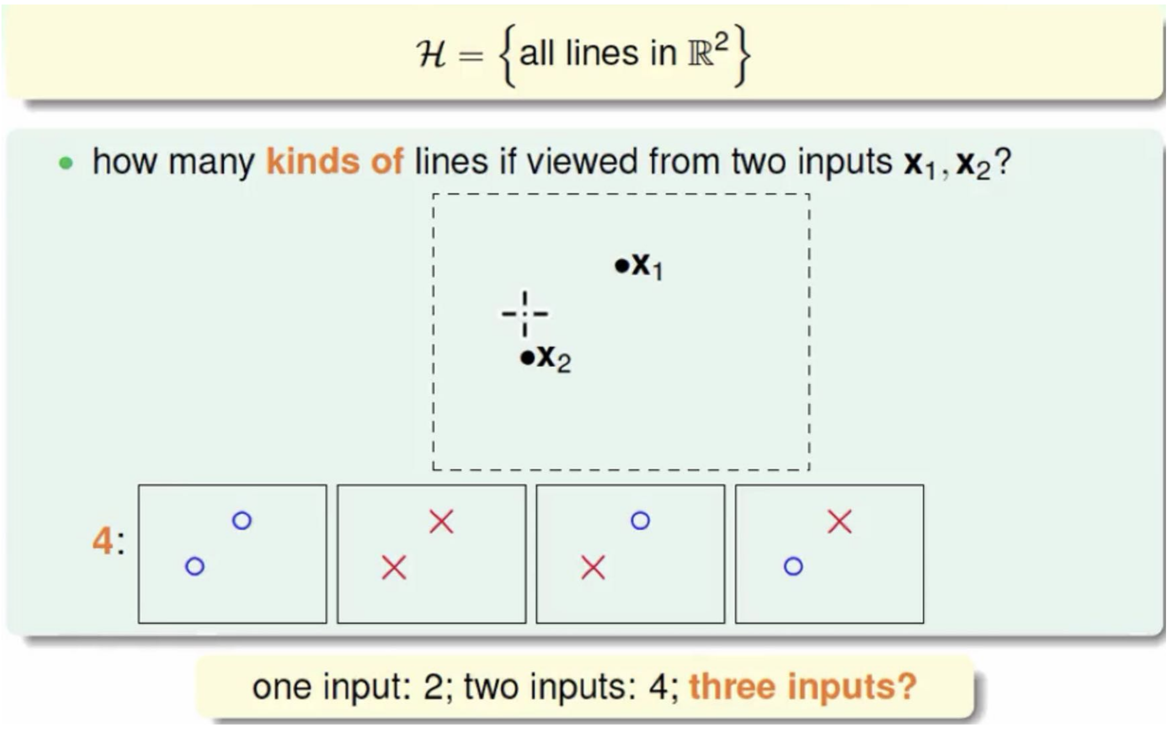
  2 for most $\mathcal{D}$, $E_{\text{in}}(h_1) = E_{\text{in}}(h_2)$
- union bound **over-estimating**



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

2.2 How Many Lines Are There?





For three points, we have fewer than 8 (when degenerate)

2.3 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:

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

lines in 2D

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 :-)

3 Effective Number of Hypotheses

3.1 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**

3.2 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

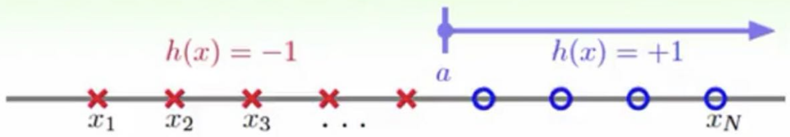
how to 'calculate' the growth function?

lines in 2D

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

3.2.1 Easy examples

Growth Function for Positive Rays

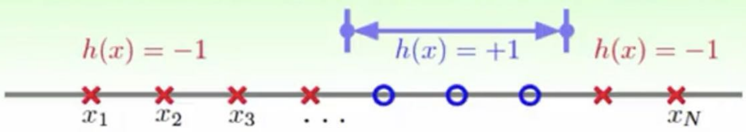


- $\mathcal{X} = \mathbb{R}$ (one dimensional)
- $\mathcal{H}$ contains h, where **each** $h(x) = \text{sign}(x - a)$ **for threshold** a
- 'positive half' of 1D perceptrons

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

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

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'

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

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

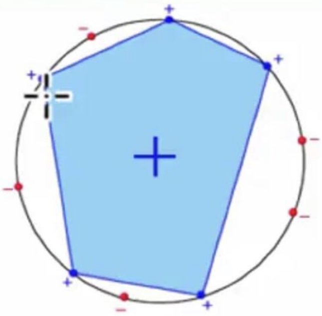
$\left(\frac{1}{2}N^2 + \frac{1}{2}N + 1\right) \ll 2^N$ when N large!

x_1	x_2	x_3	x_4
o	×	×	×
o	o	×	×
o	o	o	×
o	o	o	o
×	o	×	×
×	o	o	×
×	o	o	o
×	×	o	×
×	×	o	o
×	×	×	o
×	×	×	×

- 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

4 Break point

4.1 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**

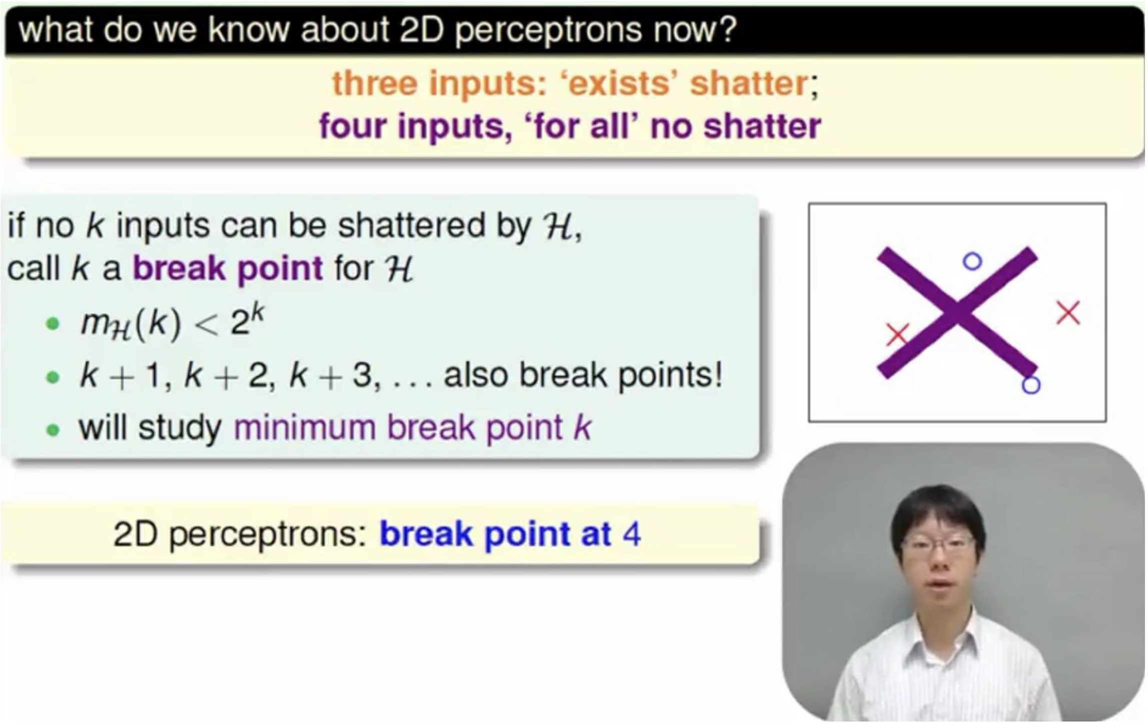
what if $m_{\mathcal{H}}(N)$ replaces M?

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

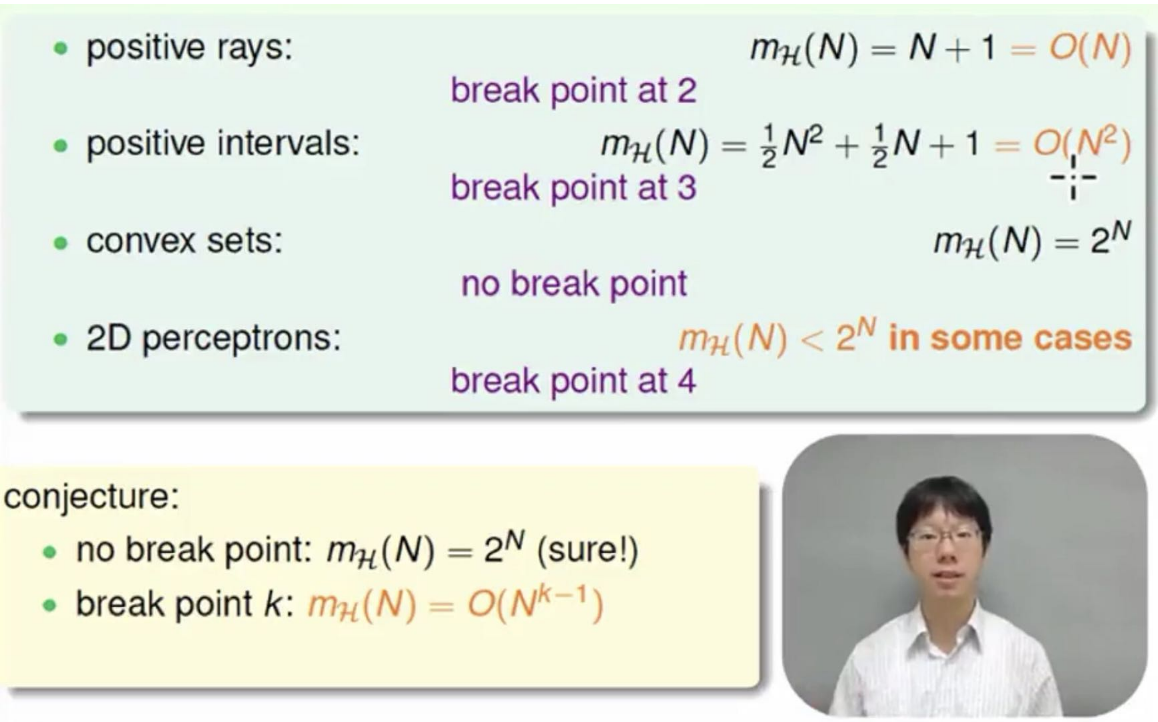
polynomial: good; **exponential: bad**

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

4.2 Break Point of H



4.3 The Four Break Points



Reference