Machine Learning from Data

Lecture 5: Spring 2021

Today's Lecture

- Training Vs Testing
  - The Two Questions of Learning
  - Theory of Generalization (Ein ≈ Eout)
  - An Effective Number of Hypotheses

Feasibility

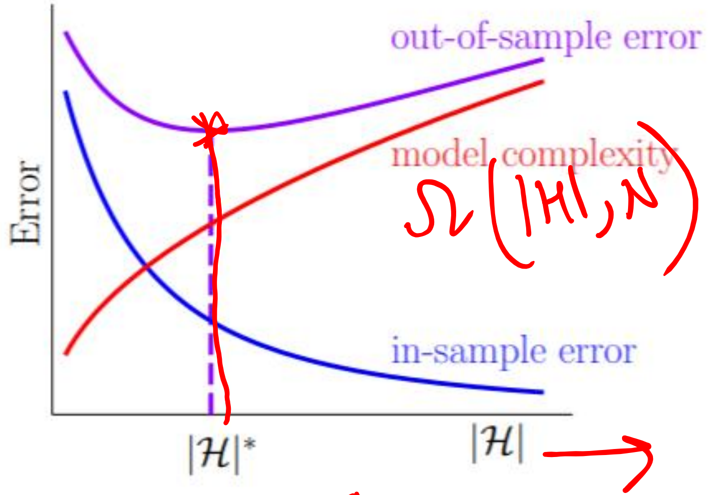
1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$? $E_{in}(g) \simeq E_{out}(g)$
2. Can we make $E_{\text{in}}(g)$ small enough?

The Hoeffding *generalization bound*:

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \underbrace{\sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}}_{\text{generalization error bar}}$$

E_{in}: training (eg. the practice exam)

E_{out}: testing (eg. the real exam)



There is a tradeoff when picking $|\mathcal{H}|$.

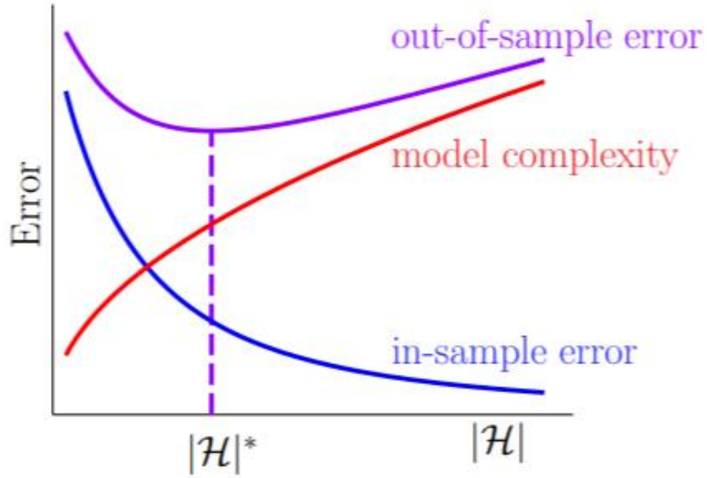
The Two Questions of Learning

error bar

$|\mathcal{H}|$

$\mathcal{H} \to \infty$

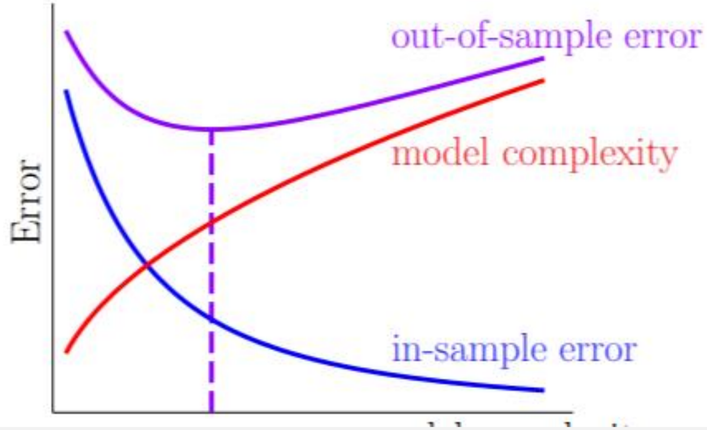
$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}$$

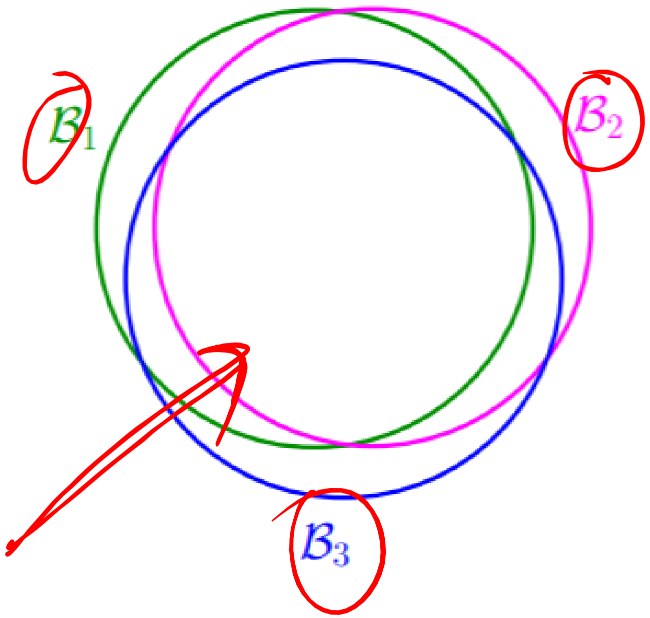


Most models

$\downarrow$

$$\boxed{E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}}{\delta}}}$$





- $\mathcal{B}_m$ are events (sets of outcomes); they can overlap.
- If the $\mathcal{B}_m$ overlap, the union bound is loose.
- If many h_m are similar, the $\mathcal{B}_m$ overlap.
- There are "effectively" fewer than $|\mathcal{H}|$ hypotheses,.
- We can replace $|\mathcal{H}|$ by something smaller.

Meaning of cardinality of H? $|H|$

How did H come in?

$|E_{in}(g) - E_{out}(g)| > \epsilon$, $g \in H$

B_m is a bad event

"$|E_{in}(h_m) - E_{out}(h_m)| > \epsilon$"

$$P[B_g] \le P[\text{any } B_m] \le \underbrace{\sum_{m=1}^{|H|} P[B_m]}$$

Effective Number of Hypothesis

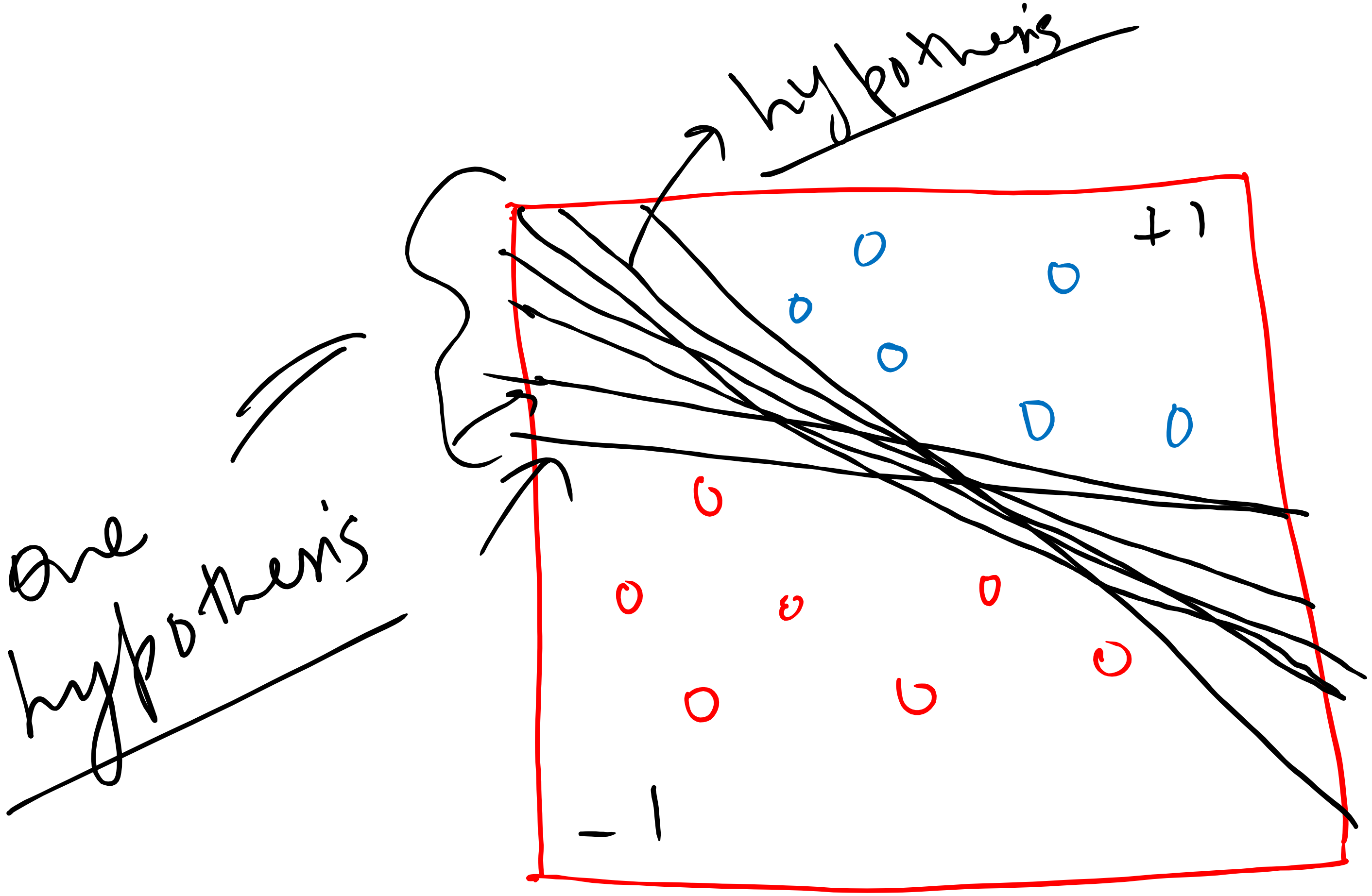
Similarity

- We need a way to measure the diversity of H.

$|H|$

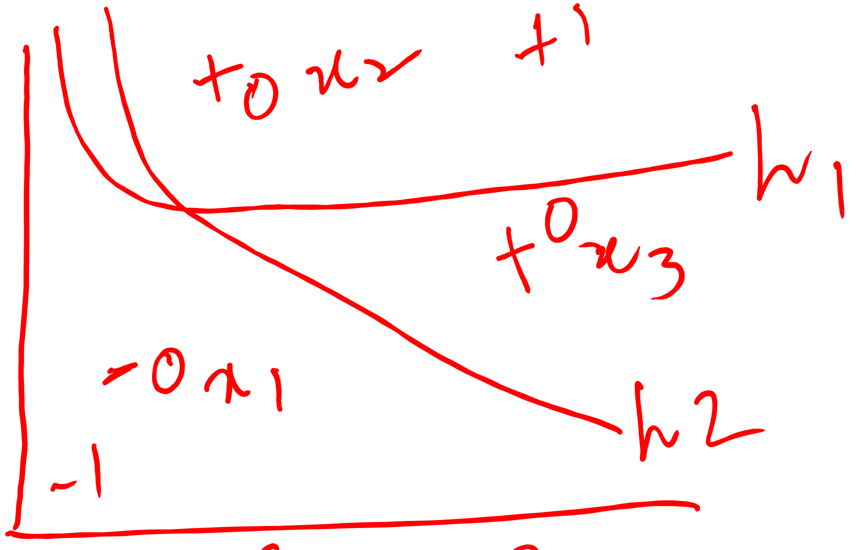
infinite

- A simple idea:
  - Fix any set of N data points.
  - If H is diverse it should be able to implement all functions . . . on these N points.


hypothesis
one hypothesis
+1
-1

The Growth Function

•



$$2^3 = 8$$

dichotomy

	x_1	x_2	x_3
$h1$	$\mp$	$\pm$	$\mp$
h_2	$-$	$+$	$+$
$\vdots$			
h_9	$-$	$+$	$+$

Another Example

(N-points)

$\boxed{2^N}$

$\delta \rightarrow$ dichotomy

$\mathcal{H}$

	x_1	x_2	$\dots$	x_N
h_1	$-$	$-$		$+$
h_2	$+$	$-$		$+$
$\vdots$				

Restriction of the hypothesis set $\mathcal{H}$ to the points $x_1, x_2 \dots x_N$

$$\mathcal{H}(x_1, x_2 \dots x_N) = \{\delta \mid \delta \text{ is implemented by } h \in \mathcal{H}\}$$

Quantify the complexity of H on $x_1, x_2 \cdots x_N$

using $\boxed{\left| H(x_1, \ldots x_N) \right| \leq 2^N}$?

Q1: What $x_1, x_2 \ldots x_N$ should we use to compute the complexity of a hypothesis set?

Q2: Is this bound for complexity useful?

$\{$:

Ans] Yes,

Worst case Analysis

$$m_{\mathcal{H}}(N) = \max_{x_1, x_2 \dots x_N} \left| \mathcal{H}(x_1, x_2 \dots x_N) \right| \leq 2^N$$

growth function!

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

Ans 2

Error bar: $\sqrt{\frac{1}{2N} \ln \frac{2|H|}{\delta}} \rightarrow \underline{\underline{m_H(N)}}$

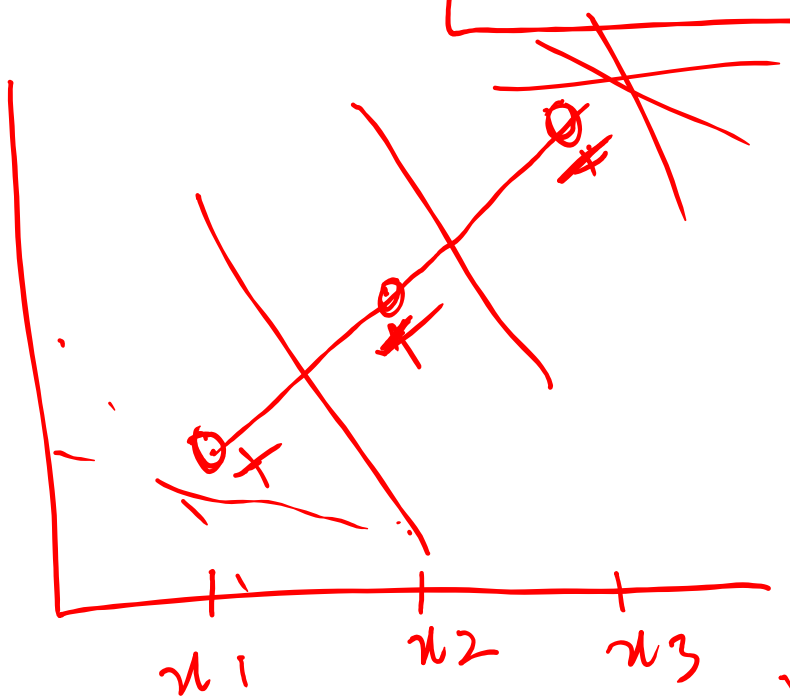
$$\sqrt{\frac{1}{2N} \ln \frac{2 \cdot 2^{N}}{\delta}}$$

$$\sqrt{\frac{1}{2N} d \ln \frac{2N}{\delta}} \qquad \frac{\ln N}{N}$$

$N \rightarrow \infty$ (limit $\frac{\ln N}{N} = 0$ as $N \rightarrow \infty$)

Dichotomies in 2-D Perceptron

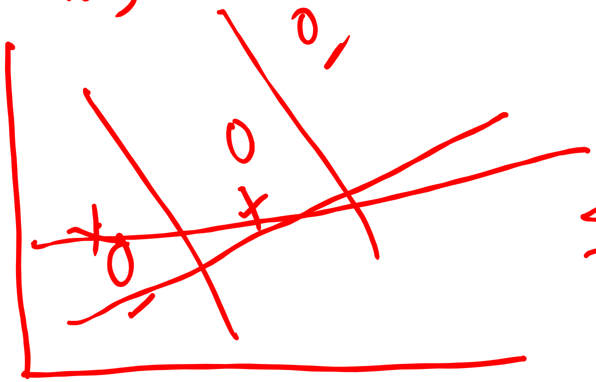
-



x_1	x_2	x_3
+	+	+
−	+	+
−	−	+
−	−	−
+	−	−
+	+	−

3 points $= 2^N$

$|H(x_1, x_2, x_3)| = 6$



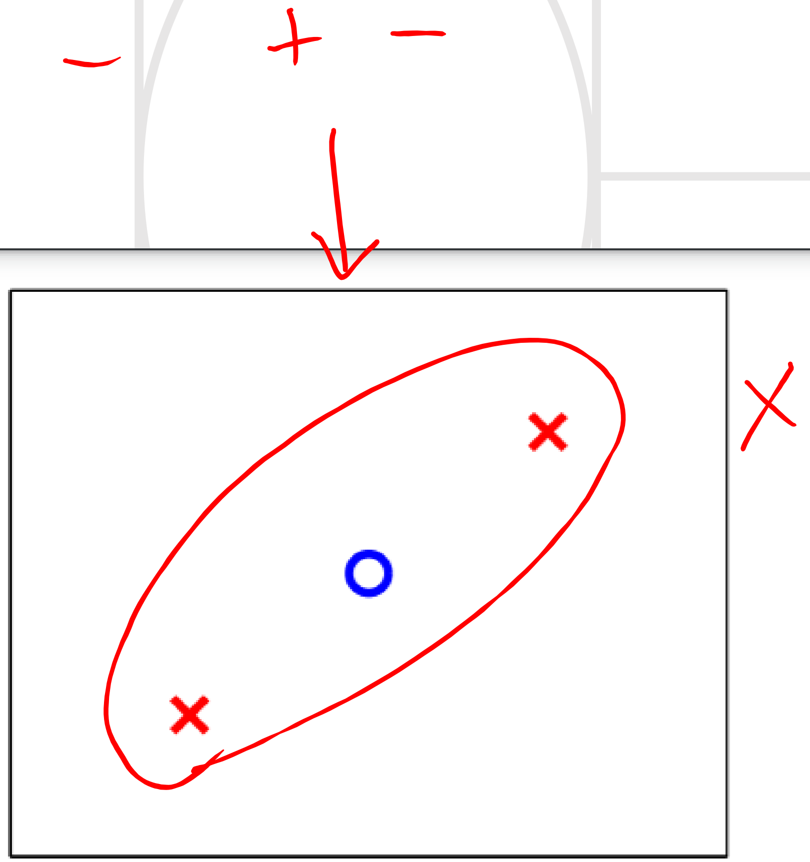
$\Rightarrow$

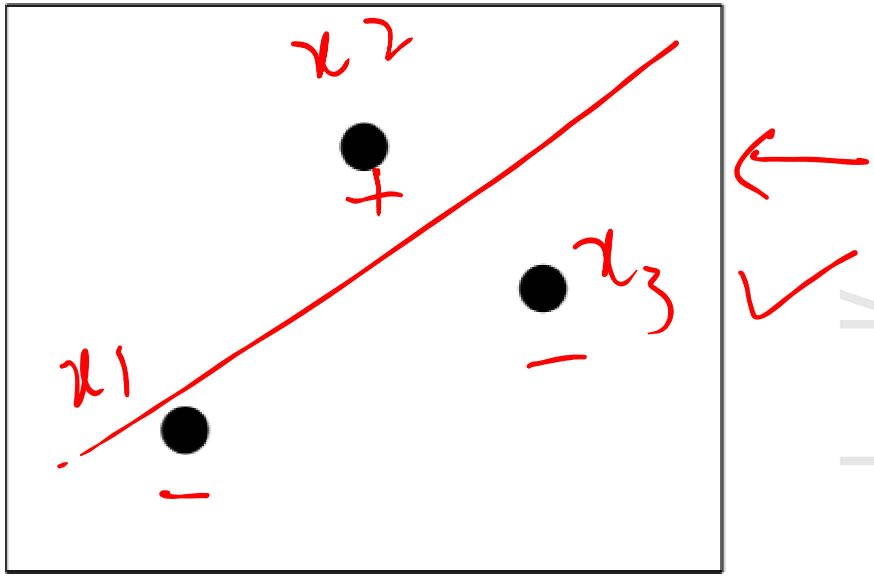
−	+	−
+	−	+

Example - Growth Functions

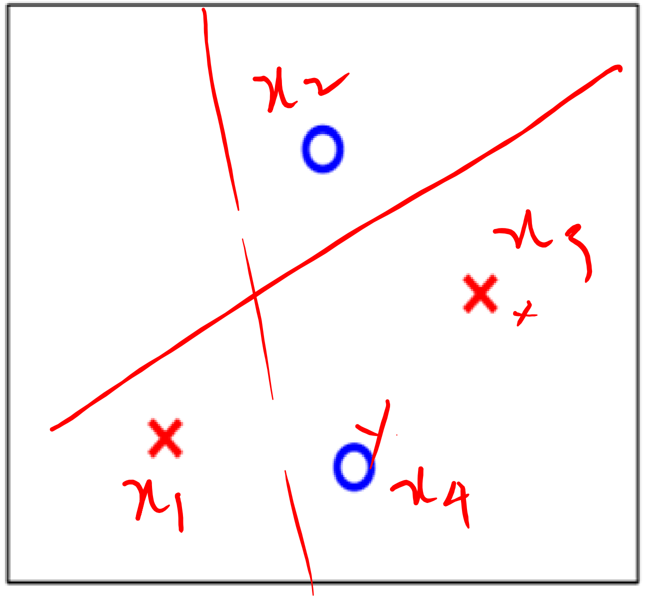
2-D Perceptron Model

For 3 points $= 2^3 = 8$





$m_H(N) = 8$
$= 2^N$



$\mathcal{H} \rightarrow$ Perceptron

$2^4 = 16$

$$\begin{matrix} - & + & - & + \\ + & - & + & - \end{matrix}$$

$m_{\mathcal{H}}(4) = 14 < 2^N$

$14 < 16$

$3 \longrightarrow \underbrace{4 \text{ points}} \rightarrow$ deficiency

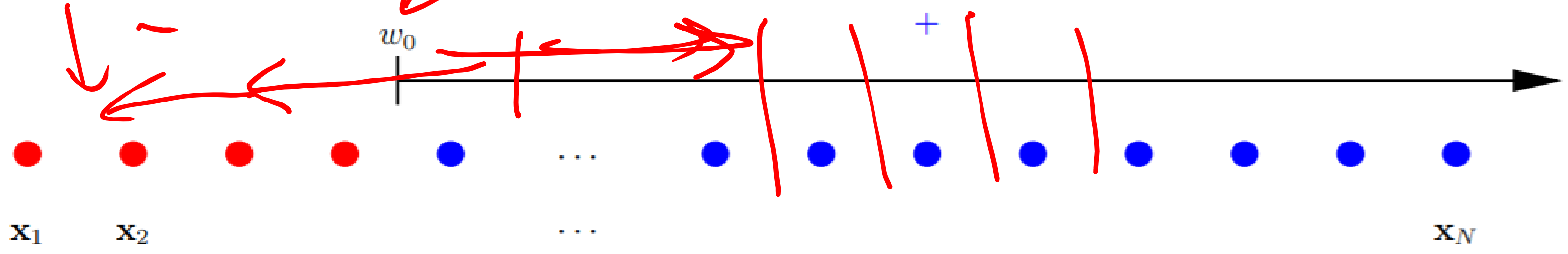
2^N

Value = 4 is important

1-D Positive Ray $\mathcal{H}$

$h(x) = \text{sign}(x - w_0)$

Threshold



N points

N points

N+1 regions

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

N	1	2	3	4
$m_H(N)$	2	3	4	5
2^N	2	4	8	16

Broken at $N=2$

Worst Case 2^N

$3 < 4$

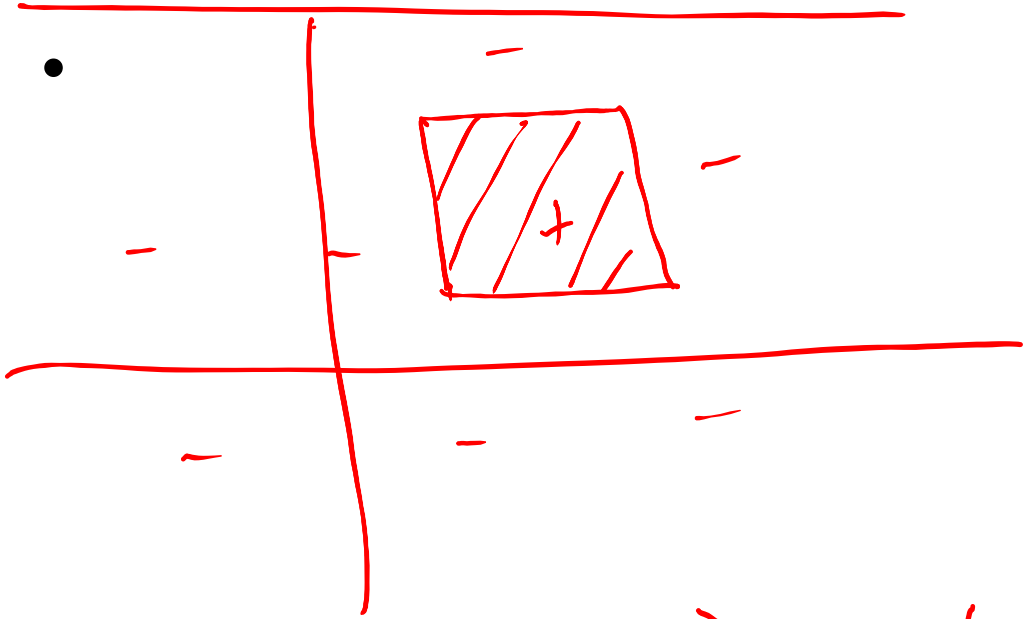
Perceptron

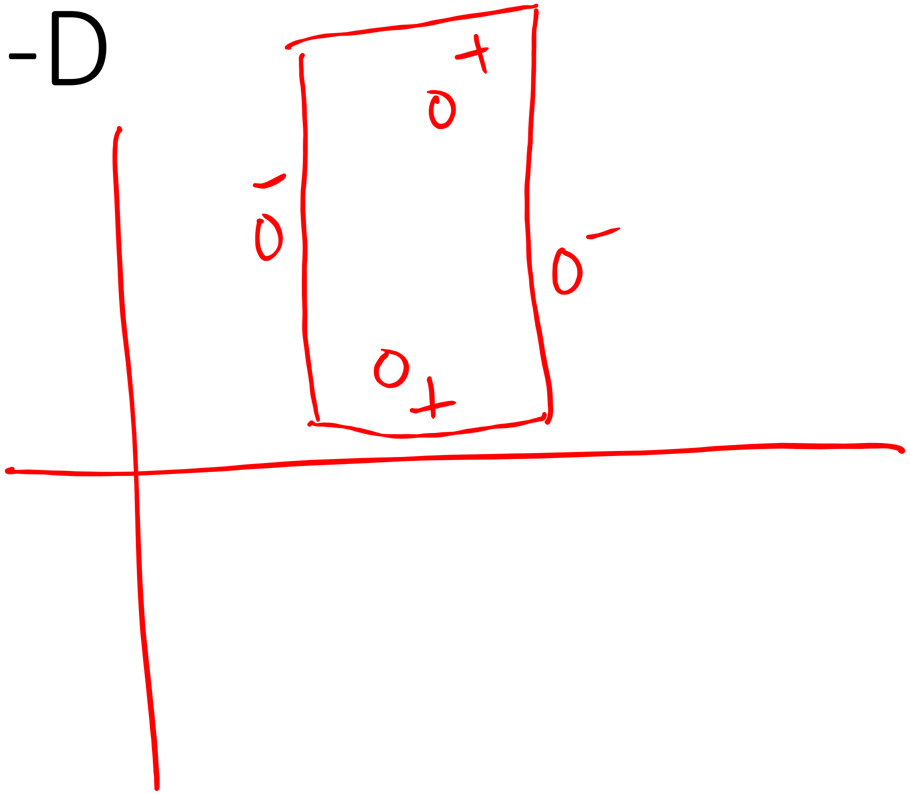
N	1	2	3	4	5
$m_H(N)$	2	4	8	14 (<16)	
2^N	2	4	8	16	

*

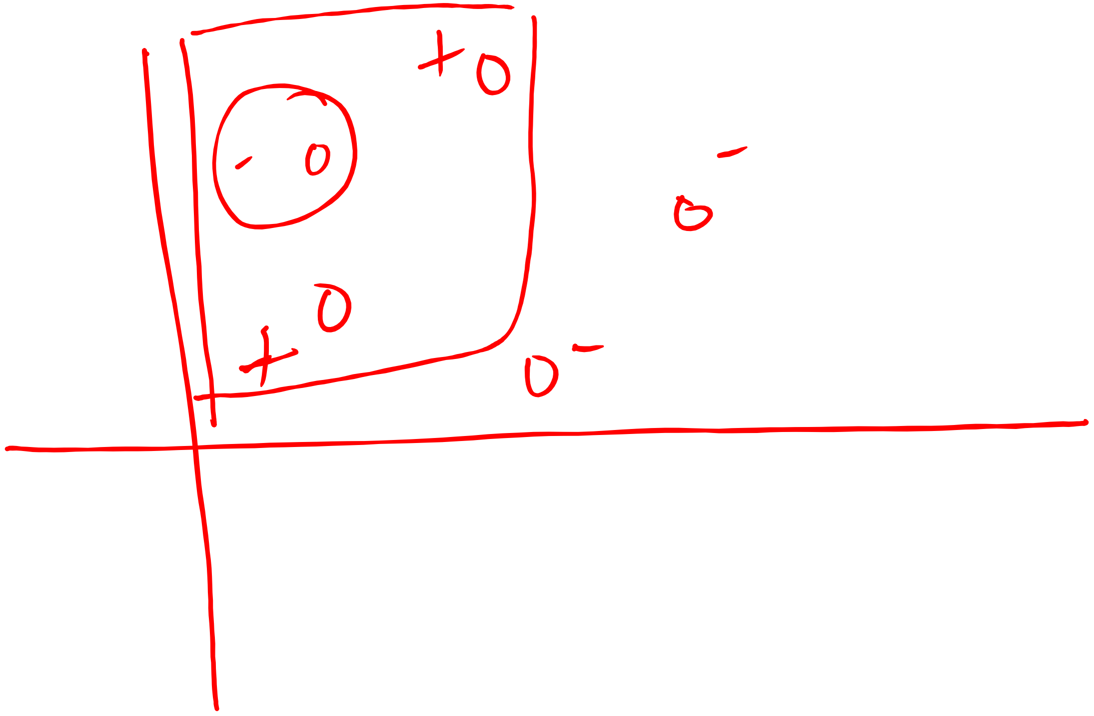
Positive Rectangles in 2-D

-





$$m_H(N) = m_H(4) = 16 \quad (2^N)$$



$$\underbrace{m_{H}(5)} < 2^{5}$$

Summarize

Effective no. → $\mathcal{H}, N$

2^N

$N = 4$

	1	2	3	4	5	...
2-D perceptron	2	4	8	14	...	
1-D pos. ray	2	3	4	5	...	
2-D pos. rectangles	2	4	8	16	$< 2^5$	...

$\mathcal{H}$

- $m_{\mathcal{H}}(N)$ drops below 2^N – there is hope for the generalization bound.
- A break point is any n for which $m_{\mathcal{H}}(n) < 2^n$.

Definition: Shatter a Data Set

$\mathcal{X}$

- $\underbrace{x_1 \quad x_2 \cdots\cdots \quad x_N}$

2^N

$H \xrightarrow[\text{implement}]{\text{can}}$ All possible dichotomies (2^N)

Finding 4 dichotomies that cannot shatter

$2^N = 8$

Combinatorial Puzzle

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
○	●	●

any pair of 2 points

○ −

● +

2 points are shattered

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○ −	○ −
○	○ −	● +
○	● +	○ −
○	● +	● +

No Pair of Points is Shattered

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
●	○	○

If N = 4 how many possible dichotomys with no 2 points shattered?

$2^4 = 16$

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
●	○	○

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
○	○	○	○
○	○	○	●
	⋮		

Thanks!