Machine learning II

Learning theory

Souhaib Ben Taieb

February 23, 2022

University of Mons

Table of contents

Table of contents

Training vs Testing

Testing:

$$P[|E_{\text{in}} - E_{\text{out}}| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$

Training:

$$P[|E_{\text{in}} - E_{\text{out}}| > \varepsilon] \leq 2Me^{-2\varepsilon^2 N}$$

We would like to replace M by a quantity that is not useless with an infinite hypothesis set.

Review

The statement we would like to make is

$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon]$ is small for the final hypothesis g.

We know that

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

$$\implies$$

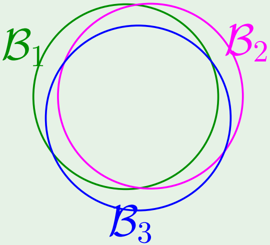
$$\begin{aligned} & |E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\ \text{or } & |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\ \ldots & \\ \text{or } & |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon \end{aligned}$$

Overlap between bad events

Where did the M come from?

The $\mathcal{B}$ad events $\mathcal{B}_m$ are

$$\text{“}|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\text{”}$$

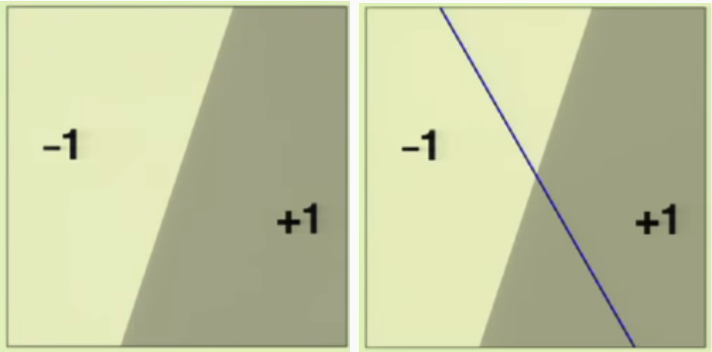


The union bound:

$$\mathbb{P}[\mathcal{B}_1 \textbf{ or } \mathcal{B}_2 \textbf{ or } \cdots \textbf{ or } \mathcal{B}_M]$$

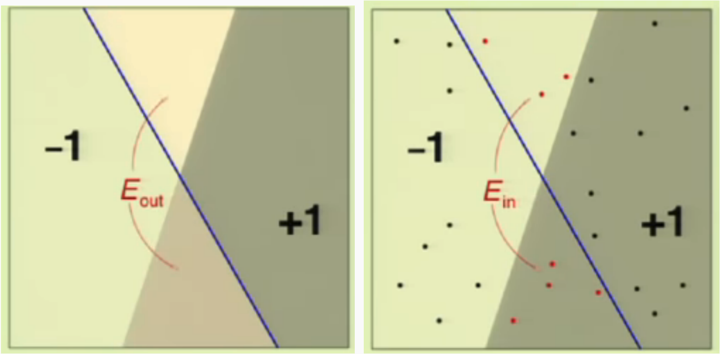
$$\leq \underbrace{\mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \cdots + \mathbb{P}[\mathcal{B}_M]}_{\text{no overlaps: } M \text{ terms}}$$

Can we improve on M?



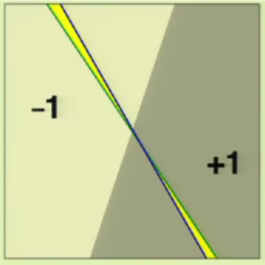
What is E_{out}? What is E_{in}?

What is E_{out}? E_{in}?



Consider another (very similar) hypothesis. How will E_{out} and E_{in} change?

ΔE_{out} and ΔE_{in}

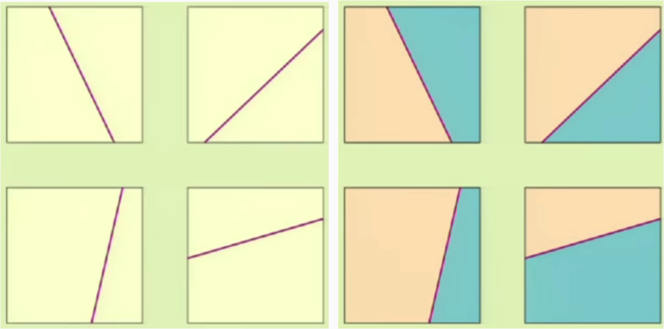


- ΔE_{out} is the change in E_{out} (yelllow area)
- ΔE_{in} is the change in labels of data points
- $|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \varepsilon$ happens as often as $|E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \varepsilon$

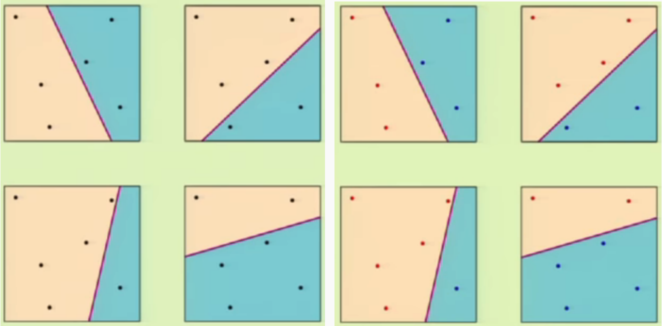
Can we improve on M? Yes, bad events are very overlapping!

Table of contents

What can we replace M with?

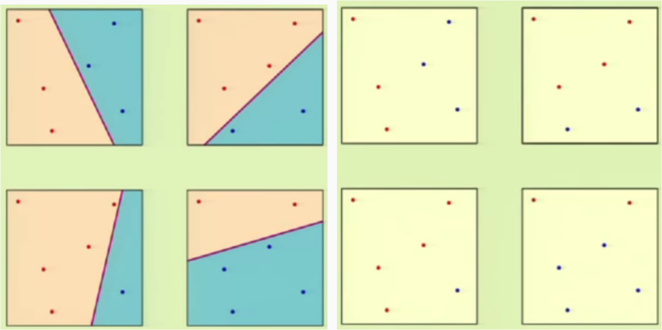


What can we replace M with?



What can we replace M with?

Instead of the whole input space, we consider a (finite) set of input points, and count the number of **dichotomies**.



The number of dichotomies is a candidate for replacing M.

Table of contents

Dichotomies: mini-hypotheses

- A hyothesis $h : \mathcal{X} \rightarrow \{-1, +1\}$
- A **dichotomy** $h : \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \rightarrow \{-1, +1\}$

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{X}$. The dichotomies generated by $\mathcal{H}$ on these points are defined by

$$\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_N)) | h \in \mathcal{H}\}.$$

For any $\mathcal{H}$, $\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \subseteq \{-1, +1\}^N$ (the set of all possible dichotomies on any N points).

The number of dichotomies is at most 2^N:

$$|\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)| \leq |\{-1, +1\}^N| \leq 2^N.$$

The number of dichotomies is a candidate for replacing M.

The growth function

The growth function counts the most dichotomies on any N points:

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

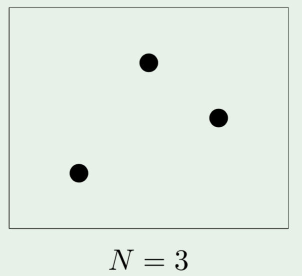
The growth function satisfies:

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

If $\mathcal{H}$ is capable of generating all possible dichotomies on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, then $\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \{-1, 1\}^N$ and we say that $\mathcal{H}$ can **shatter** $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

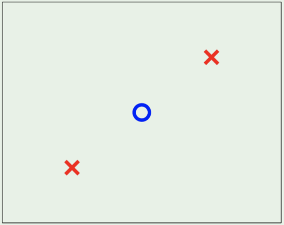
$m_{\mathcal{H}}(N)$ is also known as the N-th **shattering coefficient** of $\mathcal{H}$.

The growth function for the perceptron ($N = 3$)



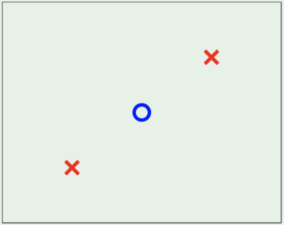
$m_{\mathcal{H}}(3) = ?$

The growth function for the perceptron ($N = 3$)



$N = 3$

The growth function for the perceptron ($N = 3$)



$N = 3$

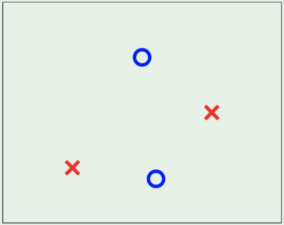
The growth function counts the most dichotomies on **any** N points:

$$m_{\mathcal{H}}(3) = 8$$

The growth function for the perceptron ($N = 4$)

$$m_{\mathcal{H}}(4) = ?$$

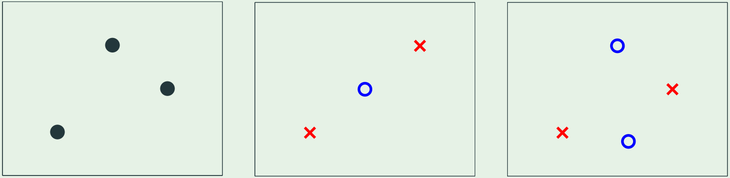
The growth function for the perceptron ($N = 4$)



$N = 4$

The growth function for the perceptron

Applying $m_{\mathcal{H}}(N)$ definition - perceptrons

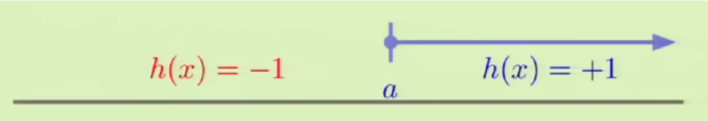


$N = 3$ $N = 3$ $N = 4$

$$m_{\mathcal{H}}(3) = 8 \qquad m_{\mathcal{H}}(4) = 14$$

Table of contents

Example 1: positive rays

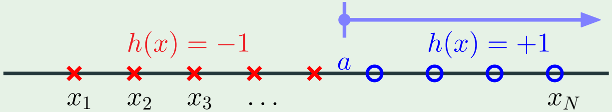


$$h(x) = \text{sign}(x - a)$$

$$m_{\mathcal{H}}(N) = ?$$

Example 1: positive rays

Example 1: positive rays

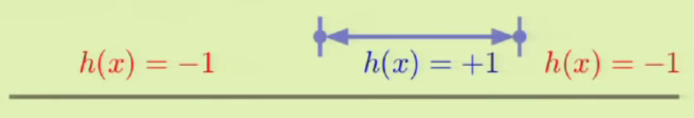


$\mathcal{H}$ is set of $h\colon \mathbb{R} \to \{-1, +1\}$

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

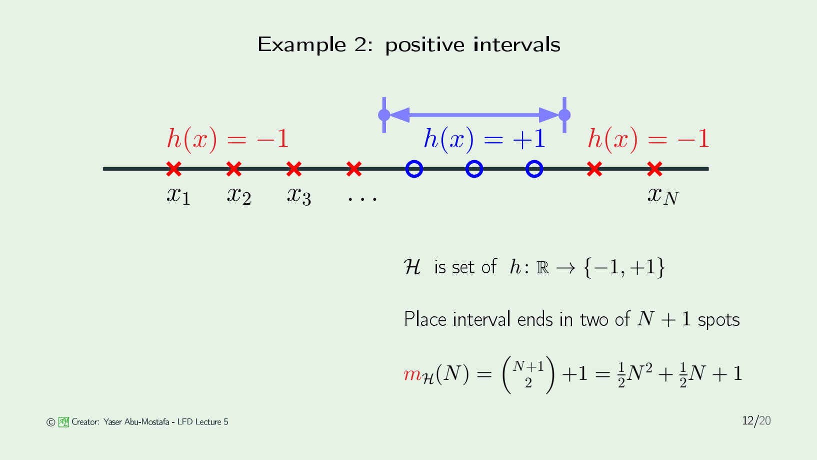
$m_{\mathcal{H}}(N) = N + 1$

Example 2: positive intervals



$$m_{\mathcal{H}}(N) = ?$$

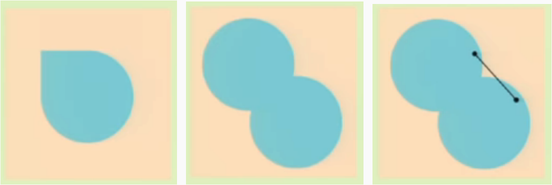
Example 2: positive intervals



Note that $m_{\mathcal{H}}(N)$ grows as the square of N, faster then the linear $m_{\mathcal{H}}(N)$ of the "simpler" positive ray case.

Example 3: convex sets

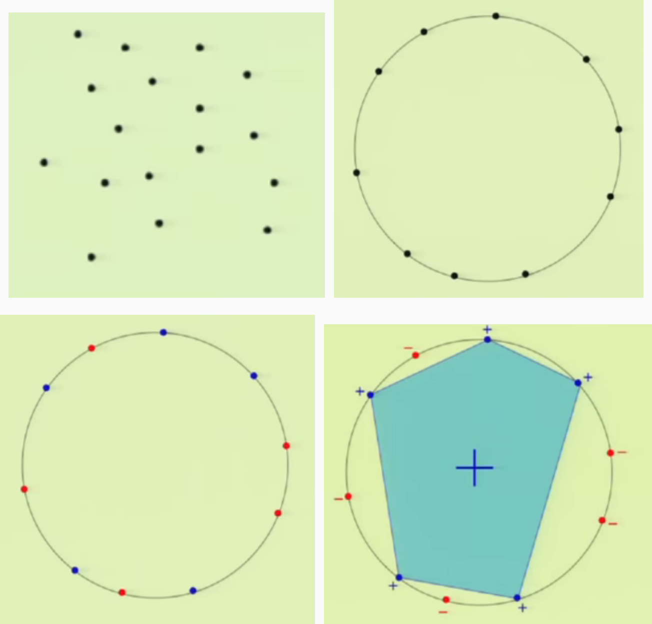
- $\mathcal{H}$ consists of all hypotheses in two dimensions $h : \mathbb{R}^2 \to \{-1, +1\}$ that are positive inside some convex set and negative elsewhere.
- A set is convex if the line segment connecting any two points in the set lies entirely within the set.



Example 3: convex sets

- Because we chose the N points at random in the plane, many of the points are “internal”, and we are not able to shatter all the points with convex hypotheses.
- Is there another choice for the points that provide more hypotheses?

Example 3: convex sets



Example 3: convex sets

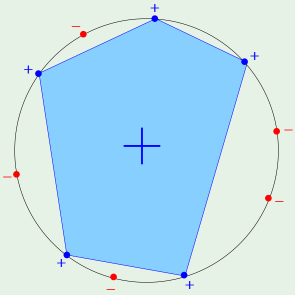
Example 3: convex sets

$\mathcal{H}$ is set of $h\colon \mathbb{R}^2 \to \{-1,+1\}$

$h(\mathbf{x}) = +1$ is convex

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

The N points are 'shattered' by convex sets



The 3 growth functions

The 3 growth functions

- $\mathcal{H}$ is positive rays:

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

- $\mathcal{H}$ is positive intervals:

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

- $\mathcal{H}$ is convex sets:

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

$\rightarrow$ More complex $\mathcal{H}$ gives a bigger growth function

Back to the big picture

$$P[|E_{\text{in}} - E_{\text{out}}| > \varepsilon] \leq 2Me^{-2\varepsilon^2 N}$$

In future lectures, we will see that $m_{\mathcal{H}}(N)$ can replace M. Our bound is now finite even for an infinite hypothesis set!

As a function of N, what is a good property of $m_{\mathcal{H}}(N)$?

Back to the big picture

$$P[|E_{\text{in}} - E_{\text{out}}| > \varepsilon] \leq 2Me^{-2\varepsilon^2 N}$$

In future lectures, we will see that $m_{\mathcal{H}}(N)$ can replace M. Our bound is now finite even for an infinite hypothesis set!

As a function of N, what is a good property of $m_{\mathcal{H}}(N)$?

If $m_{\mathcal{H}}(N)$ is polynomial (in N), this is good for learning!

In fact, for any real constants a and b such that $a > 1$,

$$\lim_{N\to\infty} \frac{N^b}{a^N} = 0.$$

We will prove that $m_{\mathcal{H}}(N)$ is polynomial. The key notion to prove it is the **break point**.

Table of contents

Main result

No break point $\Longrightarrow$ $m_{\mathcal{H}}(N) = 2^N$

Any break point $\Longrightarrow$ $m_{\mathcal{H}}(N)$ is **polynomial** in N

The break point

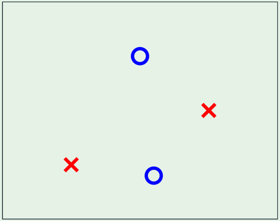
Break point of $\mathcal{H}$

Definition:

If no data set of size k can be shattered by $\mathcal{H}$, then k is a *break point* for $\mathcal{H}$

$$m_{\mathcal{H}}(k) < 2^k$$

For 2D perceptrons, $k = 4$



A bigger data set cannot be shattered either

In general, it is easier to find a break point for $\mathcal{H}$ than to compute the full growth function for that $\mathcal{H}$.

Break point - the three examples

The 3 growth functions

- $\mathcal{H}$ is positive rays:

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

- $\mathcal{H}$ is positive intervals:

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

- $\mathcal{H}$ is convex sets:

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

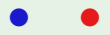
Is there a break point? What is it?

Break point - the three examples

Break point - the 3 examples

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

break point $k =$ 2



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

break point $k =$ 3

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

break point $k =$'∞'

Main result

No break point $\implies$ $m_{\mathcal{H}}(N) = 2^N$

Any break point $\implies$ $m_{\mathcal{H}}(N)$ is **polynomial** in N

A puzzle (N = 3)

Let us assume $k = 2$ is a break point. What are the number of dichotomies on $N = 3$ points? (Use $\circ$ for -1 and $\bullet$ for $+1$).

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

A puzzle (N = 3)

Let us assume $k = 2$ is a break point. What are the number of dichotomies on $N = 3$ points? (Use $\circ$ for -1 and $\bullet$ for $+1$).

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

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\circ$	$\circ$	$\circ$
$\circ$	$\circ$	$\bullet$

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\circ$	$\circ$	$\circ$
$\circ$	$\circ$	$\bullet$
$\circ$	$\bullet$	$\circ$
$\circ$	$\bullet$	$\bullet$

A puzzle (N = 3)

Let us assume $k = 2$ is a break point. What are the number of dichotomies on $N = 3$ points? (Use $\circ$ for -1 and $\bullet$ for $+1$).

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

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

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

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

Two pairs of points are shattered $\rightarrow$ this contradicts the fact that $k = 2$ is a break point.

A puzzle (N = 3)

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

No pair of points is shattered. Maximum 4 dichotomies.

If $k = 2$ is a break point, the maximum number of dichotomies on $N = 3$ points is 4.

A puzzle (N = 4)

What about $N = 4$?

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

A puzzle (N = 4)

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

Try to add a 6th dichotomy.

A puzzle (N = 4)

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

If $k = 2$ is a break point, the maximum number of dichotomies on $N = 4$ points is 5.