


MAC 0460 / 5832

Introduction to Machine Learning

08 – Is learning feasible ? (Infinite \mathcal{H}) 

- dichotomy
- growth function
- break point
-

IME/USP (10/05/2021)

h : hypothesis

$E_{in}(h)$: empirical error \leftarrow computed on D

$E_{out}(h)$: true error \leftarrow unknown

Our question: Does $E_{in}(h)$ say anything about $E_{out}(h)$?

What is the probability that h is a “bad event” ?

$$P\left(\underbrace{|E_{in}(h) - E_{out}(h)|}_{\text{error difference}} > \epsilon\right) = ?$$

Probability of a “bad” event (fixed h)

(Hoeffding)

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq \underline{\underline{2e^{-2\epsilon^2 N}}}$$

Probability of a “bad” event (g selected from a set of M hypothesis)

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq \underline{\underline{2Me^{-2\epsilon^2 N}}}$$

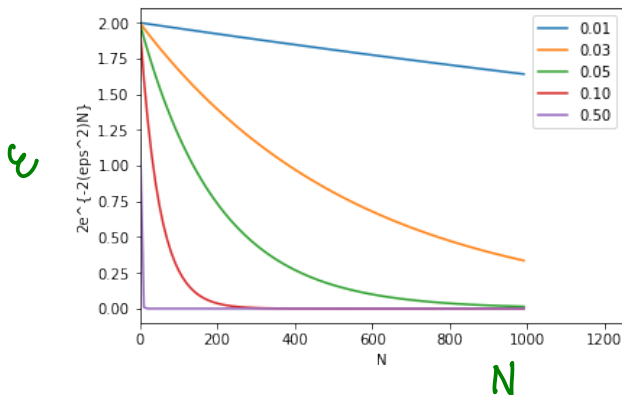
Compare the experiment of **tossing one coin N times** with the experiment of **tossing M coins, N times each**. The chance of a coin resulting in N heads is much larger for the second case.

Recall: Bound variation in function of N

The smaller is ϵ , the larger is the number of samples needed to keep the probability of “bad” events small

(Each color represents a different value of ϵ)

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq \underline{\underline{2e^{-2\epsilon^2 N}}}$$



(verification scenario)

The good news: we can make the Hoeffding bound arbitrarily small

4

(learning scenario)

If M is infinite, the bound $2Me^{-2\epsilon^2 N}$ will be large (meaningless)

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}$$

∞
↑

Today's question: Can we replace M ?

Where did the M come from?

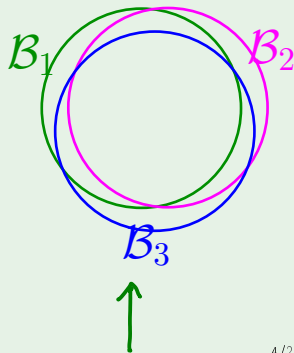
The \mathcal{B}_{ad} events \mathcal{B}_m are

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

The union bound:

$$\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \dots \text{ or } \mathcal{B}_M]$$

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



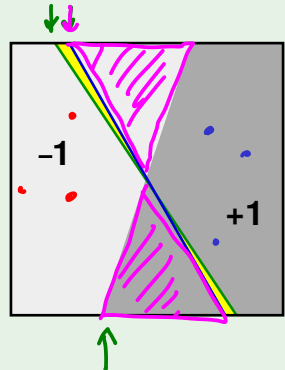
Can we improve on M ?

Yes, bad events are *very* overlapping!

ΔE_{out} : change in +1 and -1 areas

ΔE_{in} : change in labels of data points

$$\underbrace{|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)|}_{\Delta E_{\text{in}}} \approx \underbrace{|E_{\text{in}}(h_2) - E_{\text{out}}(h_2)|}_{\Delta E_{\text{out}}}$$



The choice of g from \mathcal{H} is affected by D (training data)

Usually there are many similar hypothesis h_j that classify samples in D in the exact same way

If in such a group of hypothesis, there is one that corresponds to a “bad” event, would it not be reasonable to think that other similar hypothesis also correspond to a “bad” event ?

To improve the bound, we will replace the Union bound with one that takes the overlap into consideration

For that, we will define a “number” that characterizes the complexity of \mathcal{H}

Important definitions:

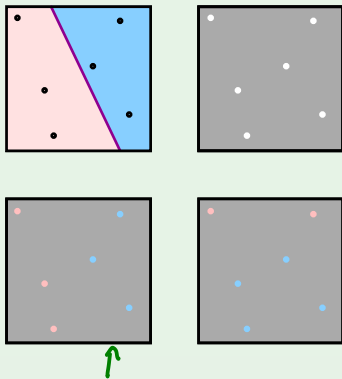
- dichotomy ✓
- growth function ✓
- break point (the “number”) ✓

What can we replace M with?

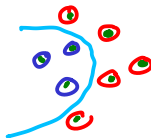
Instead of the whole input space,

we consider a finite set of input points,

and count the number of *dichotomies*



Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ (N points)



Let \mathcal{H} be a hypothesis space

Dichotomies generated by \mathcal{H} :

any bipartition of X as $X_{-1} \cup X_{+1}$ that agrees with some hypothesis $h \in \mathcal{H}$

$$\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = \left\{ \underbrace{(h(\mathbf{x}_1), h(\mathbf{x}_2), \dots, h(\mathbf{x}_N))}_{\text{bipartition}} \mid h \in \mathcal{H} \right\}$$

Dichotomies: mini-hypotheses

A hypothesis $h : \mathcal{X} \rightarrow \{-1, +1\}$

A dichotomy $h : \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \rightarrow \{-1, +1\}$

Number of hypotheses $|\mathcal{H}|$ can be infinite ✓

Number of dichotomies $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)|$ is at most 2^N

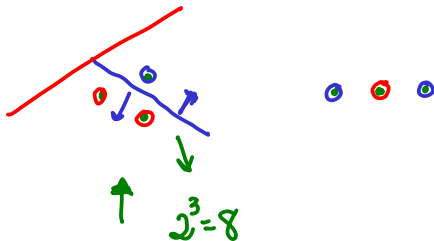
Candidate for replacing M

Why the number of dichotomies $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)|$ is at most 2^N ?

$$\mathcal{X} = \{x_1, x_2, \dots, x_N\}$$

If you consider another set of points, say, $\mathcal{X}' = \{x'_1, x'_2, \dots, x'_N\}$,

1. is $\mathcal{H}(x_1, x_2, \dots, x_N) = \mathcal{H}(x'_1, x'_2, \dots, x'_N)$?
2. is $|\mathcal{H}(x_1, x_2, \dots, x_N)| = |\mathcal{H}(x'_1, x'_2, \dots, x'_N)|$?



The growth function

The growth function counts the most dichotomies on any N points

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

The growth function satisfies:

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

Let's apply the definition.

Growth function for the perceptron

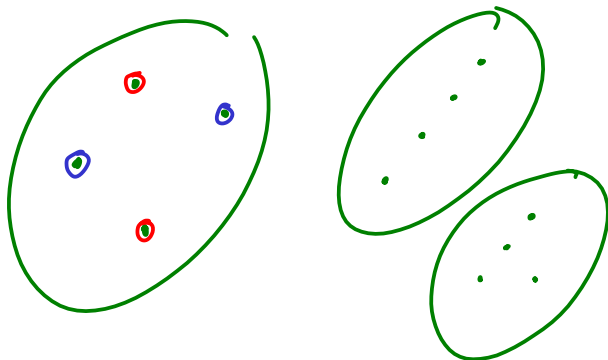
$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \dots, \mathbf{x}_N)|$$

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

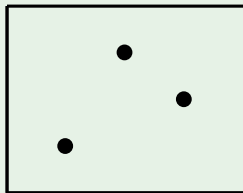
$$m_{\mathcal{H}}(3) = 8 \quad ?$$

\therefore

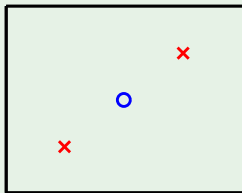
$$m_{\mathcal{H}}(4) = ?$$



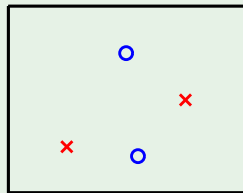
Applying $m_{\mathcal{H}}(N)$ definition - perceptrons



$N = 3$



$N = 3$



$N = 4$

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

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

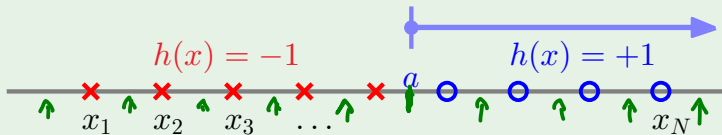


It may not be easy to compute the growth function for an arbitrary hypothesis set.

Imagine doing that for perceptrons, for each value of N !!

There are, however some simple hypothesis set for which we can write down the growth function in terms of N

Example 1: positive rays

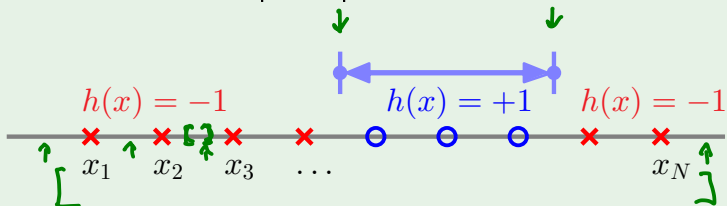


\mathcal{H} is set of $h: \mathbb{R} \rightarrow \{-1, +1\}$

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

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

Example 2: positive intervals



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

Place interval ends in two of $N + 1$ spots

$$m_{\mathcal{H}}(N) = \underbrace{\binom{N+1}{2}}_{\text{green bracket}} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

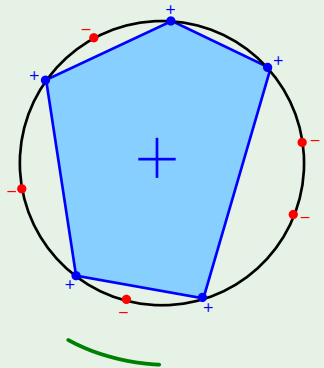
Example 3: convex sets

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

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

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

The N points are 'shattered' by convex sets



The 3 growth functions

- \mathcal{H} is positive rays:

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

- \mathcal{H} is positive intervals:

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

- \mathcal{H} is convex sets:

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

Why are we discussing growth functions ?

Back to the big picture

Remember this inequality?

$$\mathbb{P} [|E_{\text{in}} - E_{\text{out}}| > \epsilon] \leq \underbrace{2M}_{\substack{\downarrow \\ \downarrow}} e^{-2\epsilon^2 N}$$

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

$m_{\mathcal{H}}(N)$ polynomial \implies Good!

Just prove that $m_{\mathcal{H}}(N)$ is polynomial?

If the growth function is polynomial, the bound could be made arbitrarily small by choosing an adequate value of N .

Do we need to compute the growth function value for each N ?

Break point of \mathcal{H}

$$\mathcal{D} = \{x_1, \dots, x_N\}$$

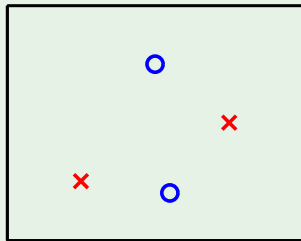
Definition:

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

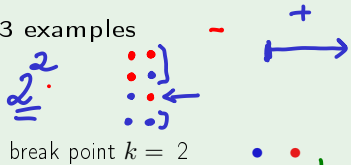


$$m_{\mathcal{H}}(3) = 2^3 = 8$$

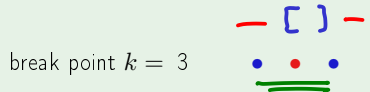
$$m_{\mathcal{H}}(4) = 14 < 2^4$$

Break point - the 3 examples

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

break point $k = \infty$

An exercise

Assume that for a certain hypothesis set \mathcal{H} the break-point is 2

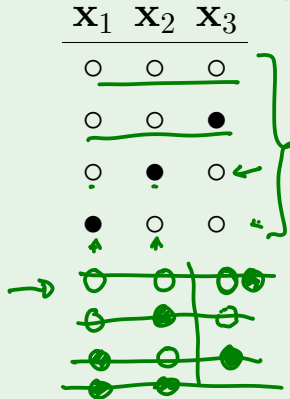
This means that \mathcal{H} can not generate the $2^2 = 4$ possible dichotomies for any subset of two samples $\{x_1, x_2\}$.

Under such supposition, how many dichotomies are possible when we consider three samples $\{x_1, x_2, x_3\}$?

Puzzle

$K=2$ is a
break point
↑

↙ 4 dichotomies



Main result

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

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

1. We started searching a replacement for M

$$P\left(|E_{in}(g) - E_{out}(g)| > \epsilon\right) \leq 2\underline{M}e^{-2\epsilon^2 N}$$

2. Dichotomies: to deal with the issue of overlapping “bad” events.
 - The complexity of \mathcal{H} is related to the number of dichotomies it can generate
3. Growth function: number of dichotomies for each N
 - Polynomial growth functions are good candidate for replacing M
 - Not always possible to write this function

4. Break-point: if it is finite, it means that the growth function is polynomial (to be demonstrated)

5. Next meeting

- if there is a finite break-point, then the growth function is polynomial
- it is valid to replace M with the growth function