MAC 0460 / 5832
Introduction to Machine Learning

09 – Is learning feasible ? (Replacing M)

IME/USP (12/05/2021)

RECAP

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

"bad" event (we would like to avoid it)

$$\underbrace{|E_{in}(h) - E_{out}(h)| > \epsilon}$$

RECAP – Bounds

Probability of a "bad" event (fixed h) (Hoeffding)

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

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

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

RECAP – Our current discussion

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

M appears because of the **union bound**, which does not take the overlaps among the "bad" events into consideration

Can we find another bound that takes the overlaps into consideration ?

(and also works for inifinite Hypothesis set?)

RECAP – 3 concepts

Dicothomies ✓

Growth function ✓

Break point ✓

RECAP – Dicothomies

- $\underbrace{X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}}$ (N points)
- $\underbrace{\mathcal{H}}$: a hypothesis space

Dichotomies generated by $\mathcal{H}$:
any bipartition of X as $X_{-1} \cup X_{+1}$ that agrees with a hypothesis $h \in \mathcal{H}$

$$\boxed{\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \Big\{ \big(h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_N) \big) \mid h \in \mathcal{H} \Big\}}$$

We know that $\underbrace{|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)| \leq 2^N}$

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

Perceptron 2D: $m_{\mathcal{H}}(3) = 8 = 2^3$, $m_{\mathcal{H}}(4) = 14 < 2^4$ $m_{\mathcal{H}}(5) = ?$

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

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

Perceptron 2D: $k = 4$ is a break point $14 < 2^4$

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 $k = +\infty$

Outline

- Proof that $m_{\mathcal{H}}(N)$ is polynomial

- Proof that $m_{\mathcal{H}}(N)$ can replace M

Bounding $m_{\mathcal{H}}(N)$

To show: $m_{\mathcal{H}}(N)$ is polynomial

We show: $m_{\mathcal{H}}(N) \leq \cdots \leq \cdots \leq$ a polynomial

Key quantity:

$B(N, k)$: Maximum number of dichotomies on N points, with break point k

$B(N, k)$: Maximum number of dichotomies on N points, with break point k

Example of last meeting: Supposing $k = 2$ is a break point, we computed $B(3, k = 2) = 4$

N=3

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

		x_1	x_2	x_3
S_1	α	o	●	o
		●	o	o
S_2^+	β	o	o	●
S_2^-	B	o	o	o

(If we know $k = 2$ is a break point, then for $N = 3$ only 4 dicothomies out of 2^3 can be generated)

Computing $B(N, k)$ is too troublesome

Let us bound $B(N, k)$!

Recursive bound on $B(N, k)$

Consider the following table:

$$\underbrace{B(N,k)} = \underbrace{\alpha + 2\beta}$$

$\alpha + \beta + \beta$

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
	S_1	α	$+1$	$+1$	$\dots$	$+1$	$+1$
			-1	$+1$	$\dots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	-1	-1
			-1	$+1$	$\dots$	-1	$+1$
S_2	S_2^+	β	$+1$	-1	$\dots$	$+1$	$+1$
			-1	-1	$\dots$	$+1$	$+1$
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	$+1$	$+1$
			-1	-1	$\dots$	-1	$+1$
	S_2^-	β	$+1$	-1	$\dots$	$+1$	-1
			-1	-1	$\dots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	$+1$	-1
			-1	-1	$\dots$	-1	-1

$$B(N, k) = \alpha + 2\beta = \underbrace{\boxed{\alpha + \beta}}_{} + \underbrace{\beta}_{} \le \underbrace{\quad}_{} + \underbrace{\quad}_{}$$

Estimating α and β

Focus on $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N-1}$ columns:

$$\underbrace{\alpha + \beta \ \le \ B(N-1, k)}_{}$$

$$\alpha + \beta > B(N-1, k)$$

$N-1$

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
			$+1$	$+1$	$\ldots$	$+1$	$+1$
			-1	$+1$	$\ldots$	$+1$	-1
	S_1	α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	-1	-1
			-1	$+1$	$\ldots$	-1	$+1$
			$+1$	-1	$\ldots$	$+1$	$+1$
			-1	-1	$\ldots$	$+1$	$+1$
	S_2^+	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	$+1$
S_2			-1	-1	$\ldots$	-1	$+1$
			$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	$+1$	-1
	S_2^-	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	-1	-1

Estimating β by itself

Now, focus on the $S_2 = S_2^+ \cup S_2^-$ rows:

$$\beta \;\le\; B(N-1, k-1)$$

can not $\cancel{\beta = 2^{k-1}}$

N-1

k-1

x_N

	# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
		$+1$	$+1$	...	$+1$	$+1$
		-1	$+1$	...	$+1$	-1
S_1	α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$+1$	-1	...	-1	-1
		-1	$+1$	...	-1	$+1$
		$+1$	-1	...	$+1$	$+1$
		-1	-1	...	$+1$	$+1$
S_2^+	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$+1$	-1	...	$+1$	$+1$
		-1	-1	...	-1	$+1$
		$+1$	-1	...	$+1$	-1
		-1	-1	...	$+1$	-1
S_2^-	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$+1$	-1	...	$+1$	-1
		-1	-1	...	-1	-1

β

2^k

Putting it together

$$B(N,k) = \alpha + 2\beta$$

$$\boxed{\alpha + \beta \le B(N-1,k)}$$

$$\boxed{\beta \le B(N-1,k-1)}$$

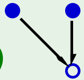
$$\boxed{B(N,k) \le B(N-1,k) + B(N-1,k-1)}$$

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
S_1		α	$+1$	$+1$	$\ldots$	$+1$	$+1$
			-1	$+1$	$\ldots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	-1	-1
			-1	$+1$	$\ldots$	-1	$+1$
S_2	S_2^+	β	$+1$	-1	$\ldots$	$+1$	$+1$
			-1	-1	$\ldots$	$+1$	$+1$
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	$+1$
			-1	-1	$\ldots$	-1	$+1$
	S_2^-	β	$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	-1	-1

Numerical computation of $B(N,k)$ bound

$$B(N,k) \le \boxed{\boxed{B(N-1,k)} + B(N-1,k-1)}$$

$B(2,2) +$
$B(2,1)$



$B(N=1,k)$

		k						
		1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1	3	4	4	4	4	..
	3	1	4	7	8	8	8	..
N	4	1	5	11	..	..	..	..
	5	1	6	:	.			
	6	1	7	:		.		
	:	:	:	:			.	

$B(N,k=1)$

$B(3,2)$
N k

$2^1 = 2$

Analytic solution for $B(N,k)$ bound

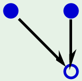
$$B(N,k) \le B(N-1,k) + B(N-1,k-1)$$

Theorem:

$$\boxed{B(N,k) \le \sum_{i=0}^{k-1} \binom{N}{i}}$$

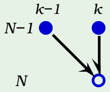
1. Boundary conditions: easy

		k						
		1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1						
	3	1						
N	4	1						
	5	1						
	6	1						
	:	:						



2. The induction step

$$\boxed{\sum_{i=0}^{k-1}\binom{N}{i} = \sum_{i=0}^{k-1}\binom{N-1}{i} + \sum_{i=0}^{k-2}\binom{N-1}{i}} \; ?$$



$$= 1+\sum_{i=1}^{k-1}\binom{N-1}{i} + \sum_{i=1}^{k-1}\binom{N-1}{i-1}$$

$$= 1+\sum_{i=1}^{k-1}\left[\binom{N-1}{i} + \binom{N-1}{i-1}\right]$$

$$= 1+\sum_{i=1}^{k-1}\binom{N}{i} = \sum_{i=0}^{k-1}\binom{N}{i} \; \checkmark$$

It is polynomial!

For a given $\mathcal{H}$, the break point k is fixed

$$\boxed{m_{\mathcal{H}}(N) \leq \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{maximum power is } N^{k-1}}}$$

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Three examples

$$\sum_{i=0}^{k-1} \binom{N}{i}$$

- $\mathcal{H}$ is positive rays: (break point $k = 2$)

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

- $\mathcal{H}$ is positive intervals: (break point $k = 3$)

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

- $\mathcal{H}$ is 2D perceptrons: (break point $k = 4$)

$$m_{\mathcal{H}}(N) = \; ? \;\; \leq \;\; \tfrac{1}{6}N^3 + \tfrac{5}{6}N + 1$$

When there is a break point k,

the effective number of hypothesis

is bounded by a polynomial of order N^{k-1}

Outline

- Proof that $m_{\mathcal{H}}(N)$ is polynomial

- Proof that $m_{\mathcal{H}}(N)$ can replace M

What we want

Instead of:

$$\mathbb{P}[\ \underbrace{|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon}\ ] \ \leq \ \underbrace{2 \quad {\color{red}M} \quad e^{-2\epsilon^2 N}}$$

We want:

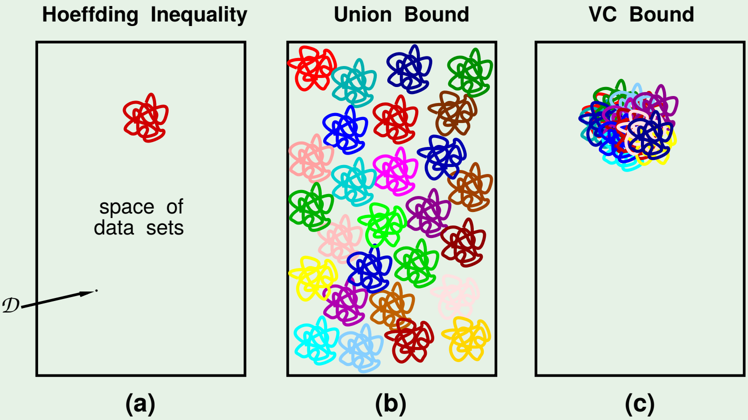
$$\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \cancel{\leq} \ 2 \ \underbrace{{\color{red}m_{\mathcal{H}}}(N)} \ e^{-2\epsilon^2 N}$$

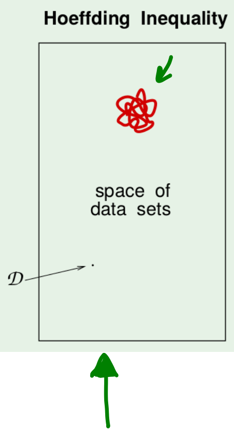
Pictorial proof ☺

- How does $m_{\mathcal{H}}(N)$ relate to overlaps?

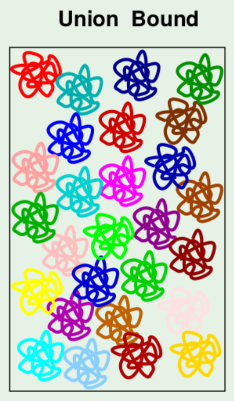
- What to do about E_{out}?

- Putting it together

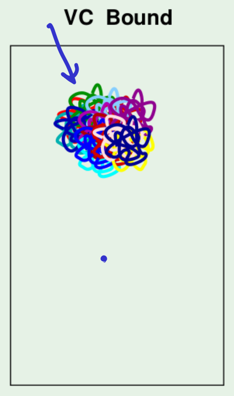

Hoeffding Inequality
Union Bound
VC Bound
space of
data sets
$\mathcal{D}$
(a)
(b)
(c)



- The canvas is the space of all possible datasets of size N
- Each point in the canvas is a dataset of size N
- Given a hypothesis h, one can compute $E_{in}(h)$ with respect to each dataset
- The red points are the "bad" events for h (i.e., $|E_{in}(h) - E_{out}(h)| > \epsilon$)
- According to Hoeffding the probability of "bad" event of h is bounded (thus only a small area of the canvas is painted red)

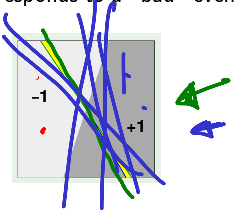


- When we have multiple hypothesis, we should consider the probability of "bad" events associated to all of them
- Each color in the canvas corresponds to points that are the "bad" events for a specific h (i.e., $|E_{in}(h) - E_{out}(h)| > \epsilon$)
- Since we are considering the union bound (no overlaps between "bad" events) a large area of the canvas is colored (as "bad" events)



- It is very reasonable to think that one dataset corresponds to a "bad" event for multiple hypothesis

- For instance, the two separating lines could have $E_{in} = 0$ and both have large error (the same dataset corresponds to a "bad" event for both)



- Considering the overlaps, the canvas painting should look like the one at the left, suggesting a bound larger than the original Hoeffding bound but much smaller than the union bound

Many hypotheses share the same dichotomy on a given $\mathcal{D}$, since there are finitely many dichotomies even with an infinite number of hypotheses. Any statement based on $\mathcal{D}$ alone will be simultaneously true or simultaneously false for all the hypotheses that look the same on that particular $\mathcal{D}$. What

The growth function "groups hypotheses" according to their behavior on D

This establishes the link between overlaps and dicothomies

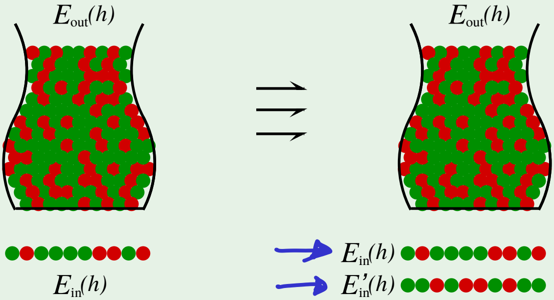
The event $|E_{in}(h) - E_{out}(h)| > \epsilon$ depends not only on D, but also on the entire space $\mathcal{X}$

Since we are "grouping" hypotheses based on their behavior on D, how do we deal with $E_{out}(h)$??

$\mathcal{D}$

$\mathcal{D}'$

What to do about E_{out}



$$|E_{in}(h) - E_{out}(h)| \approx |E_{in}(h) - E'_{in}(h)|$$

VC inequality

$$P\left(\left|E_{in}(g) - E_{out}(g)\right| > \varepsilon \right) \leq 2\, \underline{m_{\mathcal{H}}(N)}\, e^{-2\varepsilon^2 N}$$

$$\boxed{P\left(\left|E_{in}(g) - E_{out}(g)\right| > \varepsilon \right) \leq 4\, m_{\mathcal{H}}(\underline{2N})\, e^{-\frac{1}{8}\varepsilon^2 N}}$$

$\underline{2N}$:

- hypotheses are grouped based on their behavior on D, but their behavior outside D is not the same
- to track $\underline{|E_{in}(h) - E_{out}(h)| > \epsilon}$, we track $\underline{|E_{in}(h) - E'_{in}(h)| > \epsilon}$ (relative to D and D', both of size N)

N N

$\underline{4}$ and $\underline{\frac{1}{8}}$:

- these are factors to account for the uncertainties added when we replace $\underbrace{|E_{in}(h) - E_{out}(h)| > \epsilon \text{ with } |E_{in}(h) - E'_{in}(h)| > \epsilon}$

Summary

- The growth function (counts number of dichotomies) is polynomially bounded if $\mathcal{H}$ has a break point

- The growth function can replace M

- Main result: VC inequality

$$\boxed{P\left(\left| E_{in}(g) - E_{out}(g) \right| > \varepsilon \right) \leq 4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\varepsilon^2 N}}$$

  Again, we have a bound that can be made small enough by taking a sufficiently large N

- Next meeting: (i) Do we need to have the growth function ? (ii) Sample complexity