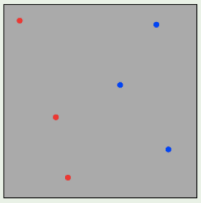
CSE4088 Introduction to Machine Learning

Theory of Generalization

Review of Last Lecture

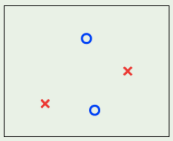
- Dichotomies



- Growth function

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

- Break point



- Maximum # of dichotomies

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

Outline

- Show that the growth function with a break point is polynomial.

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

- Show that we can put the growth function in Hoeffding's inequality in place of M.

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

Recursive bound on $B(N,k)$

Consider the following table:

We will try to put as many points as we can under The constraint that there is a break point.

x_1	x_2	...	x_{N-1}	x_N
$+1$	$+1$	...	$+1$	$+1$
-1	$+1$	...	$+1$	-1

Recursive bound on $B(N,k)$

Consider the following table:

S_1: the rows that appear only once as far as $x_1,\dots,x_N$ are concerned. Rows with only one extension x_N.

S_2+: rows with the positive extension.

S_2-: same rows with the negative extension.

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

	# of rows	x_1	x_2	...	x_{N-1}	x_N
S_1	α	$+1$	$+1$	...	$+1$	$+1$
		-1	$+1$	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	-1	-1
		-1	$+1$	...	-1	$+1$
S_2	β	$+1$	-1	...	$+1$	$+1$
S_2^+		-1	-1	...	$+1$	$+1$
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	$+1$
		-1	-1	...	-1	$+1$
S_2	β	$+1$	-1	...	$+1$	-1
S_2^-		-1	-1	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	-1
		-1	-1	...	-1	-1

Estimating α and β

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

Different rows are shown in black.

We can not have all possible patterns on any k columns of the small matrix.

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

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

Estimating β by itself

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

If we had k-1 columns with all possible patterns, then we would add x_N with +1 and -1 we would have all possible k patterns in the big matrix, which we know does not exist. Therefore the break point for beta rows is k-1.

$$\beta \leq B(N-1, k-1)$$

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

Putting it together

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

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

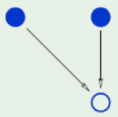
$$\beta \leq B(N-1,k-1)$$

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

	# of rows	x_1	x_2	...	x_{N-1}	x_N
S_1	α	$+1$	$+1$	...	$+1$	$+1$
		-1	$+1$	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	-1	-1
		-1	$+1$	...	-1	$+1$
S_2		$+1$	-1	...	$+1$	$+1$
S_2^+	β	-1	-1	...	$+1$	$+1$
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	$+1$
		-1	-1	...	-1	$+1$
S_2^-	β	$+1$	-1	...	$+1$	-1
		-1	-1	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	-1
		-1	-1	...	-1	-1

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

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



N \ 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	..
4	1	5	11	..	..	..	..
5	1	6	:	.			
6	1	7	:		.		
:	:	:	:			.	

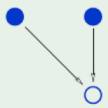
Analytic solution for $B(N,k)$ bound

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

Theorem:

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

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



1. Boundary conditions: easy

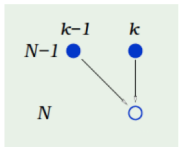
2. The induction step

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

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

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

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



It is polynomial!

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

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

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 $k = 4$)
$$m_{\mathcal{H}}(N) = \ ? \ \leq \tfrac{1}{6}N^3 + \tfrac{5}{6}N + 1$$

Outline

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

What we want

Instead of:

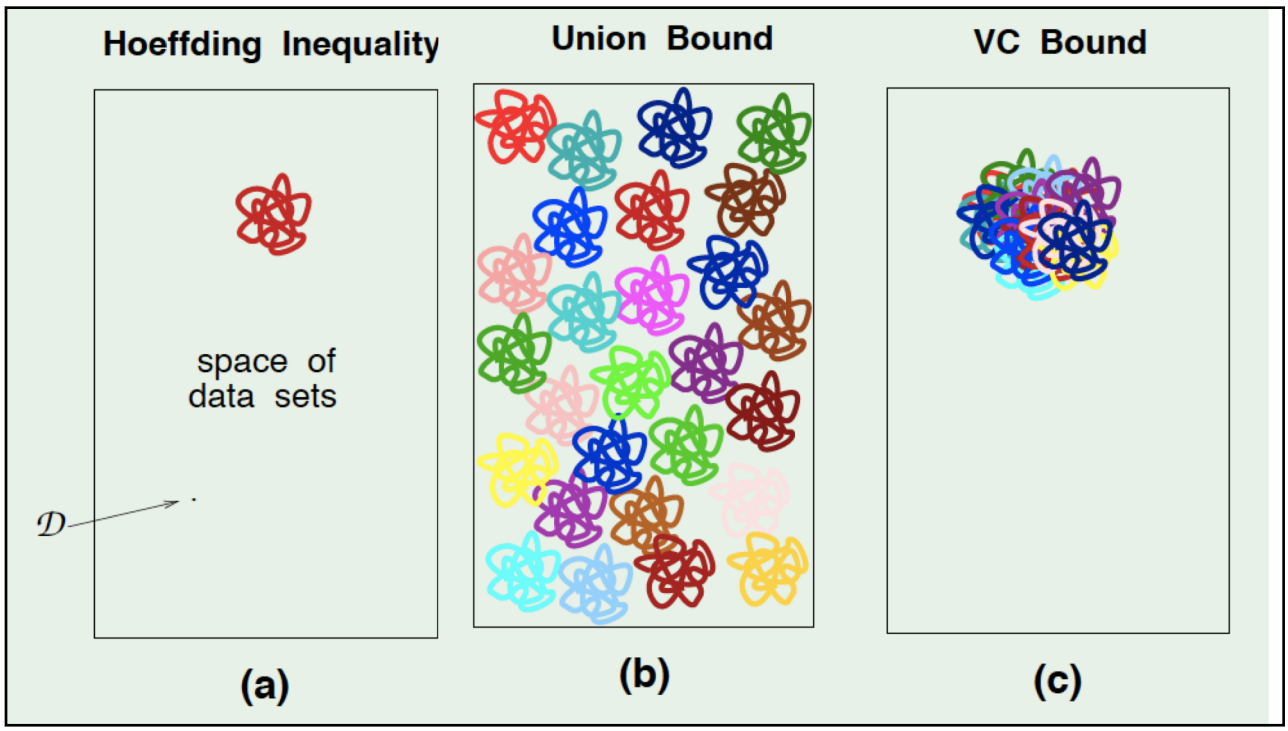
$$\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq\ 2 \quad M \quad e^{-2\epsilon^2 N}$$

We want:

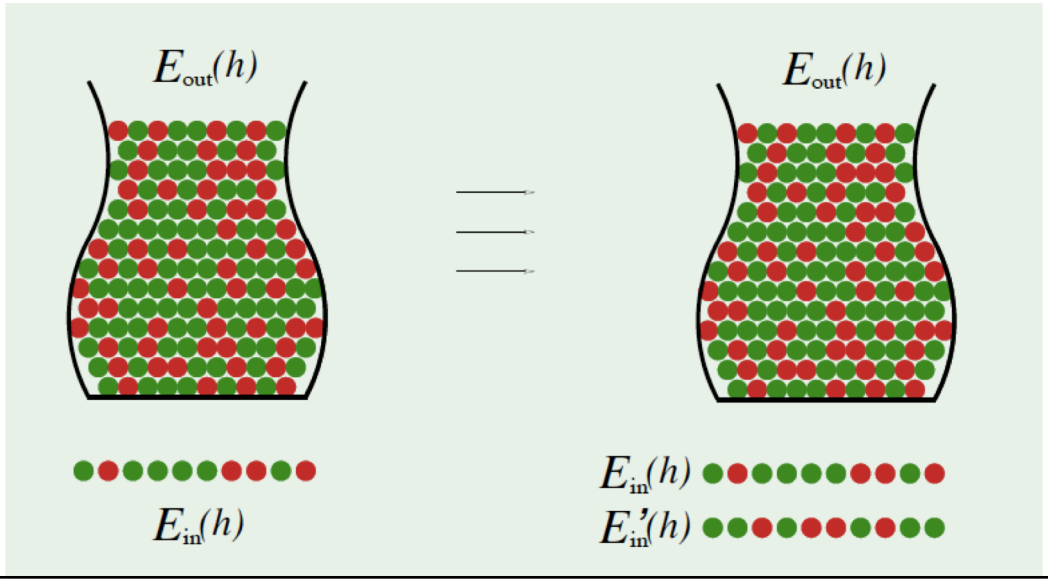
$$\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq\ 2\ m_{\mathcal{H}}(N)\ e^{-2\epsilon^2 N}$$

Pictorial proof ☺

- How does $m_{\mathcal{H}}(N)$ relate to overlaps?
- What to do about E_{out}?
- Putting it together



What to do about E_{out}



Putting it together

Not quite:

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

but rather:

$$\mathbb{P}[\,|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\,] \;\leq\; 4\;\; m_{\mathcal{H}}(2N)\;\; e^{-\,\frac{1}{8}\,\epsilon^2 N}$$

The Vapnik-Chervonenkis Inequality

MOST IMPORTANT THEORETICAL RESULT IN MACHINE LEARNING