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

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)

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

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.

Our current discussion

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

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

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

(and also works for inifinite Hypothesis set?)

Dicothomies

Let

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

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

$$\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \left\{ \left(h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_N) \right) \mid h \in \mathcal{H} \right\}$$

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

Growth function

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

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

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

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

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

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: $m_{\mathcal{H}}(N) \le \cdots \le \cdots \le$ 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 $B(3, k) = 4$

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

Recursive bound on $B(N,k)$

Consider the following table:

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

		# 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

Estimating α and β

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

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

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

Estimating β by itself

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

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

		# 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

Putting it together

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

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

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

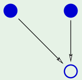
		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
			$+1$	$+1$	$\dots$	$+1$	$+1$
			-1	$+1$	$\dots$	$+1$	-1
	S_1	α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	-1	-1
			-1	$+1$	$\dots$	-1	$+1$
			$+1$	-1	$\dots$	$+1$	$+1$
			-1	-1	$\dots$	$+1$	$+1$
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

$$B(N,k) \le$$

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

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

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



		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	:		.		
	:	:	:	:			.	

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

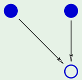
$B(N,k) \leq B(N-1,k) + B(N-1,k-1)$

Theorem:

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

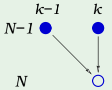
1. Boundary conditions: easy

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



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

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

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:

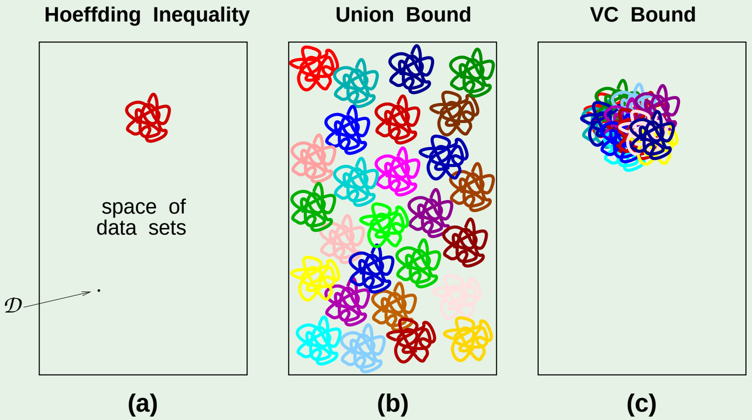
$$\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq\ 2\ m_{\mathcal{H}}(N)\ e^{-2\epsilon^2 N}$$

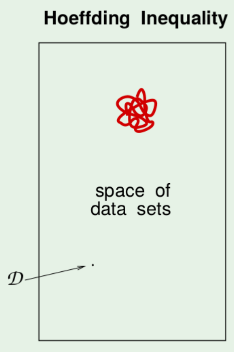
Pictorial proof ☺

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

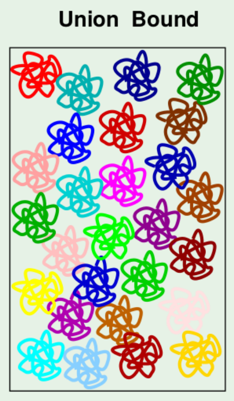
- What to do about E_{out}?

- Putting it together

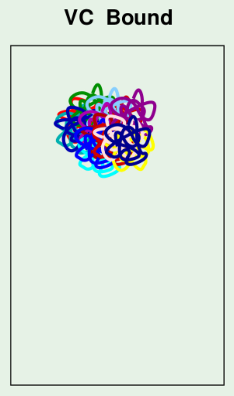

Hoeffding Inequality
Union Bound
VC Bound
space of
data sets
𝒟
(a)
(b)
(c)



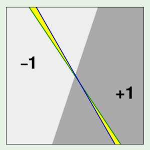
- The canvas is the space of all possible datasets of size N
- Each point in the canvas is a dataset of size N
- Given a hypothesis h, one can compute $E_{in}(h)$ with respect to each dataset
- The red points are the "bad" events for h (i.e., $|E_{in}(h) - E_{out}(h)| > \epsilon$)
- According to Hoeffding 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 correspond 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 "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 tha the original Hoeffding bound but much smaller than the union bound

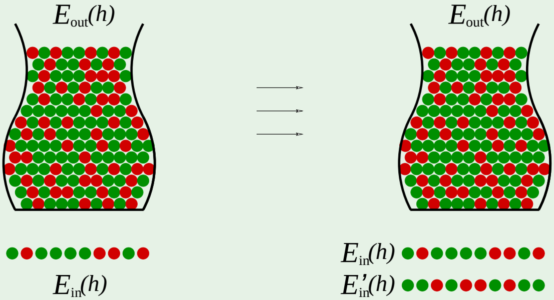
Many hypotheses share the same dichotomy on a given $\mathcal{D}$, since there are finitely many dichotomies even with an infinite number of hypotheses. Any statement based on $\mathcal{D}$ alone will be simultaneously true or simultaneously false for all the hypotheses that look the same on that particular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

The growth function groups hypotheses based on their behavior on D

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

Now we need to consider the event $|E_{in}(h) - E_{out}(h)| > \epsilon$ in terms of a group of hypothesis ...

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

VC inequality

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

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

$2N$:

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

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

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

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

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