CS260: Machine Learning Algorithms

Lecture 7: VC Dimension

Cho-Jui Hsieh
UCLA

Jan 30, 2019

Reducing M to finite number

Where did the M come from?

- The $\mathcal{B}$ad events $\mathcal{B}_m$:
  "$|E_{\text{tr}}(h_m) - E(h_m)| > \epsilon$" with probability $\leq 2e^{-2\epsilon^2 N}$

Where did the M come from?

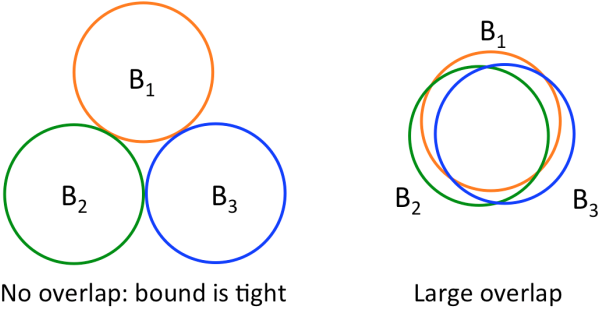
- The $\mathcal{B}$ad events $\mathcal{B}_m$:

  "$|E_{\text{tr}}(h_m) - E(h_m)| > \epsilon$" with probability $\leq 2e^{-2\epsilon^2 N}$

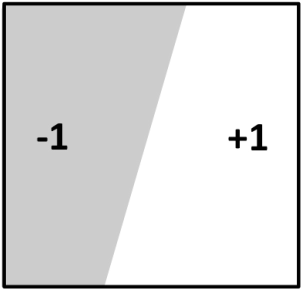
- The union bound:

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

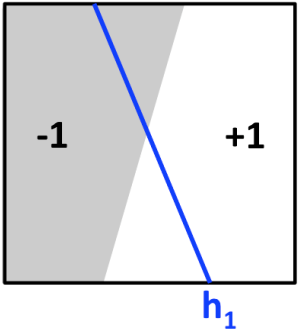
  $$\leq \underbrace{\mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \cdots + \mathbb{P}[\mathcal{B}_M]}_{\text{consider worst case: no overlaps}} \leq 2Me^{-2\epsilon^2 N}$$



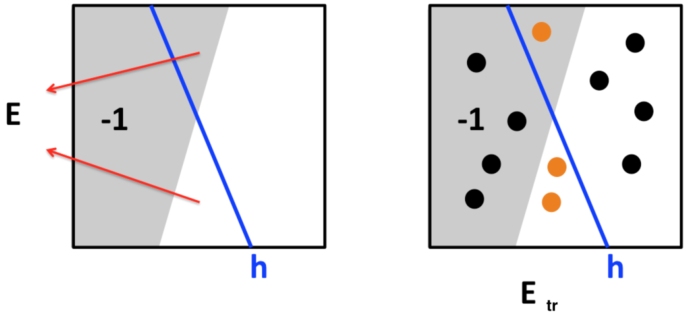
Can we improve on M?



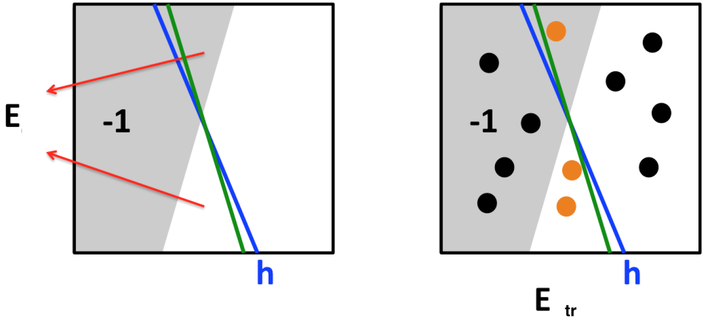
Can we improve on M?



Can we improve on M?



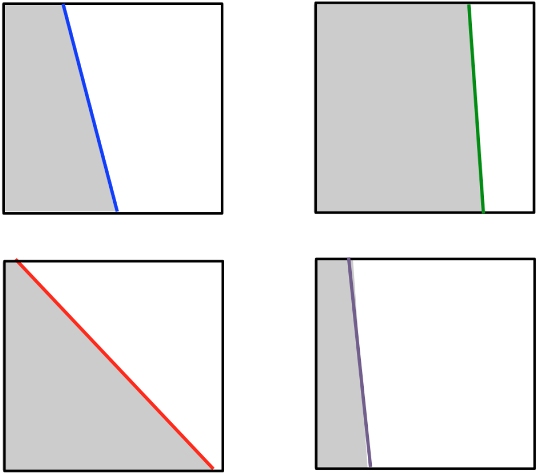
Can we improve on M?



- The event that $|E_{\text{tr}}(h_1) - E(h_1)| > \epsilon$ and $|E_{\text{tr}}(h_2) - E(h_2)| > \epsilon$ are largely overlapped.

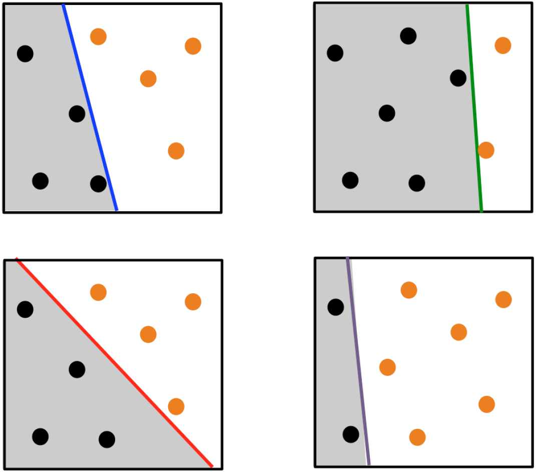
What can we replace M with?

Instead of the whole input space



What can we replace M with?

Instead of the whole input space
Let's consider a finite set of input points

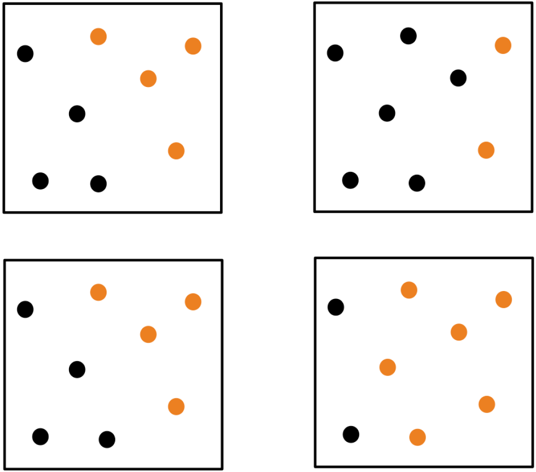


What can we replace M with?

Instead of the whole input space

Let's consider a finite set of input points

How many patterns of colors can you get?



Dichotomies: mini-hypotheses

- A hypothesis: $h : \mathcal{X} \to \{-1, +1\}$
- A dichotomy: $h : \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\} \to \{-1, +1\}$

Dichotomies: mini-hypotheses

- A hypothesis: $h : \mathcal{X} \to \{-1, +1\}$
- A dichotomy: $h : \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\} \to \{-1, +1\}$
- Number of hypotheses $|\mathcal{H}|$ can be infinite
- Number of dichotomies $|\mathcal{H}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)|$:

  at most 2^N

Dichotomies: mini-hypotheses

- A hypothesis: $h : \mathcal{X} \to \{-1, +1\}$
- A dichotomy: $h : \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\} \to \{-1, +1\}$
- Number of hypotheses $|\mathcal{H}|$ can be infinite
- Number of dichotomies $|\mathcal{H}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)|$:

  at most 2^N

  $\Rightarrow$Candidate for replacing M

The growth function

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

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

The growth function

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

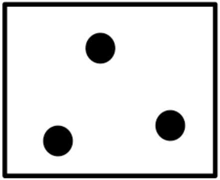
$$m_{\mathcal{H}}(N) = \max_{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N \in \mathcal{X}} |\mathcal{H}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)|$$

- The growth function satisfies:

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

Growth function for linear classifiers

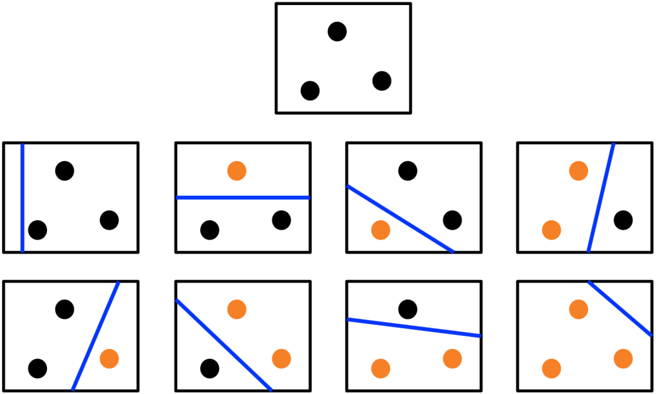
Compute $m_{\mathcal{H}}(3)$ in 2-D space



What's $|\mathcal{H}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)|$?

Growth function for linear classifiers

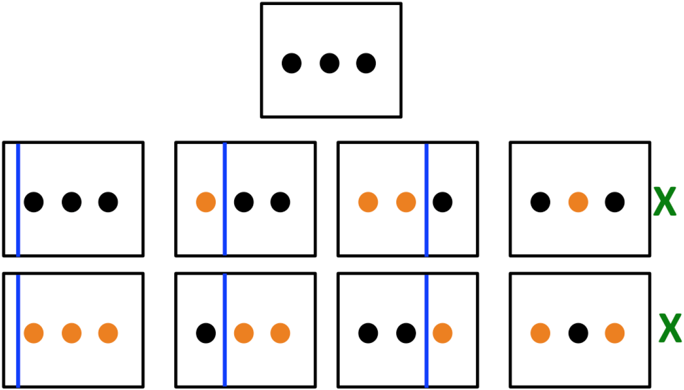
Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)



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

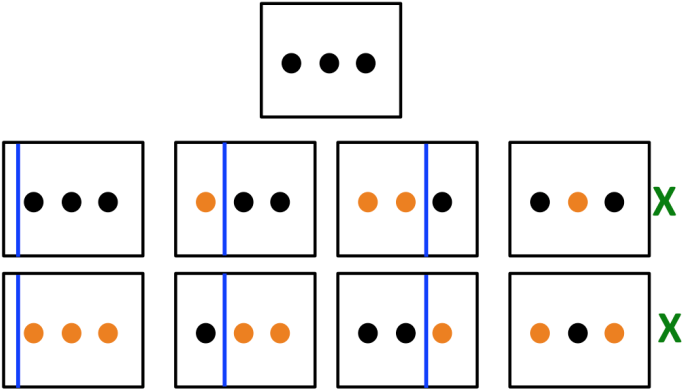
Growth function for linear classifiers

Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)



Growth function for linear classifiers

Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)



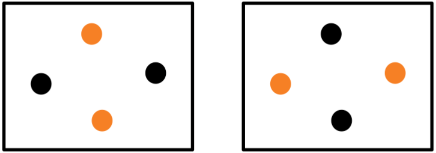
Doesn't matter because we only counts the most dichotomies

Growth function for linear classifiers

- What's $m_{\mathcal{H}}(4)$?

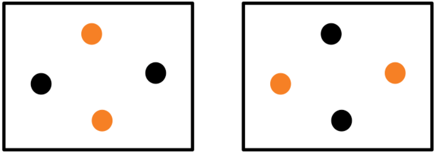
Growth function for linear classifiers

- What's $m_{\mathcal{H}}(4)$?
- (At least) missing two dichotomies:



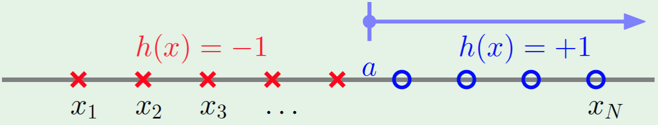
Growth function for linear classifiers

- What's $m_{\mathcal{H}}(4)$?
- (At least) missing two dichotomies:



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

Example I: positive rays

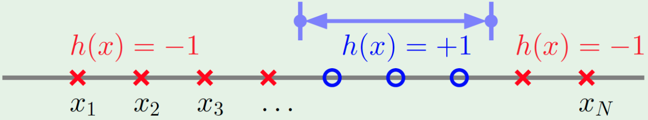


$\mathcal{H}$ is set of $h \colon \mathbb{R} \to \{-1, +1\}$

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

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

Example II: positive intervals



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

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

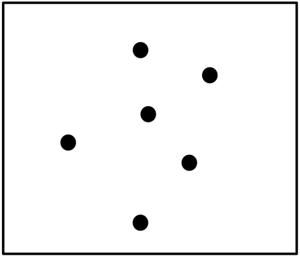
$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

Example III: convex sets

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

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

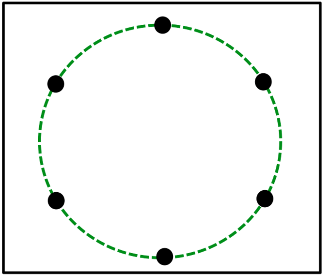
- How many dichotomies can we generate?



Example III: convex sets

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

  $h(\boldsymbol{x}) = +1$ is convex
- How many dichotomies can we generate?

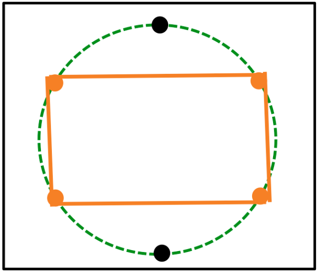


Example III: convex sets

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

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

- How many dichotomies can we generate?



Example III: convex sets

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

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

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

  $\Rightarrow$ We say the N points are "shattered" by h

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:

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

What's next?

- Remember the inequality

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

What's next?

- Remember the inequality

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

- What happens if we replace M by $m_{\mathcal{H}}(N)$?
  $m_{\mathcal{H}}(N)$ polynomial $\Rightarrow$ Good!

What's next?

- Remember the inequality

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

- What happens if we replace M by $m_{\mathcal{H}}(N)$?
  $m_{\mathcal{H}}(N)$ polynomial $\Rightarrow$ Good!
- How to show $m_{\mathcal{H}}(N)$ is polynomial?

When will $m_{\mathcal{H}}(N)$ be polynomial

Break point of $\mathcal{H}$

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

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

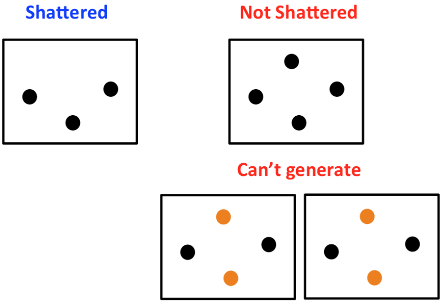
- VC dimension of $\mathcal{H}$: $k - 1$ (the most points $\mathcal{H}$ can shatter)

Break point of $\mathcal{H}$

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

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

- VC dimension of $\mathcal{H}$: $k-1$ (the most points $\mathcal{H}$ can shatter)
- For 2-D perceptron: $k=4$, VC dimension $=3$



Break point - examples

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

  Break point $k = 2$, $d_{VC} = 1$

Break point - examples

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

  Break point $k = 2$, $d_{VC} = 1$
- Positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

  Break point $k = 3$, $d_{VC} = 2$

Break point - examples

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

  Break point $k = 2$, $d_{VC} = 1$
- Positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

  Break point $k = 3$, $d_{VC} = 2$
- Convex set: $m_{\mathcal{H}}(N) = 2^N$

  Break point $k = \infty$, $d_{VC} = \infty$

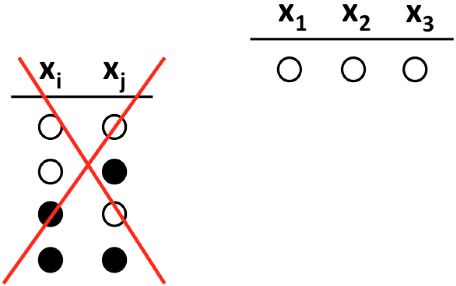
We will show

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

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

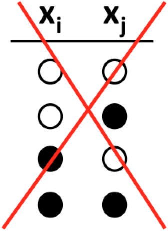
Puzzle

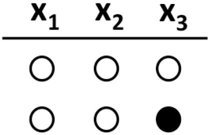
- Break point is $k = 2$



Puzzle

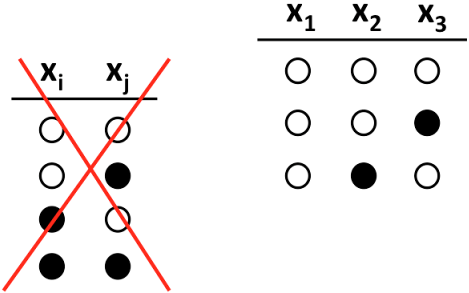
- Break point is $k = 2$





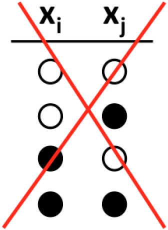
Puzzle

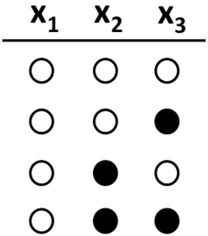
- Break point is $k = 2$



Puzzle

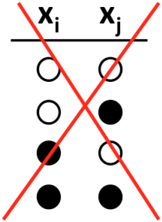
- Break point is $k = 2$

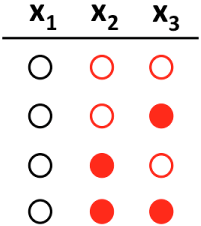




Puzzle

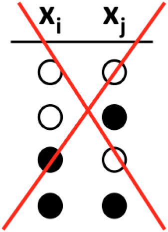
- Break point is $k = 2$

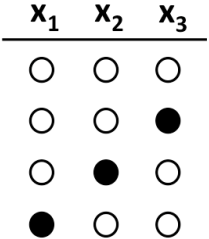




Puzzle

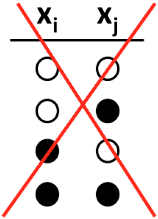
- Break point is $k = 2$

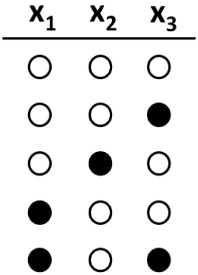




Puzzle

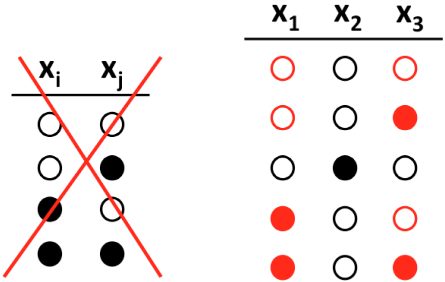
- Break point is $k = 2$





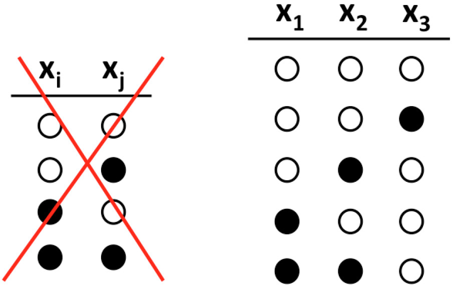
Puzzle

- Break point is $k = 2$



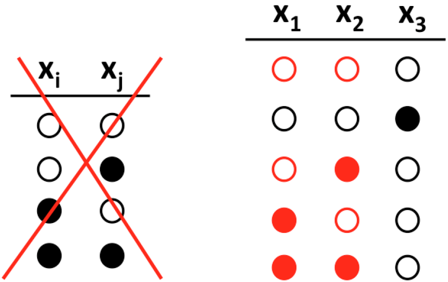
Puzzle

- Break point is $k = 2$



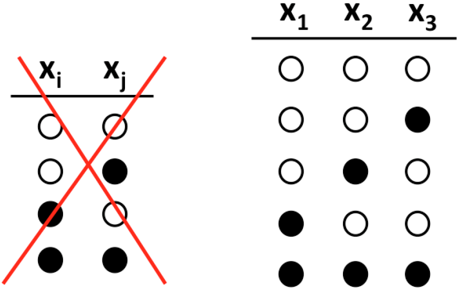
Puzzle

- Break point is $k = 2$



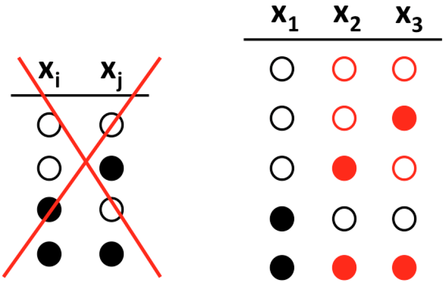
Puzzle

- Break point is $k = 2$



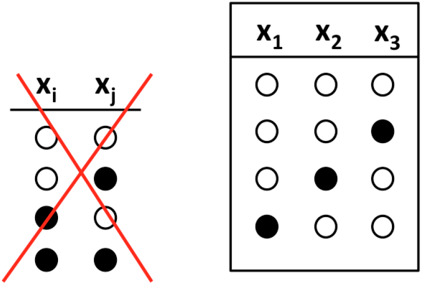
Puzzle

- Break point is $k = 2$



Puzzle

- Break point is $k = 2$



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

- Key quantity:

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

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

- Key quantity:

  $B(N, k)$: Maximum number of dichotomies on N points, with break point k
- If the hypothesis space has break point k, then

$$m_{\mathcal{H}}(N) \leq B(N, k)$$

Recursive bound on $B(N, k)$

- For any "valid" set of dichotomies, reorganize rows by
  - S_1: pattern of $x_1, \cdots, x_{N-1}$ only appears once
  - S_2^+, S_2^-: pattern of $x_1, \cdots, x_{N-1}$ appears twice

		# 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

Recursive bound on $B(N, k)$

- Focus on $x_1, x_2, \cdots, x_{N-1}$ columns:
  $\alpha + \beta \leq B(N-1, k)$

	# 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: 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

Recursive bound on $B(N, k)$

- Now focus on the $S_2 = S_2^+ \cup S^- + 2$ rows
  $\beta \leq B(N-1, k-1)$

	# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
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: 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

Recursive bound on $B(N, k)$

$$\begin{aligned} B(N,k) &= \alpha + \beta + \beta \\ &\leq B(N-1,k) + B(N-1,k-1) \end{aligned}$$

What's the upper bound for $B(N, k)$?

Recursive bound on $B(N, k)$

$$
\begin{aligned}
B(N,k) &= \alpha + \beta + \beta \\
&\le B(N-1,k) + B(N-1,k-1)
\end{aligned}
$$

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

Recursive bound on $B(N, k)$

$$
\begin{aligned}
B(N,k) &= \alpha + \beta + \beta \\
&\leq B(N-1,k) + B(N-1,k-1)
\end{aligned}
$$

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

Recursive bound on $B(N, k)$

$$\begin{aligned} B(N,k) &= \alpha + \beta + \beta \\ &\leq B(N-1,k) + B(N-1,k-1) \end{aligned}$$

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

Recursive bound on $B(N, k)$

$$
\begin{aligned}
B(N,k) &= \alpha + \beta + \beta \\
&\le B(N-1,k) + B(N-1,k-1)
\end{aligned}
$$

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

Recursive bound on $B(N,k)$

$$
\begin{aligned}
B(N,k) &= \alpha + \beta + \beta \\
&\le B(N-1,k) + B(N-1,k-1)
\end{aligned}
$$

		k					
		1	2	3	4	5	……
	1	1	2	2	2	2	……
	2	1	3	4	4	4	……
N	3	1					
	4	1					
	5	1					
	.	.					
	.	.					

Recursive bound on $B(N, k)$

$$
\begin{aligned}
B(N,k) &= \alpha + \beta + \beta \\
&\leq B(N-1,k) + B(N-1,k-1)
\end{aligned}
$$

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

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

$B(N, k)$ is upper bounded by $C(N, k)$:

$$\begin{aligned} C(N, 1) &= 1, \;\; N = 1, 2, \cdots \\ C(1, k) &= 2, \;\; k = 2, 3, \cdots \\ C(N, k) &= C(N-1, k) + C(N-1, k-1) \end{aligned}$$

- Theorem: $C(N, k) = \sum_{i=0}^{k-1} \binom{N}{i}$

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

$B(N, k)$ is upper bounded by $C(N, k)$:

$$
\begin{aligned}
&C(N, 1) = 1, \;\; N = 1, 2, \cdots \\
&C(1, k) = 2, \;\; k = 2, 3, \cdots \\
&C(N, k) = C(N-1, k) + C(N-1, k-1)
\end{aligned}
$$

- Theorem: $C(N, k) = \sum_{i=0}^{k-1} \binom{N}{i}$
- Boundary conditions: (easy to check)

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

$B(N,k)$ is upper bounded by $C(N,k)$:

$$\begin{aligned}
&C(N,1) = 1,\ \ N = 1,2,\cdots \\
&C(1,k) = 2,\ \ k = 2,3,\cdots \\
&C(N,k) = C(N-1,k) + C(N-1,k-1)
\end{aligned}$$

- Theorem: $C(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$
- Boundary conditions: (easy to check)
- Induction:

$$\underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{select } < k \text{ from } N \text{ items}} = \underbrace{\sum_{i=0}^{k-1} \binom{N-1}{i}}_{N\text{-th item not chosen}} + \underbrace{\sum_{i=0}^{k-2} \binom{N-1}{i}}_{N\text{-th item chosen}}$$

It is polynomial!

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

$$m_{\mathcal{H}}(N) \le \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{Polynomial with degree } k-1}$$

It is polynomial!

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

$$m_{\mathcal{H}}(N) \leq \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{Polynomial with degree } k}$$

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

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

It is polynomial!

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

$$m_{\mathcal{H}}(N) \leq \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{Polynomial with degree } k}$$

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

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

It is polynomial!

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

$$m_{\mathcal{H}}(N) \leq \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{Polynomial with degree } k}$$

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

$$m_{\mathcal{H}}(N) \leq \frac{1}{6}N^3 + \frac{5}{6}N + 1$$

Replace M by $m_{\mathcal{H}}(N)$

- Original bound:

$$\mathbf{P}[\exists h \in \mathcal{H} \text{ s.t. } |E_{\text{tr}}(h) - E(h)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

- Replace M by $m_{\mathcal{H}}(N)$

$$\underbrace{\mathbf{P}[\exists h \in \mathcal{H} \text{ s.t. } |E_{\text{tr}}(h) - E(h)| > \epsilon]}_{\text{BAD}} \leq 2 \cdot 2m_{\mathcal{H}}(2N) \cdot e^{-\frac{1}{8}\epsilon^2 N}$$

Vapnik-Chervonenkis (VC) bound

VC Dimension

Definition

- The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\mathrm{VC}}(\mathcal{H})$, is

  the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$

  "the most points $\mathcal{H}$ can shatter"

Definition

- The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\mathrm{VC}}(\mathcal{H})$, is

$$\text{the largest value of } N \text{ for which } m_{\mathcal{H}}(N) = 2^N$$

"the most points $\mathcal{H}$ can shatter"

- $N \leq d_{\mathrm{VC}}(\mathcal{H}) \Rightarrow \mathcal{H}$ can shatter N points

Definition

- The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\mathrm{VC}}(\mathcal{H})$, is

$$\text{the largest value of } N \text{ for which } m_{\mathcal{H}}(N) = 2^N$$

"the most points $\mathcal{H}$ can shatter"

- $N \leq d_{\mathrm{VC}}(\mathcal{H}) \Rightarrow \mathcal{H}$ can shatter N points
- $k > d_{\mathrm{VC}}(\mathcal{H}) \Rightarrow \mathcal{H}$ cannot be shattered
- The smallest **break point** is 1 above VC-dimension

The growth function

- In terms of a break point k:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

- In terms of the VC dimension d_{VC}:

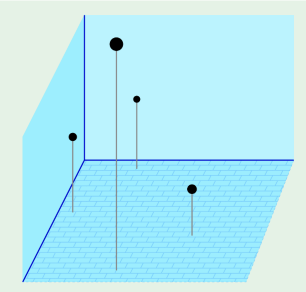
$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{\text{VC}}} \binom{N}{i}$$

VC dimension of linear classifiers

- For $d = 2$, $d_{\text{VC}} = 3$

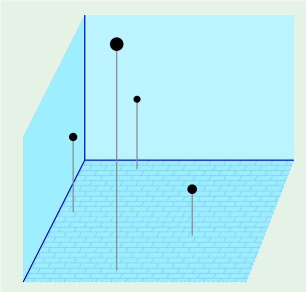
VC dimension of linear classifiers

- For $d = 2$, $d_{\mathrm{VC}} = 3$
- What if $d > 2$?



VC dimension of linear classifiers

- For $d = 2$, $d_{\mathrm{VC}} = 3$
- What if $d > 2$?

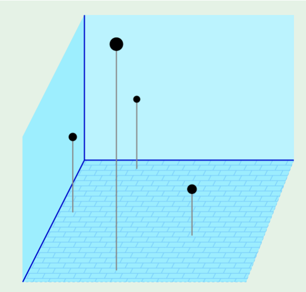


- In general,

$$d_{\mathrm{VC}} = d + 1$$

VC dimension of linear classifiers

- For $d = 2$, $d_{\text{VC}} = 3$
- What if $d > 2$?



- In general,

$$d_{\text{VC}} = d + 1$$

- We will prove $d_{\text{VC}} \geq d + 1$ and $d_{\text{VC}} \leq d + 1$

VC dimension of linear classifiers

- To prove $d_{\mathrm{VC}} \geq d + 1$

VC dimension of linear classifiers

- To prove $d_{\mathrm{VC}} \geq d + 1$
- A set of $N = d + 1$ points in $\mathbb{R}^d$ shattered by the linear hyperplane

$$
\mathrm{X} = \begin{bmatrix} -\mathbf{x}_1^\intercal- \\ -\mathbf{x}_2^\intercal- \\ -\mathbf{x}_3^\intercal- \\ \vdots \\ -\mathbf{x}_{d+1}^\intercal- \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix}
$$

VC dimension of linear classifiers

- To prove $d_{\mathrm{VC}} \geq d+1$
- A set of $N = d+1$ points in $\mathbb{R}^d$ shattered by the linear hyperplane

$$\mathrm{X} = \begin{bmatrix} -\mathbf{x}_1^{\intercal}- \\ -\mathbf{x}_2^{\intercal}- \\ -\mathbf{x}_3^{\intercal}- \\ \vdots \\ -\mathbf{x}_{d+1}^{\intercal}- \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix}$$

- X is invertible!

Can we shatter the dataset?

- For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, can we find $\boldsymbol{w}$ satisfying

$$\mathrm{sign}(X\boldsymbol{w}) = y$$

Can we shatter the dataset?

- For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, can we find $\boldsymbol{w}$ satisfying

$$\text{sign}(X\boldsymbol{w}) = y$$

- Easy! Just set $\boldsymbol{w} = X^{-1}\mathbf{y}$

Can we shatter the dataset?

- For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, can we find $\boldsymbol{w}$ satisfying

$$\text{sign}(X\boldsymbol{w}) = y$$

- Easy! Just set $\boldsymbol{w} = X^{-1}\mathbf{y}$
- So, $d_{\text{VC}} \geq d + 1$

VC dimension of linear classifiers

- To show $d_{\text{VC}} \leq d + 1$, we need to show

We cannot shatter any set of $d + 2$ points

VC dimension of linear classifiers

- To show $d_{\text{VC}} \leq d+1$, we need to show

We cannot shatter any set of $d+2$ points

- For any $d+2$ points

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{d+1}, \boldsymbol{x}_{d+2}$$

- More points than dimensions $\Rightarrow$ linear dependent

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i$$

where not all a_i's are zeros

VC dimension of linear classifiers

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i$$

- Now we construct a dichotomy that cannot be generated:

$$y_i = \begin{cases} \text{sign}(a_i) & \text{if } i \neq j \\ -1 & \text{if } i = j \end{cases}$$

VC dimension of linear classifiers

$$\boldsymbol{x}_j = \sum_{i \neq j} a_i \boldsymbol{x}_i$$

- Now we construct a dichotomy that cannot be generated:

$$y_i = \begin{cases} \text{sign}(a_i) & \text{if } i \neq j \\ -1 & \text{if } i = j \end{cases}$$

- For all $i \neq j$, assume the labels are correct: $\text{sign}(a_i) = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_i)$
$\Rightarrow a_i \boldsymbol{w}^T \boldsymbol{x}_i > 0$

VC dimension of linear classifiers

$$\mathbf{x}_j = \sum_{i \neq j} a_i \mathbf{x}_i$$

- Now we construct a dichotomy that cannot be generated:

$$y_i = \begin{cases} \text{sign}(a_i) & \text{if } i \neq j \\ -1 & \text{if } i = j \end{cases}$$

- For all $i \neq j$, assume the labels are correct: $\text{sign}(a_i) = \text{sign}(\mathbf{w}^T \mathbf{x}_i)$
  $\Rightarrow a_i \mathbf{w}^T \mathbf{x}_i > 0$
- For j-th data,

$$\mathbf{w}^T \mathbf{x}_j = \sum_{i \neq j} a_i \mathbf{w}^T \mathbf{x}_i > 0$$

- Therefore, $y_j = \text{sign}(\mathbf{w}^T \mathbf{x}_j) = +1$ (cannot be -1)

Putting it together

- We proved for d-dimensional linear hyperplane

$$d_{\mathrm{VC}} \leq d + 1 \text{ and } d_{\mathrm{VC}} \geq d + 1 \Rightarrow \qquad d_{\mathrm{VC}} = d + 1$$

Putting it together

- We proved for d-dimensional linear hyperplane

$$d_{\mathrm{VC}} \leq d+1 \text{ and } d_{\mathrm{VC}} \geq d+1 \Rightarrow \qquad d_{\mathrm{VC}} = d+1$$

- Number of parameters $w_0, \cdots, w_d$

  $d+1$ parameters!

Putting it together

- We proved for d-dimensional linear hyperplane

$$d_{\mathrm{VC}} \leq d+1 \text{ and } d_{\mathrm{VC}} \geq d+1 \Rightarrow \qquad d_{\mathrm{VC}} = d+1$$

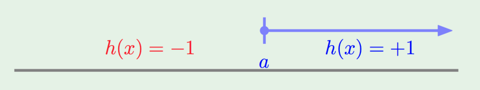
- Number of parameters $w_0, \cdots, w_d$

  $d+1$ parameters!

- Parameters create degrees of freedom

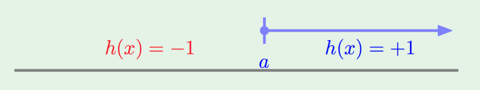
Examples

- Positive rays: 1 parameters, $d_{\mathrm{VC}} = 1$

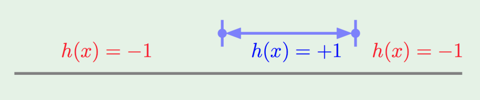


Examples

- Positive rays: 1 parameters, $d_{\mathrm{VC}} = 1$

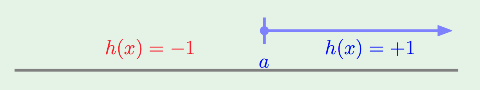


- Positive intervals: 2 parameters, $d_{\mathrm{VC}} = 2$

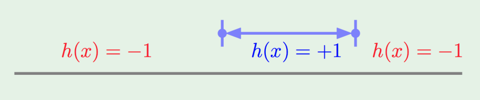


Examples

- Positive rays: 1 parameters, $d_{\mathrm{VC}} = 1$



- Positive intervals: 2 parameters, $d_{\mathrm{VC}} = 2$



- Not always true $\cdots$

  d_{VC} measures the effective number of parameters

Number of data points needed

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

- If we want certain ϵ and δ, how does N depend on d_{VC}?

Number of data points needed

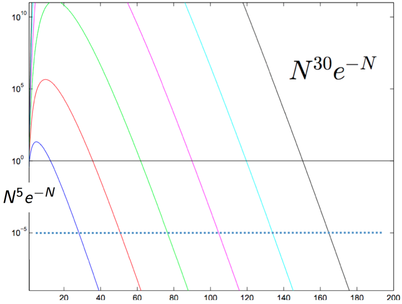
$$\mathbf{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

- If we want certain ϵ and δ, how does N depend on d_{VC}?
- Need $N^d e^{-N} =$ small value

Number of data points needed

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

- If we want certain ϵ and δ, how does N depend on d_{VC}?
- Need $N^d e^{-N}$ = small value



N is almost linear with d_{VC}

Regularization

The polynomial model

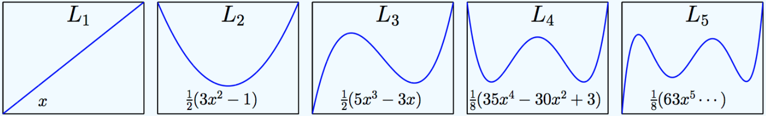
- $\mathcal{H}_Q$: polynomials of order Q

$$\mathcal{H}_Q = \{\sum_{q=0}^{Q} w_q L_q(x)\}$$

- Linear regression in the $\mathcal{Z}$ space with

$$z = [1,\ L_1(x),\ \cdots,\ L_Q(x)]$$

Legendre polynomials:



Unconstrained solution

- Input $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N) \to (\boldsymbol{z}_1, y_1), \cdots, (\boldsymbol{z}_N, y_N)$
- Linear regression:

$$\text{Minimize}: E_{\text{tr}}(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{w}^T \boldsymbol{z}_n - y_n)^2$$

$$\text{Minimize}: \frac{1}{N} (Z\boldsymbol{w} - \mathbf{y})^T (Z\boldsymbol{w} - \mathbf{y})$$

- Solution $\boldsymbol{w}_{\text{tr}} = (Z^T Z)^{-1} Z^T \mathbf{y}$

Constraining the weights

- Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$
  (with $w_q = 0$ for $q > 2$)

Constraining the weights

- Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$ (with $w_q = 0$ for $q > 2$)
- Soft-order constraint: $\sum_{q=0}^{Q} {w_q}^2 \leq C$

Constraining the weights

- Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$ (with $w_q = 0$ for $q > 2$)
- Soft-order constraint: $\sum_{q=0}^{Q} {w_q}^2 \leq C$
- The problem given soft-order constraint:

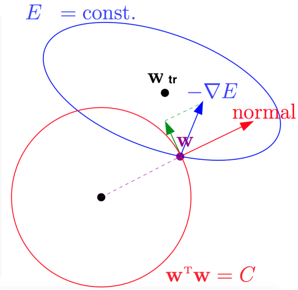
$$\text{Minimize} \frac{1}{N}(Z\boldsymbol{w} - \mathbf{y})^T(Z\boldsymbol{w} - \mathbf{y}) \quad \text{s.t.} \qquad \underbrace{\boldsymbol{w}^T\boldsymbol{w} \leq C}_{\text{smaller hypothesis space}}$$

- Solution $\boldsymbol{w}_{\text{reg}}$ instead of $\boldsymbol{w}_{\text{tr}}$

Equivalent to the unconstrained version

- Constrained version:

$$\min_{\boldsymbol{w}} E_{\text{tr}}(\boldsymbol{w}) = \frac{1}{N}(Z\boldsymbol{w} - \mathbf{y})^T(Z\boldsymbol{w} - \mathbf{y}) \quad \text{s.t.} \quad \boldsymbol{w}^T\boldsymbol{w} \leq C$$



- Optimal when

$$\nabla E_{\text{tr}}(\boldsymbol{w}_{\text{reg}}) \propto -\boldsymbol{w}_{\text{reg}}$$

Why? If $-\nabla E_{\text{tr}}(\boldsymbol{w})$ and $\boldsymbol{w}$ are not parallel, can decrease $E_{\text{tr}}(\boldsymbol{w})$ without violating the constraint

Equivalent to the unconstrained version

- Constrained version:

$$\min_{\boldsymbol{w}} E_{\text{tr}}(\boldsymbol{w}) = \frac{1}{N}(Z\boldsymbol{w} - \mathbf{y})^T(Z\boldsymbol{w} - \mathbf{y}) \ \text{ s.t. } \ \boldsymbol{w}^T\boldsymbol{w} \leq C$$

- Optimal when

$$\nabla E_{\text{tr}}(\boldsymbol{w}_{\text{reg}}) \propto -\boldsymbol{w}_{\text{reg}}$$

- Assume $\nabla E_{\text{tr}}(\boldsymbol{w}_{\text{reg}}) = -2\frac{\lambda}{N}\boldsymbol{w}_{\text{reg}}$
  $\Rightarrow \nabla E_{\text{tr}}(\boldsymbol{w}_{\text{reg}}) + 2\frac{\lambda}{N}\boldsymbol{w}_{\text{reg}} = 0$

Equivalent to the unconstrained version

- Constrained version:

$$\min_{\boldsymbol{w}} E_{\text{tr}}(\boldsymbol{w}) = \frac{1}{N}(Z\boldsymbol{w} - \mathbf{y})^T(Z\boldsymbol{w} - \mathbf{y}) \quad \text{s.t.} \quad \boldsymbol{w}^T\boldsymbol{w} \leq C$$

- Optimal when

$$\nabla E_{\text{tr}}(\boldsymbol{w}_{\text{reg}}) \propto -\boldsymbol{w}_{\text{reg}}$$

- Assume $\nabla E_{\text{tr}}(\boldsymbol{w}_{\text{reg}}) = -2\frac{\lambda}{N}\boldsymbol{w}_{\text{reg}}$
  $\Rightarrow \nabla E_{\text{tr}}(\boldsymbol{w}_{\text{reg}}) + 2\frac{\lambda}{N}\boldsymbol{w}_{\text{reg}} = 0$
- $\boldsymbol{w}_{\text{reg}}$ is also the solution of unconstrained problem

$$\min_{\boldsymbol{w}} E_{\text{tr}}(\boldsymbol{w}) + \frac{\lambda}{N}\boldsymbol{w}^T\boldsymbol{w}$$

(Ridge regression!)

Equivalent to the unconstrained version

- Constrained version:

$$\min_{\mathbf{w}} E_{\text{tr}}(\mathbf{w}) = \frac{1}{N}(Z\mathbf{w} - \mathbf{y})^T(Z\mathbf{w} - \mathbf{y}) \quad \text{s.t.} \quad \mathbf{w}^T\mathbf{w} \le C$$

- Optimal when

$$\nabla E_{\text{tr}}(\mathbf{w}_{\text{reg}}) \propto -\mathbf{w}_{\text{reg}}$$

- Assume $\nabla E_{\text{tr}}(\mathbf{w}_{\text{reg}}) = -2\frac{\lambda}{N}\mathbf{w}_{\text{reg}}$
  $\Rightarrow \nabla E_{\text{tr}}(\mathbf{w}_{\text{reg}}) + 2\frac{\lambda}{N}\mathbf{w}_{\text{reg}} = 0$
- $\mathbf{w}_{\text{reg}}$ is also the solution of unconstrained problem

$$\min_{\mathbf{w}} E_{\text{tr}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$$

  (Ridge regression!)

  $C \uparrow \quad \lambda \downarrow$

Ridge regression solution

$$\min_{\boldsymbol{w}} E_{\text{reg}}(\boldsymbol{w}) = \frac{1}{N}\left((Z\boldsymbol{w} - \mathbf{y})^T(Z\boldsymbol{w} - \mathbf{y}) + \lambda \boldsymbol{w}^T\boldsymbol{w}\right)$$

- $\nabla E_{\text{reg}}(\boldsymbol{w}) = 0 \quad \Rightarrow \quad Z^T Z(\boldsymbol{w} - \mathbf{y}) + \lambda \boldsymbol{w} = 0$

Ridge regression solution

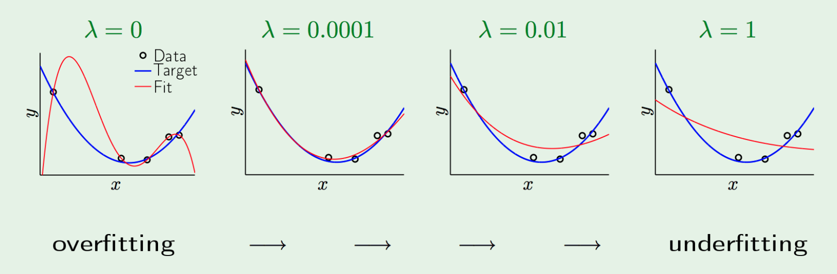
$$\min_{\mathbf{w}} E_{\text{reg}}(\mathbf{w}) = \frac{1}{N}\left((Z\mathbf{w} - \mathbf{y})^T(Z\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^T\mathbf{w}\right)$$

- $\nabla E_{\text{reg}}(\mathbf{w}) = 0 \quad \Rightarrow \quad Z^T Z(\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w} = 0$
- So, $\mathbf{w}_{\text{reg}} = (Z^T Z + \lambda I)^{-1} Z^T \mathbf{y}$ (with regularization)

  as opposed to $\mathbf{w}_{\text{tr}} = (Z^T Z)^{-1} Z^T \mathbf{y}$ (without regularization)

The result

$$\min_{\mathbf{w}} \; E_{\text{tr}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$$



Equivalent to "weight decay"

- Consider the general case

$$\min_{\boldsymbol{w}} E_{\text{tr}}(\boldsymbol{w}) + \frac{\lambda}{N}\boldsymbol{w}^T\boldsymbol{w}$$

Equivalent to "weight decay"

- Consider the general case

$$\min_{\mathbf{w}} E_{\text{tr}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$$

- Gradient descent:

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta\left(\nabla E_{\text{tr}}(\mathbf{w}_t) + 2\frac{\lambda}{N}\mathbf{w}_t\right)\\ &= \mathbf{w}_t \underbrace{\left(1 - 2\eta\frac{\lambda}{N}\right)}_{\text{weight decay}} - \eta\nabla E_{\text{tr}}(\mathbf{w}_t)\end{aligned}$$

Variations of weight decay

- Emphasis of certain weights:

$$\sum_{q=0}^{Q} \gamma_q w_q^2$$

  - Example 1: $\gamma_q = 2^q \;\Rightarrow\;$ low-order fit
  - Example 2: $\gamma_q = 2^{-q} \;\Rightarrow\;$ high-order fit

Variations of weight decay

- Emphasis of certain weights:

$$\sum_{q=0}^{Q} \gamma_q w_q^2$$

  - Example 1: $\gamma_q = 2^q \;\Rightarrow\;$ low-order fit
  - Example 2: $\gamma_q = 2^{-q} \;\Rightarrow\;$ high-order fit

- General Tikhonov regularizer:

$$\boldsymbol{w}^T H \boldsymbol{w}$$

with a positive semi-definite H

General form of regularizer

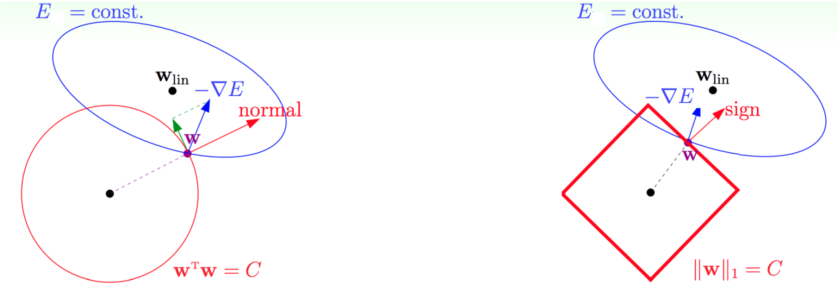
- Calling the regularizer $\Omega = \Omega(h)$, we minimize

$$E_{\text{reg}}(h) = E_{\text{tr}}(h) + \frac{\lambda}{N}\Omega(h)$$

- In general, $\Omega(h)$ can be any measurement for the "size" of h

L2 vs L1 regularizer

- L1-regularizer: $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = \sum_q |w_q|$
- Usually leads to a sparse solution
  (only few w_q will be nonzero)



Conclusions

- VC dimension
- Regularization

Questions?