



Slides adapted from Rob Schapire

# Introduction to Machine Learning

Machine Learning: Jordan Boyd-Graber

University of Maryland

RADEMACHER COMPLEXITY

## Recap

- Rademacher complexity provides nice guarantees

$$R(h) \leq \hat{R}(h) + \mathcal{R}_m(H) + \mathcal{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right) \quad (1)$$

- But in practice hard to compute for real hypothesis classes
- Is there a relationship with simpler combinatorial measures?

## Growth Function

Define the **growth function**  $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$  for a hypothesis set  $H$  as:

$$\forall m \in \mathbb{N}, \Pi_H(m) \equiv \max_{\{x_1, \dots, x_m\} \in X} |\{(h(x_1), \dots, h(x_m)) : h \in H\}| \quad (2)$$

## Growth Function

Define the **growth function**  $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$  for a hypothesis set  $H$  as:

$$\forall m \in \mathbb{N}, \Pi_H(m) \equiv \max_{\{x_1, \dots, x_m\} \in X} |\{(h(x_1), \dots, h(x_m)) : h \in H\}| \quad (2)$$

i.e., the number of ways  $m$  points can be classified using  $H$ .

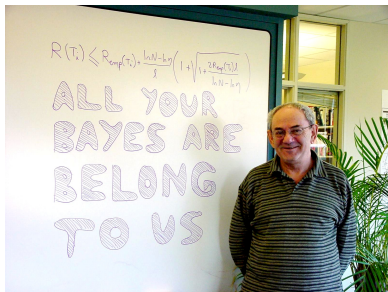
## Rademacher Complexity vs. Growth Function

If  $G$  is a function taking values in  $\{-1, +1\}$ , then

$$\mathcal{R}_m(G) \leq \sqrt{\frac{2 \ln \Pi_G(m)}{m}} \quad (3)$$

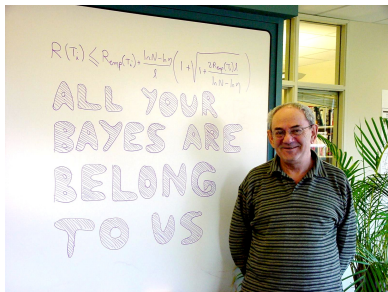
Uses Masart's lemma

## Vapnik-Chervonenkis Dimension



$$\text{VC}(H) \equiv \max \{m : \Pi_H(m) = 2^m\} \quad (4)$$

## Vapnik-Chervonenkis Dimension



$$VC(H) \equiv \max \{m : \Pi_H(m) = 2^m\} \quad (4)$$

The size of the largest set that can be fully shattered by  $H$ .

## VC Dimension for Hypotheses

- Need upper and lower bounds
- Lower bound: example
- Upper bound: Prove that no set of  $d + 1$  points can be shattered by  $H$  (harder)



## Intervals

What is the VC dimension of  $[a, b]$  intervals on the real line.

## Intervals

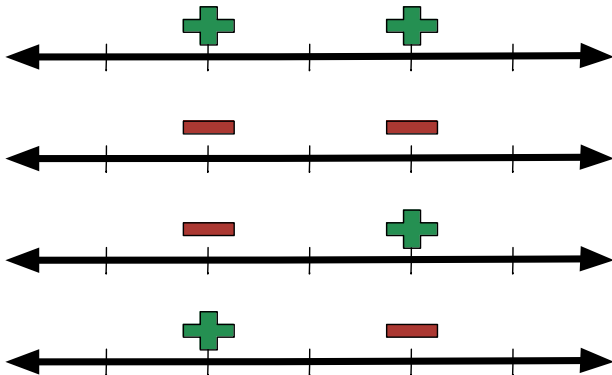
What is the VC dimension of  $[a, b]$  intervals on the real line.

- What about two points?

## Intervals

What is the VC dimension of  $[a, b]$  intervals on the real line.

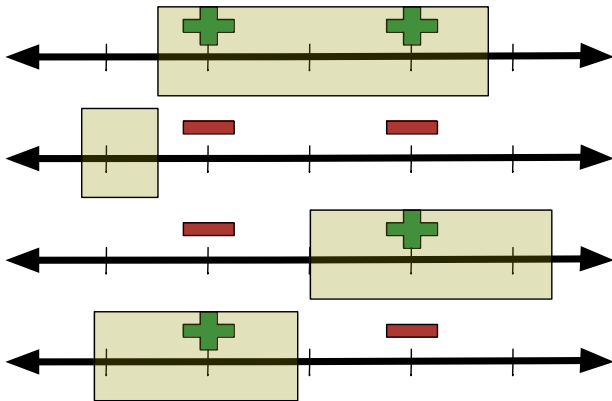
- What about two points?



## Intervals

What is the VC dimension of  $[a, b]$  intervals on the real line.

- What about two points?



## Intervals

What is the VC dimension of  $[a, b]$  intervals on the real line.

- Two points can be perfectly classified, so VC dimension  $\geq 2$

## Intervals

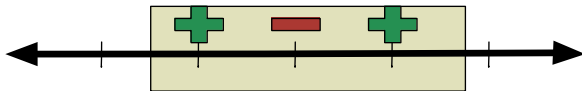
What is the VC dimension of  $[a, b]$  intervals on the real line.

- Two points can be perfectly classified, so VC dimension  $\geq 2$
- What about three points?

## Intervals

What is the VC dimension of  $[a, b]$  intervals on the real line.

- Two points can be perfectly classified, so VC dimension  $\geq 2$
- What about three points?



## Intervals

What is the VC dimension of  $[a, b]$  intervals on the real line.

- Two points can be perfectly classified, so VC dimension  $\geq 2$
- What about three points?
- **No set** of three points can be shattered



## Intervals

What is the VC dimension of  $[a, b]$  intervals on the real line.

- Two points can be perfectly classified, so VC dimension  $\geq 2$
- What about three points?
- **No set** of three points can be shattered
- Thus, VC dimension of intervals is 2

## Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \quad (5)$$

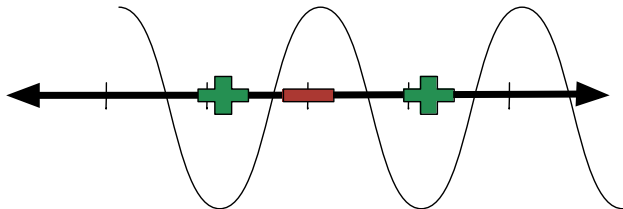
- Can you shatter three points?

## Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \quad (5)$$

- Can you shatter three points?



## Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \quad (5)$$

- Can you shatter four points?

## Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \quad (5)$$

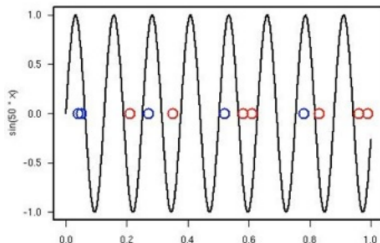
- How many points can you shatter?

## Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \quad (5)$$

- Thus, VC dim of sine on line is  $\infty$



## Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

## Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

### Theorem

**Sauer's Lemma** *Let  $H$  be a hypothesis set with VC dimension  $d$ . Then*

$\forall m \in \mathbb{N}$

$$\Pi_H(m) \leq \sum_{i=0}^d \binom{m}{i} \equiv \Phi_d(m) \quad (6)$$



## Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

### Theorem

**Sauer's Lemma** *Let  $H$  be a hypothesis set with VC dimension  $d$ . Then*

$\forall m \in \mathbb{N}$

$$\Pi_H(m) \leq \sum_{i=0}^d \binom{m}{i} \equiv \Phi_d(m) \quad (6)$$

This is good because the sum when multiplied out becomes

$\binom{m}{i} = \frac{m \cdot (m-1) \dots}{i!} = \mathcal{O}(m^d)$ . When we plug this into the learning error limits:  
 $\log(\Pi_H(2m)) = \log(\mathcal{O}(m^d)) = \mathcal{O}(d \log m)$ .

## Proof of Sauer's Lemma

Prelim:

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$$

This comes from Pascal's Triangle

$$\binom{m}{k} = 0 \quad \text{if} \quad \begin{cases} k < 0 \\ k > m \end{cases}$$

This convention is consistent with Pascal's Triangle

## Proof of Sauer's Lemma

Prelim:

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \quad \text{This comes from Pascal's Triangle}$$
$$\binom{m}{k} = 0 \quad \text{if } \begin{cases} k < 0 \\ k > m \end{cases} \quad \text{This convention is consistent with Pascal's Triangle}$$

We'll proceed by induction. Our two base cases are:

- If  $m = 0$ ,  $\Pi_H(m) = 1$ . You have no data, so there's only one (degenerate) labeling
- If  $d = 0$ ,  $\Pi_H(m) = 1$ . If you can't even shatter a single point, then it's a fixed function

## Induction Step

Assume that it holds for all  $m'$ ,  $d'$  for which  $m' + d' < m + d$ . We are given  $H$ ,  $|S| = m$ ,  $S = \langle x_1, \dots, x_m \rangle$ , and  $d$  is the VC dimension of  $H$ .

## Induction Step

Assume that it holds for all  $m'$ ,  $d'$  for which  $m' + d' < m + d$ . We are given  $H$ ,  $|S| = m$ ,  $S = \langle x_1, \dots, x_m \rangle$ , and  $d$  is the VC dimension of  $H$ .

### Build two new hypothesis spaces

	$\mathcal{H}$		$\mathcal{H}_1$		$\mathcal{H}_2$		
	$x_1, \dots, x_m$		$x_1, \dots, x_{m-1}$		$x_1, \dots, x_{m-1}$		
h1	0 1 1 0 0	→	h1	0 1 1 0	→	h1	0 1 1 0
h2	0 1 1 0 1	↗					
h3	0 1 1 1 0	→	h3	0 1 1 1			
h4	1 0 0 1 0	→	h4	1 0 0 1	→	h4	1 0 0 1
h5	1 0 0 1 1	↗					
h6	1 1 0 0 1	→	h6	1 1 0 0			

Encodes where the extended set has differences on the first  $m$  points.

## What is VC dimension of $H_1$ and $H_2$ ?

- If a set is shattered by  $H_1$ , then it is also shattered by  $H$

$$\text{VC-dim}(H_1) \leq \text{VC-dim}(H) = d \quad (7)$$

## What is VC dimension of $H_1$ and $H_2$ ?

- If a set is shattered by  $H_1$ , then it is also shattered by  $H$

$$\text{VC-dim}(H_1) \leq \text{VC-dim}(H) = d \quad (7)$$

- If a set  $T$  is shattered by  $H_2$ , then  $T \cap \{x_m\}$  is shattered by  $H$  since there will be two hypotheses in  $H$  for every element of  $H_2$  by adding  $x_m$

$$\text{VC-dim}(H_2) \leq d - 1 \quad (8)$$

## Bounding Growth Function

$$|\Pi_H(\mathcal{S})| = |H_1| + |H_2| \tag{9}$$

$$\leq \sum_{i=0}^d \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \tag{10}$$

$$\tag{11}$$



## Bounding Growth Function

$$|\Pi_H(S)| = |H_1| + |H_2| \tag{9}$$

$$\leq \sum_{i=0}^d \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \tag{10}$$

$$\tag{11}$$

We can rewrite this as  $\sum_{i=0}^d \binom{m-1}{i-1}$  because  $\binom{x}{-1} = 0$ .

## Bounding Growth Function

$$|\Pi_H(S)| = |H_1| + |H_2| \tag{9}$$

$$\leq \sum_{i=0}^d \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \tag{10}$$

$$= \sum_{i=0}^d \left[ \binom{m-1}{i} + \binom{m-1}{i-1} \right] \tag{11}$$

$$\tag{12}$$

## Bounding Growth Function

$$|\Pi_H(\mathcal{S})| = |H_1| + |H_2| \quad (9)$$

$$\leq \sum_{i=0}^d \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \quad (10)$$

$$= \sum_{i=0}^d \left[ \binom{m-1}{i} + \binom{m-1}{i-1} \right] \quad (11)$$

$$= \sum_{i=0}^d \binom{m}{i} \quad (12)$$

$$(13)$$

## Pascal's Triangle

## Bounding Growth Function

$$|\Pi_H(\mathcal{S})| = |H_1| + |H_2| \quad (9)$$

$$\leq \sum_{i=0}^d \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \quad (10)$$

$$= \sum_{i=0}^d \left[ \binom{m-1}{i} + \binom{m-1}{i-1} \right] \quad (11)$$

$$= \sum_{i=0}^d \binom{m}{i} \quad (12)$$

$$= \Phi_d(m) \quad (13)$$

## Wait a minute ...

Is this combinatorial expression really  $\mathcal{O}(m^d)$ ?

$$\begin{aligned}\sum_{i=0}^d \binom{m}{i} &\leq \sum_{i=0}^d \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\ &\leq \sum_{i=0}^m \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\ &= \left(\frac{m}{d}\right)^d \sum_{i=0}^m \binom{m}{i} \left(\frac{d}{m}\right)^i \\ &= \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \leq \left(\frac{m}{d}\right)^d e^d.\end{aligned}$$

## Generalization Bounds

Combining our previous generalization results with Sauer's lemma, we have that for a hypothesis class  $H$  with VC dimension  $d$ , for any  $\delta > 0$  with probability at least  $1 - \delta$ , for any  $h \in H$ ,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \quad (14)$$

**Whew!**

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension

## Whew!

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension
- And works well in practice . . .



**Whew!**

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension
- And works well in practice . . . Support Vector Machines

**Whew!**

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension
- And works well in practice . . . Support Vector Machines
- In class: more VC dimension examples