Classification: VC Dimension

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 7

Slides adapted from Rob Schapire

Recap

- Rademacher complexity provides nice guarantees

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

- 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:

$$\forall m \in \mathbb{N}, \Pi_H(m) \equiv \max_{\{x_1,\ldots,x_m\} \in X} \left| \{(h(x_1), \ldots, h(x_m) : h \in H\} \right| \tag{2}$$

Growth Function

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

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

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

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

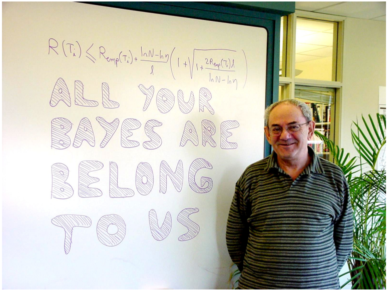
Rademacher Complexity vs. Growth Function

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

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

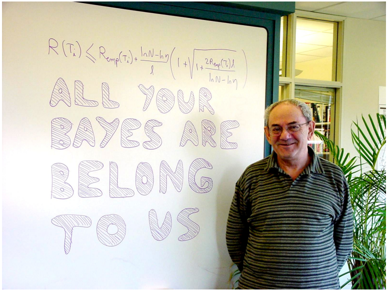
You'll prove this

Vapnik-Chervonenkis Dimension



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

Vapnik-Chervonenkis Dimension



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

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

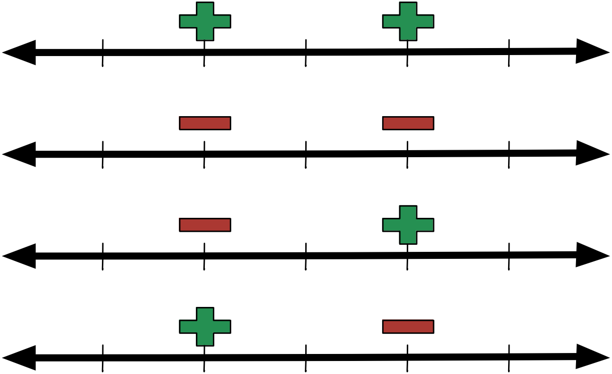
What is the VC dimension of $[a, b]$ intervals on the real line.

- What about two points?

Intervals

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

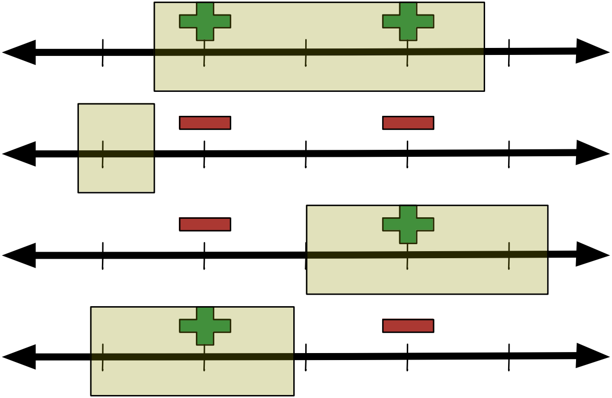
- What about two points?



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 ≥ 2

Intervals

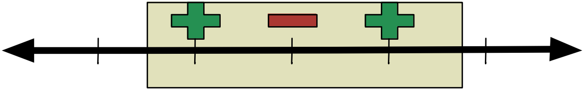
What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 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 ≥ 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 ≥ 2
- What about three points?
- Three points cannot 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 ≥ 2
- What about three points?
- Three points cannot 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

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

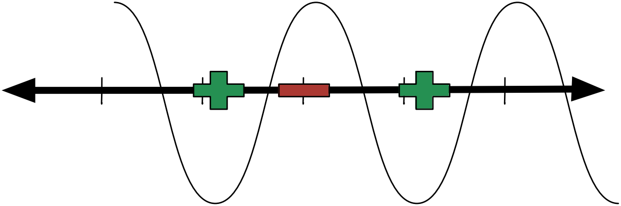
- Can you shatter three points?

Sine Functions

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

$$\{t \longrightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{5}$$

- Can you shatter three points?



Sine Functions

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

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

- Can you shatter four points?

Sine Functions

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

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

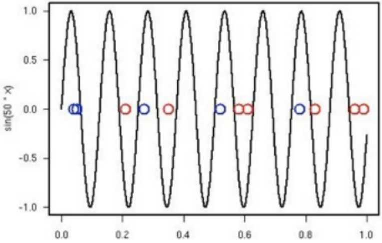
- How many points can you shatter?

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 ∞



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

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

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

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

This is good because the sum when multiplied out becomes $\binom{m}{i} = \frac{m \cdot (m-1) \dots}{i!} = \mathcal{O}\left(m^d\right)$. When we plug this into the learning error limits: $\log(\Pi_H(2m)) = \log(\mathcal{O}\left(m^d\right)) = \mathcal{O}\left(d \log m\right)$.

Proof of Sauer's Lemma

Prelim:

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \quad \text{This comes from Pascal's Triangle}$$

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

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, $|S| = m$, $S = \langle x_1, \ldots, x_m \rangle$, 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

$\mathcal{H}$							$\mathcal{H}_1$						$\mathcal{H}_2$				
	$\mathbf{x}_1, \dots, \mathbf{x}_m$							$\mathbf{x}_1, \dots, \mathbf{x}_{m-1}$						$\mathbf{x}_1, \dots, \mathbf{x}_{m-1}$			
h1	0	1	1	0	0	$\rightarrow$	h1	0	1	1	0	$\rightarrow$	h1	0	1	1	0
h2	0	1	1	0	1	$\nearrow$											
h3	0	1	1	1	0	$\rightarrow$	h3	0	1	1	1						
h4	1	0	0	1	0	$\rightarrow$	h4	1	0	0	1	$\rightarrow$	h4	1	0	0	1
h5	1	0	0	1	1	$\nearrow$											
h6	1	1	0	0	1	$\rightarrow$	h6	1	1	0	0						

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

Bounding Growth Function

$$
\begin{aligned}
|\Pi_H(S)| &= |H_1| + |H_2| && (7) \\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (8) \\
& && (9)
\end{aligned}
$$

Bounding Growth Function

$$|\Pi_H(S)| = |H_1| + |H_2| \tag{7}$$

$$\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \tag{8}$$

$$\tag{9}$$

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

Bounding Growth Function

$$
\begin{aligned}
|\Pi_H(S)| &= |H_1| + |H_2| && (7)\\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (8)\\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] && (9)\\
& && (10)
\end{aligned}
$$

Bounding Growth Function

$$|\Pi_H(S)| = |H_1| + |H_2| \tag{7}$$

$$\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \tag{8}$$

$$= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] \tag{9}$$

$$= \sum_{i=0}^{d} \binom{m}{i} \tag{10}$$

$$\tag{11}$$

Pascal's Triangle

Bounding Growth Function

$$
\begin{aligned}
|\Pi_H(S)| &= |H_1| + |H_2| && (7)\\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (8)\\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] && (9)\\
&= \sum_{i=0}^{d} \binom{m}{i} && (10)\\
&= \Phi_d(m) && (11)
\end{aligned}
$$

Wait a minute . . .

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

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

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 $h \in H$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{12}$$

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