Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 27

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Logistics

- Friday: Practice/prep session for Quiz 3
- Monday: In-class Quiz 3
- Please also fill out the FCQs for me an Saumya

Learning Objectives

- Intro. to learning theory and VC dimension

Motivation

- Remember bias-variance trade-off?
  - The more complex/flexible a model, the more likely it is to overfit
  - The more training data we have, the less likely a model is to overfit
- What we have not talked about yet
  - How can we measure how complex/flexible a model is?
  - Given a measure of the complexity/flexibility of a model, how much data do we need?

Introduction: Complexity of a Model

- Let's think of a simple classifier: A decision boundary in a 2D space $h(x) = ax_1 + bx_2 + c$ where a, b, and c are trainable parameters
- We call a learned model a **hypothesis** $h(x)$
- The class of all hypothesis is called the **Hypothesis Space** H
- If a, b, and c are assumed to be double-precision variables, then it's usually represented in 64 bits.
- For binary classification, H then consists of at most $2^{3\times 64=192}$ different hypotheses

Introduction: Complexity of a Model

- Copying from previous slide:
  - A decision boundary in a 2D space $h(x) = ax_1 + bx_2 + c$ where a, b, and c are trainable parameters
  - For binary classification, H then consists of at most $2^{3\times 64=192}$ different hypotheses

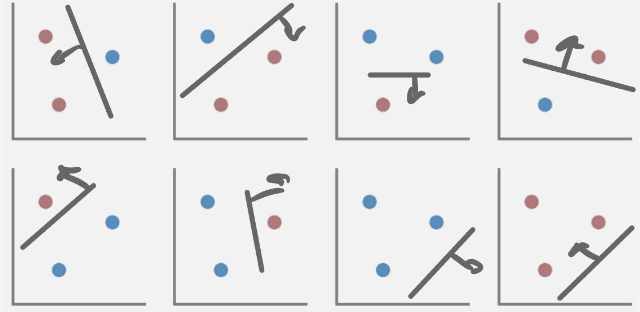
Introduction: Complexity of a Model

- Copying from previous slide:
  - A decision boundary in a 2D space $h(x) = ax_1 + bx_2 + c$ where a, b, and c are trainable parameters
  - For binary classification, H then consists of at most $2^{3\times 64=192}$ different hypotheses
- Is this helpful? Because we can equivalently express h as $h(x) = (a-d)x_1 + (b-e)x_2 + (c-f)$ which increases the number of parameters
- Alternatively, we can use the idea of **shattering**
- Def of **shattering**: A set of points S is shattered by Hypothesis Class H if H can correctly classify ALL possible labels of S

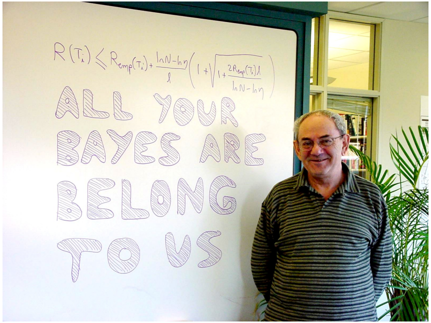
Introduction: Complexity of a Model

Def of **shattering**: A set of points S is shattered by Hypothesis Class H if H can correctly classify ALL possible labels of S

e.g., When $|S| = 3$ plotted as below, we need to look into all possible label combinations of S (assuming binary labels)

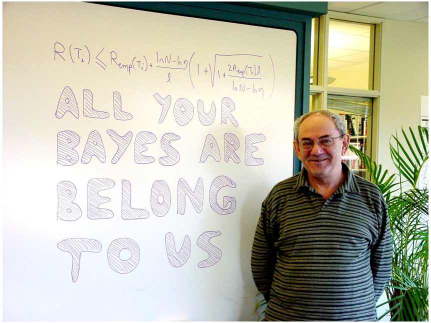


Vapnik-Chervonenkis Dimension



$$\text{VC}(H) \equiv \max\{|S| : H \text{ shatters } S\} \text{ for some } S \quad (1)$$

Vapnik-Chervonenkis Dimension



$$\text{VC}(H) \equiv \max \{|S| : H \text{ shatters } S\} \text{ for some } S \tag{1}$$

i.e., the size of the largest set S that can be fully shattered by H.

Finding VC Dimension for Hypotheses

- Need upper and lower bounds
- Lower bound: if all possible class combinations of m data points can be shattered
- Upper bound: Prove that no set of $m + 1$ data points can be shattered by H (harder)

Example of VC Dimension and Shattering: Intervals

What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

Example of VC Dimension and Shattering: Intervals

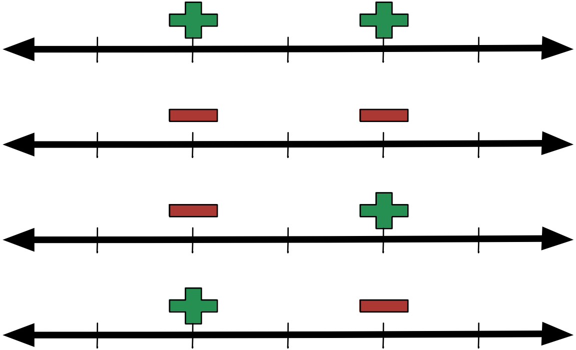
What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

$$h(x) = \begin{cases} 0, & \text{if } a \leq x \leq b \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

Example of VC Dimension and Shattering: Intervals

What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

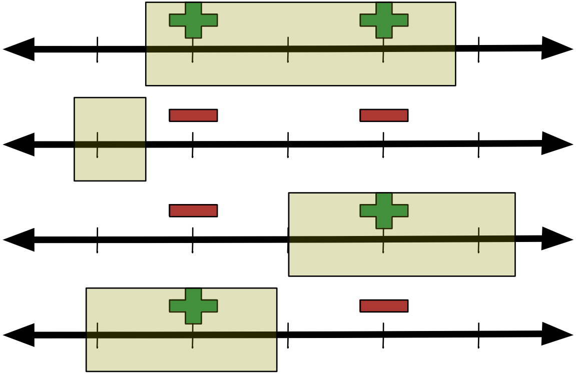
- Is the VC dimension of $h(x)$ at least 2? Check if all possible class combinations of 2 data points can be shattered



Example of VC Dimension and Shattering: Intervals

What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

- Is the VC dimension of $h(x)$ at least 2? Check if all possible class combinations of 2 data points can be shattered



Example of VC Dimension and Shattering: Intervals

What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

- Two points can be perfectly classified, so VC dimension ≥ 2

Example of VC Dimension and Shattering: Intervals

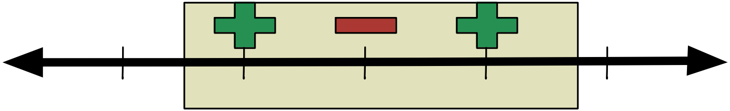
What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

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

Example of VC Dimension and Shattering: Intervals

What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

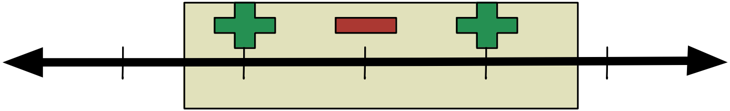
- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?



Example of VC Dimension and Shattering: Intervals

What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

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

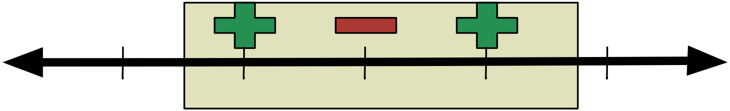


- **No set** of three points can be shattered

Example of VC Dimension and Shattering: Intervals

What is the VC dimension of $[a, b]$ intervals on the real line with $h(x)$

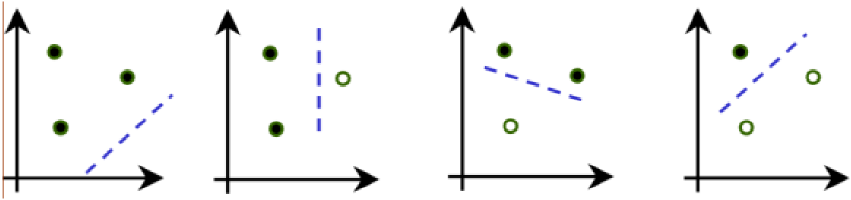
- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?



- **No set** of three points can be shattered
- Thus, VC dimension of this $h(x)$ (intervals) is 2

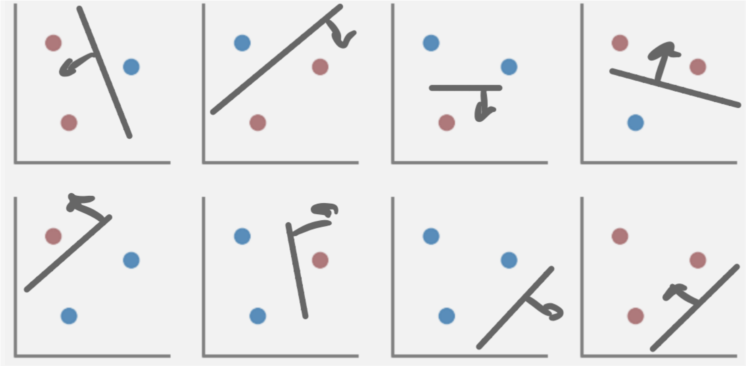
Example of VC Dimension: Hyperplanes

What is the VC dimension of a decision boundary in a 2D space i.e., $h(x) = ax_1 + bx_2 + c$ (the blue line in the plot below)?



What are other possible examples for 3 data points?

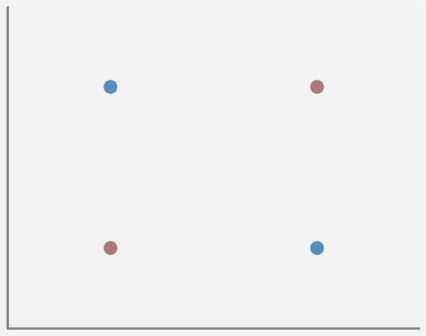
Example of VC Dimension: Hyperplanes



So we can shatter 3 data points with a decision boundary $h(x)$. Therefore, VC dimension of $h(x) \geq 3$.

Example of VC Dimension: Hyperplanes

Can we shatter 4 data points with a decision boundary $h(x)$?



Nope! Therefore, VC dimension of $h(x) = 3$

Generalization Bounds with respect to VC dimension

For a hypothesis class H with VC dimension d, for any $\delta > 0$ with probability at least $1 - \delta$, for any $h \in H$,

$$\text{Generalization Error} \leq \text{Training Error} + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{2}$$

We now have a good idea of how many training examples m do we need!
Training error is a good indicator of Generalization Error if $m >> d$

(Bonus) Sin Functions

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

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{3}$$

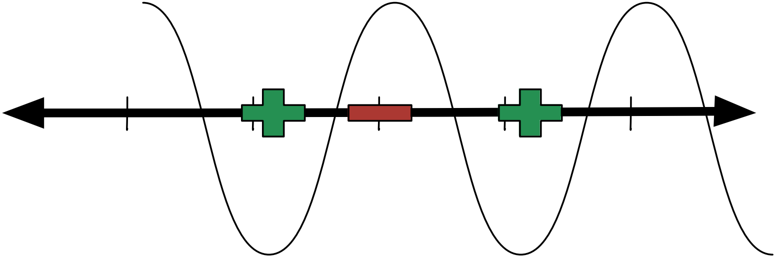
- Can you shatter three points?

(Bonus) Sin Functions

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

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{3}$$

- Can you shatter three points?



(Bonus) Sin Functions

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

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{3}$$

- How many points can you shatter?

(Bonus) Sin Functions

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

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{3}$$

- VC dim of sine on line is ∞