



More Learning Theory

Nicholas Ruozzi

University of Texas at Dallas

Based on the slides of Vibhav Gogate and David Sontag

- Probably approximately correct (PAC)
 - The only reasonable expectation of a learner is that with high probability it learns a close approximation to the target concept
 - Specify two small parameters, $0 < \epsilon, 0 < \delta < 1$
 - ϵ is the error of the approximation
 - $(1 - \delta)$ is the probability that, given M i.i.d. samples, our learning algorithm produces a classifier with error at most ϵ

- We use the observed data in order to learn a classifier
- Want to know how far the learned classifier deviates from the (unknown) underlying distribution
 - With too few samples, we will with high probability learn a classifier whose true error is quite high even though it may be a perfect classifier for the observed data
 - As we see more samples, we pick a classifier from the hypothesis space with low training error & hope that it also has low true error
 - Want this to be true with high probability – can we bound how many samples that we need?

- What we proved last time:

Theorem: For a finite hypothesis space, H , with M i.i.d. samples, and $0 < \epsilon < 1$, the probability that any consistent classifier has true error larger than ϵ is at most $|H|e^{-\epsilon M}$

- We can turn this into a sample complexity bound

- Let δ be an upper bound on the desired probability of not ϵ -exhausting the sample space
 - The probability that the version space is not ϵ -exhausted is at most $|H|e^{-\epsilon M} \leq \delta$
 - Solving for M yields

$$\begin{aligned} M &\geq -\frac{1}{\epsilon} \ln \frac{\delta}{|H|} \\ &= \left(\ln |H| + \ln \frac{1}{\delta} \right) / \epsilon \end{aligned}$$

- Our analysis for the finite case was based on $|H|$
 - If H isn't finite, this translates into infinite sample complexity
 - We can derive a different notion of complexity for infinite hypothesis spaces by considering only the number of points that can be correctly classified by some member of H
 - We will only consider the binary classification case for now

- How many points in 1-D can be correctly classified by a linear separator?
 - 2 points:



Yes!

- How many points in 1-D can be correctly classified by a linear separator?
 - 2 points:



Yes!

- How many points in 1-D can be correctly classified by a linear separator?
 - 2 points:



- How many points in 1-D can be correctly classified by a linear separator?
 - 3 points:



Yes!

- How many points in 1-D can be correctly classified by a linear separator?
 - 3 points:



NO!

- How many points in 1-D can be correctly classified by a linear separator?
 - 3 points:



- 3 points and up: for any collection of three or more there is always some choice of pluses and minuses such that that the points cannot be classified with a linear separator (in one dimension)

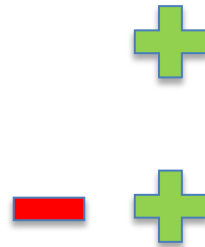
- A set of points is **shattered** by a hypothesis space H if and only if for every partition of the set of points into positive and negative examples, there exists some consistent $h \in H$
- The **Vapnik–Chervonenkis (VC) dimension** of H over inputs from X is the size of the **largest** finite subset of X shattered by H

- Common misconception:
 - VC dimension is determined by the largest shattered set of points, not the highest number such that all sets of points that size can be shattered



Cannot be shattered by a line

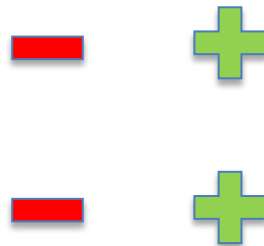
- Common misconception:
 - VC dimension is determined by the largest shattered set of points, not the highest number such that all sets of points that size can be shattered



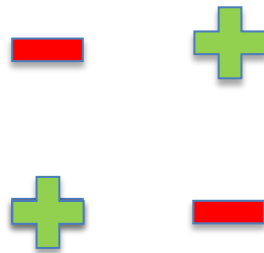
Can be shattered by a line (no matter the labels), so VC dimension is at least 3

- What is the VC dimension of 2-D space under linear separators?
 - It is at least three from the last slide
 - Can some set of four points be shattered?

- What is the VC dimension of 2-D space under linear separators?
 - It is at least three from the last slide
 - Can some set of four points be shattered?



- What is the VC dimension of 2-D space under linear separators?
 - It is at least three from the last slide
 - Can some set of four points be shattered?



- What is the VC dimension of 2-D space under linear separators?
 - It is at least three from the last slide
 - Can some set of four points be shattered?



NO! This means that the VC dimension is at most 3

- There exists a set of size $d + 1$ in a $d - dimensional$ space that can be shattered by a linear separator, but not a set of size $d + 2$
- The larger the subset of X that can be shattered, the more expressive the hypothesis space is
- If arbitrarily large finite subsets of X can be shattered, then $VC(H) = \infty$

Axis Parallel Rectangles

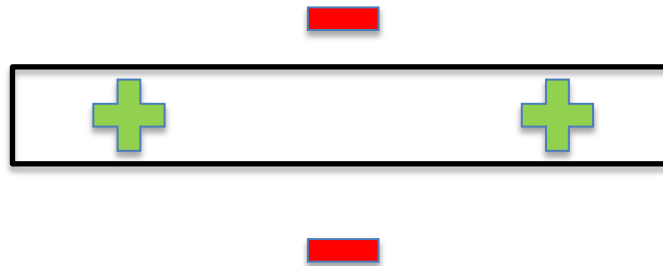


- Let X be the set of all points in \mathbb{R}^2
- Let H be the set of all axis parallel rectangles in 2-D (inside + outside -)
 - What is $VC(H)$?

Axis Parallel Rectangles



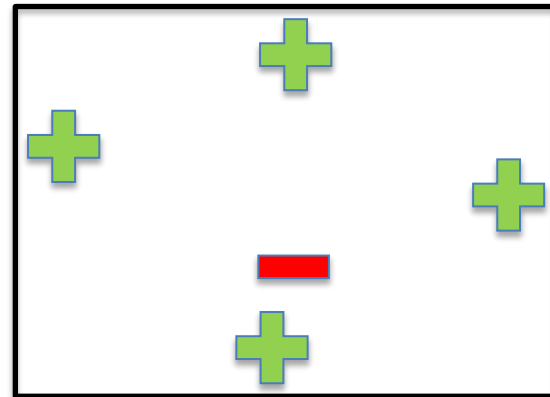
- Let X be the set of all points in \mathbb{R}^2
- Let H be the set of all axis parallel rectangles in 2-D (inside + outside -)
 - $VC(H) \geq 4$



Axis Parallel Rectangles



- Let X be the set of all points in \mathbb{R}^2
- Let H be the set of all axis parallel rectangles in 2-D
 - $VC(H) = 4$
 - A rectangle can contain at most 4 extreme points, the fifth point must be contained within the rectangle defined by these points



- VC dimension of one-level decision trees over real vectors of length 2?
- VC dimension of linear separators through the origin?
- VC dimension of a hypothesis space with exactly one hypothesis in it for binary vectors of length $n \geq 1$?

- VC dimension of one-level decision trees over real vectors of length 2?
 - Three
- VC dimension of linear separators through the origin?
- VC dimension of a hypothesis space with exactly one hypothesis in it for binary vectors of length $n \geq 1$?

- VC dimension of one-level decision trees over real vectors of length 2?
 - Three
- VC dimension of linear separators through the origin?
 - Two
- VC dimension of a hypothesis space with exactly one hypothesis in it for binary vectors of length $n \geq 1$?

- VC dimension of one-level decision trees over real vectors of length 2?
 - Three
- VC dimension of linear separators through the origin?
 - Two
- VC dimension of a hypothesis space with exactly one hypothesis in it for binary vectors of length $n \geq 1$?
 - Zero

- Given:
 - Set of data X
 - Hypothesis space H
 - Set of target concepts C
 - Training instances from unknown probability distribution over X of the form $(x, c(x))$
- Goal:
 - Learn the target concept $c \in C$

- Given:
 - A concept class C over n instances from the set X
 - A learner L with hypothesis space H
 - Two constants, $\epsilon, \delta \in (0, \frac{1}{2})$
- C is said to be PAC learnable by L using H iff for all distributions over X , learner L by sampling n instances will with probability at least $1 - \delta$ output a hypothesis $h \in H$ such that
 - $\epsilon_h \leq \epsilon$
 - Running time is polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}, n, size(c)$

Theorem: For a finite hypothesis space H , M i.i.d. samples, and $0 < \epsilon < 1$, the probability that true error of any of the best classifiers (i.e., lowest training error) is larger than its training error plus ϵ is at most $|H|e^{-2M\epsilon^2}$

- Sample complexity (for desired $\delta \geq |H|e^{-2M\epsilon^2}$)

$$M \geq \left(\ln|H| + \ln \frac{1}{\delta} \right) / 2\epsilon^2$$

- If we require that the previous error is bounded above by δ , then with probability $(1 - \delta)$, for all $h \in H$

$$\epsilon_h \leq \underbrace{\epsilon_h^{train}}_{\text{"bias"}} + \underbrace{\sqrt{\frac{1}{2M} \left(\ln |H| + \ln \frac{1}{\delta} \right)}}_{\text{"variance"}}$$

- For small $|H|$
 - High bias (may not be enough hypotheses to choose from)
 - Low variance

- If we require that the previous error is bounded above by δ , then with probability $(1 - \delta)$, for all $h \in H$

$$\epsilon_h \leq \underbrace{\epsilon_h^{train}}_{\text{"bias"}} + \underbrace{\sqrt{\frac{1}{2M} \left(\ln |H| + \ln \frac{1}{\delta} \right)}}_{\text{"variance"}}$$

- For large $|H|$
 - Low bias (lots of good hypotheses)
 - High variance

- VC dimension can be used to construct PAC bounds

$$M \geq \frac{1}{\epsilon} \left(4 \ln \frac{2}{\delta} + 8 \cdot VC(H) \ln \frac{13}{\epsilon} \right)$$

- Then, with probability at least $(1 - \delta)$ every $h \in H$ satisfies

$$\epsilon_h \leq \epsilon_h^{train} + \sqrt{\frac{1}{M} \left(VC(H) \left(\ln \left(\frac{2M}{VC(H)} \right) + 1 \right) + \ln \frac{4}{\delta} \right)}$$

- These bounds (and the preceding discussion) only work for binary classification, but there are generalizations