MORE
Learning Theory

Nicholas Ruozzi
University of Texas at Dallas

Based on the slides of Vibhav Gogate and David Sontag
Last Time

• 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 < \( \epsilon, \delta < 1 \)
    
    • \( \epsilon \) 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 \( \epsilon \)
Learning Theory

- 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 $\epsilon$ is at most $|H|e^{-\epsilon m}$

We can turn this into a sample complexity bound
Sample Complexity

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

$$m \geq -\frac{1}{\epsilon} \log \frac{\delta}{|H|}$$

$$= \left( \log |H| + \log \frac{1}{\delta} \right)/\epsilon$$
Generalizations

- How do we handle the case that there is no consistent classifier?
  - Pick the hypothesis with the lowest error on the training set, bound?

- What do we do if the hypothesis space isn’t finite?
  - Infinite sample complexity?
  - Need a way to measure the complexity of the space that isn’t based on its size
Chernoff Bounds

- **Chernoff bound**: Suppose $Y_1, \ldots, Y_m$ are i.i.d. random variables taking values in $\{0, 1\}$ such that $E_p[Y_i] = y$. For $\epsilon > 0$,

$$p \left( y - \frac{1}{m} \sum_{i} Y_i \geq \epsilon \right) \leq e^{-2m\epsilon^2}$$
Chernoff Bounds

• For $h \in H$, let $Z_i^h$ be an indicator random variable that is one if $h$ misclassifies the $i^{th}$ data point

\[
p(Z_i^h = 1) = \sum_{x,y} p(x, y)1_{h(x) \neq y} = \epsilon_h
\]

• Applying Chernoff bound to $Z_1^h, \ldots, Z_m^h$ gives

\[
p \left( \epsilon_h - \frac{1}{m} \sum_i Z_i^h \geq \epsilon \right) \leq e^{-2m\epsilon^2}
\]
Chernoff Bounds

• For $h \in H$, let $Z_i^h$ be an indicator random variable that is one if $h$ misclassifies the $i^{th}$ data point

$$p(Z_i^h = 1) = \sum_{x,y} p(x, y)1_{h(x) \neq y} = \epsilon_h$$

• Applying Chernoff bound to $Z_1^h, \ldots, Z_m^h$ gives

$$p\left(\epsilon_h - \frac{1}{m} \sum_i Z_i^h \geq \epsilon\right) \leq e^{-2m\epsilon^2}$$

This is the training error
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 $\epsilon$ is at most $|H|e^{-2m\epsilon^2}$

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

$$m \geq \left( \log |H| + \log \frac{1}{\delta} \right) / 2\epsilon^2$$
PAC Bounds

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

$$\epsilon_h \leq \epsilon_h^{train} + \sqrt{\frac{1}{2m} \left( \log |H| + \log \frac{1}{\delta} \right)}$$

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

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

\[
\epsilon_h \leq \epsilon_h^{train} + \sqrt{\frac{1}{2m} \left( \log |H| + \log \frac{1}{\delta} \right)}
\]

“bias” “variance”

– For large \( |H| \)
  • Low bias (lots of good hypotheses)
  • High variance
• Our analysis for the finite case was based on $|H|$.

  – 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 $H$. 

VC Dimension
VC Dimension

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

  – 2 points:

  ![Image with two points, one positive and one negative, indicating correct classification with "Yes!" text.]

Yes!
VC Dimension

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

  – 2 points:

    +   +   Yes!
VC Dimension

• How many points in 1-D can be correctly classified by a linear separator?
  – 2 points:
    
    Yes!
VC Dimension

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

  – 3 points:

    - + +

    Yes!
VC Dimension

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

– 3 points:

NO!
VC Dimension

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

  ![Image of three points with a linear separator]

  NO!

  – 3 points and up: for any collection of three or more there is always some choice of pluses and minuses such that the points cannot be classified with a linear separator
VC 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 $H$. 
VC Dimension

- 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
VC Dimension

- 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
VC Dimension

- 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?
VC Dimension

• 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?

![Diagram showing linear separators between points.](attachment:vc_dimension_diagram.png)
VC Dimension

• 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
VC Dimension

- There exists a linear separator that can shatter any set of size $d + 1$ in a $d - dimensional$ space, but not $d + 2$
- 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

  – 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

$VC(H) \geq 4$
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
PAC Bounds with VC Dimension

- VC dimension can be used to construct PAC bounds

\[
m \geq \frac{1}{\epsilon} \left( 4 \log \frac{2}{\delta} + 8 \cdot VC(H) \log \frac{13}{\epsilon} \right)
\]

- With probability at least \((1 - \delta)\) every \(h \in H\) satisfies

\[
\epsilon_h \leq \epsilon_h^{\text{train}} + \sqrt{\frac{1}{m} \left( VC(H) \left( \ln \left( \frac{2m}{VC(H)} \right) + 1 \right) + \ln \frac{4}{\delta} \right)}
\]

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

• 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 $c \in C$
PAC Learning

• 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 $\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c)$