Learning Theory

CISC 5800
Professor Daniel Leeds

The classifier

Function C that provides
correct label (Y) based on features (X)

C(x)=y

lion: 16
wolf: 12
monkey: 14
broker: 0
analyst: 1
dividend: 1

$\xrightarrow{C}$ jungle

lion: 0
wolf: 2
monkey: 1
broker: 14
analyst: 10
dividend: 12

$\xrightarrow{C}$ wallStreet

Goal: identifier classifier that maximizes correct labels for most inputs

2

Sample complexity

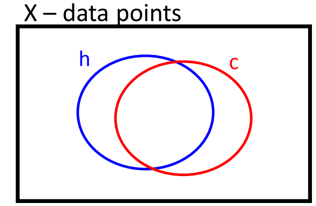
How many training examples needed to learn concept?

- *X* – set of data points
- *P(X)* – Probability of drawing data point x
- *H* – space of hypotheses H = {h : X -> classes }
- *C* – *correct* assignment C = {c : c(x) = y ∀x∈X }

3

Probability of error

X – data points



H = {h : X -> {0,1}}

True error of h: probability randomly selected data point from P(X) misclassified

$$\text{error}_{\text{true}}(h) = \Pr_{x\sim P(X)}[h(x) \neq c(x)]$$

- Hard to compute, but can prove properties of $\text{error}_{\text{true}}$

4

Example: Learner picks one of fixed number of classifiers $h \epsilon H$

Correct classifier c is some assignment of each x to a label

How many training points m needed for $\text{error}_{\text{true}}(h) < \varepsilon$?

Prob[$\text{error}_{\text{true}}(h) \leq \varepsilon$] > 1-$\delta$

"Probability learned classifier h has worse than ε error is < δ"

"Probably Approximately Correct Learning" – PAC Learning

5

Binary example: sample complexity

Note for $\varepsilon = [0,1]$, $(1-\varepsilon) \leq e^{-\varepsilon}$

What is the chance learned h is bad but classifies training data correctly?

If $\text{error}_{\text{true}}(h) > \varepsilon$:

- Prob [h correctly labels x^1] < $(1-\varepsilon) \leq e^{-\varepsilon}$
- Prob [h correctly labels x^1 and x^2 … and x^m] < $(1-\varepsilon)^m \leq e^{-m\varepsilon}$

If classifier picks one h* randomly from H

- Prob[h^* is bad] = Prob[h_1 bad] + … Prob[h_n bad]
  = Prob[$\text{error}_{\text{true}}(h^*) > \varepsilon$] < $|\mathrm{H}|\ e^{-m\varepsilon}$

Valiant, 1984

6

Binary example: sample complexity

Number of data points to reduce chance of false classification, enforce

Prob[$\text{error}_{\text{true}}(h) \leq \varepsilon$] > 1-$\delta$

1- Prob[$\text{error}_{\text{true}}(h) \leq \varepsilon$]= Prob[$\text{error}_{\text{true}}(h) > \varepsilon$]< δ

Prob[$\text{error}_{\text{true}}(h^*) > \varepsilon$] < $|\mathrm{H}|\ e^{-m\varepsilon}$; stricter bound $|\mathrm{H}|\ e^{-m\varepsilon} < \delta$

Valiant, 1984

8

Binary example: sample complexity

Number of data points to reduce chance of false classification, enforce

Prob[$\text{error}_{\text{true}}(h) \leq \varepsilon$] > 1-$\delta$

Prob[$\text{error}_{\text{true}}(h^*) > \varepsilon$] < $|\mathrm{H}|\ e^{-m\varepsilon} < \delta$

$$m > \frac{1}{\varepsilon} \ln \frac{|H|}{\delta}$$

Valiant, 1984

9

VC Dimensions

If H not finite, PAC result seems to require ∞ data points

- Overly conservative

"Dichotomy" – division of set of points S into two subsets

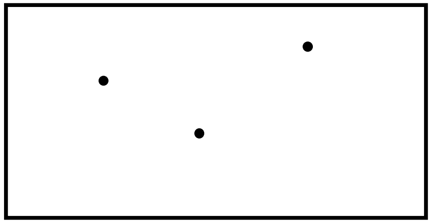
- "Shattering" – set of points is **shattered** by H iff there exists h∈H associated with every possible dichotomy

Vapnik-Chervonenkis dimension **VC(H)** is size of largest finite subset of X that can be shattered by H

10

Shattering example

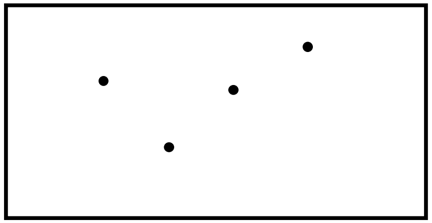
- H={rectangles} VC(3)
- S={3 specified dots}



11

Shattering example

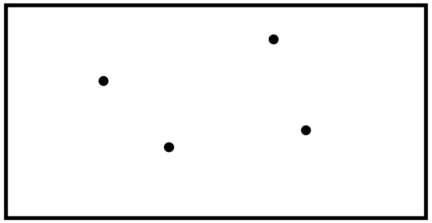
- H={rectangles} VC(3)
- S={4 specified dots}



12

Shattering example

- H={rectangles} VC(4)
- S={4 specified dots}

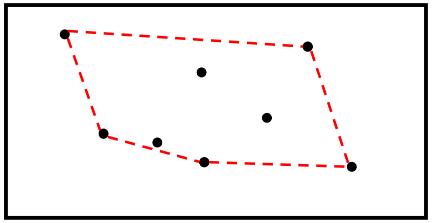


13

Shattering example

- H={ovals}
- S={8 specified dots}

VC(5) – convex hull



PAC result with infinite H

VC(H) is size of largest finite subset of X that can be shattered by H

- d=$VC(H)$
- $m \geq O\left(\frac{1}{\varepsilon}\left[d \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]\right) \sim \frac{1}{\varepsilon}\left[d \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]$

Recall: $m > \frac{1}{\varepsilon}\ln\frac{|H|}{\delta}$ for finite size H