Computational Learning Theory Part 2

VC dimension, Sample Complexity, Mistake bounds

Required reading:

- Mitchell chapter 7

Optional advanced reading:

- Kearns & Vazirani, 'Introduction to Computational Learning Theory'

Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

October 17, 2006

Last time: PAC Learning

1. Finite H, assume target function c $\in$ H

$$\underline{\text{Pr}[(\exists h \in H) s.t. (error_{train}(h) = 0) \wedge (error_{true}(h) > \epsilon)]} \;\leq\; |H| e^{-\epsilon m}$$

Suppose we want this to be at most δ. Then *m* examples suffice:

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

2. Finite H, agnostic learning: perhaps c *not* in H

with probability at least (1-δ) every h in H satisfies

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

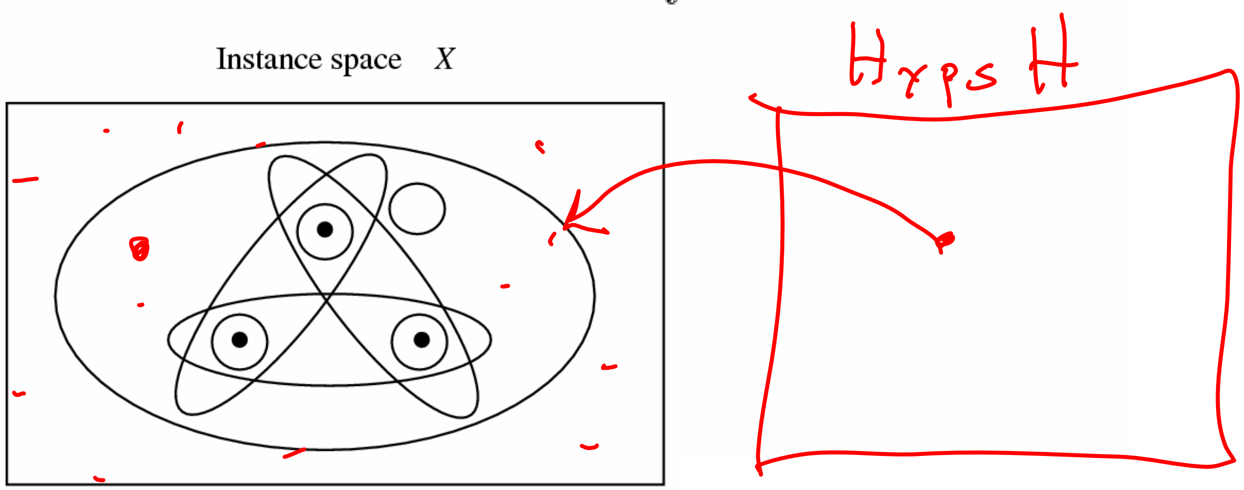
What if H is not finite?

- Can’t use our result for finite H
- Need some other measure of complexity for H
  - Vapnik-Chervonenkis (VC) dimension!

Shattering a Set of Instances

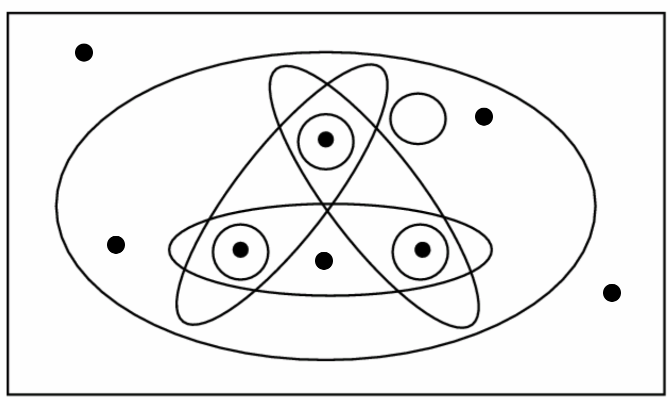
Definition: a **dichotomy** of a set S is a partition of S into two disjoint subsets.

Definition: a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.



The Vapnik-Chervonenkis Dimension

Definition: The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.



VC(H)=3

Sample Complexity based on VC dimension

How many randomly drawn examples suffice to ε-exhaust $VS_{H,D}$ with probability at least (1-δ)?

ie., to guarantee that any hypothesis that perfectly fits the training data is probably (1-δ) approximately (ε) correct

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

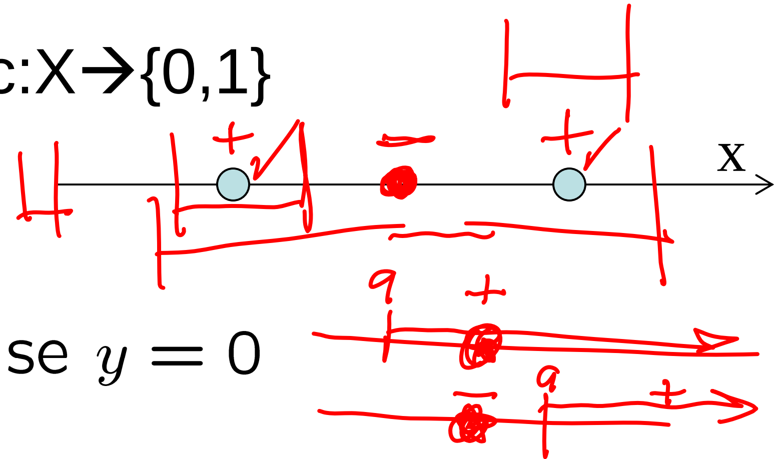
assumes zero training error

Compare to our earlier results based on |*H*|:

$$m \geq \frac{1}{\epsilon}(\ln(1/\delta) + \ln|H|)$$

VC dimension: examples

Consider X = $\Re$, want to learn c:X→{0,1}



What is VC dimension of

- Open intervals:

1 = VC(H1): if $x > a$ then $y = 1$ else $y = 0$

2 = VC(H2): if $x > a$ then $y = 1$ else $y = 0$
or, if $x > a$ then $y = 0$ else $y = 1$

- Closed intervals:

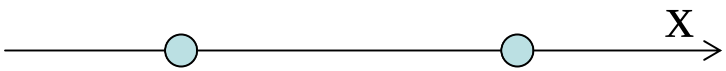
2 = H3: if $a < x < b$ then $y = 1$ else $y = 0$

H4: if $a < x < b$ then $y = 1$ else $y = 0$
or, if $a < x < b$ then $y = 0$ else $y = 1$

VC dimension: examples

Consider X = $\Re$, want to learn c:X→{0,1}

What is VC dimension of



- Open intervals:

H1: if $x > a$ then $y = 1$ else $y = 0$ VC(H1)=1

H2: if $x > a$ then $y = 1$ else $y = 0$ VC(H2)=2
or, if $x > a$ then $y = 0$ else $y = 1$

- Closed intervals:

H3: if $a < x < b$ then $y = 1$ else $y = 0$ VC(H3)=2

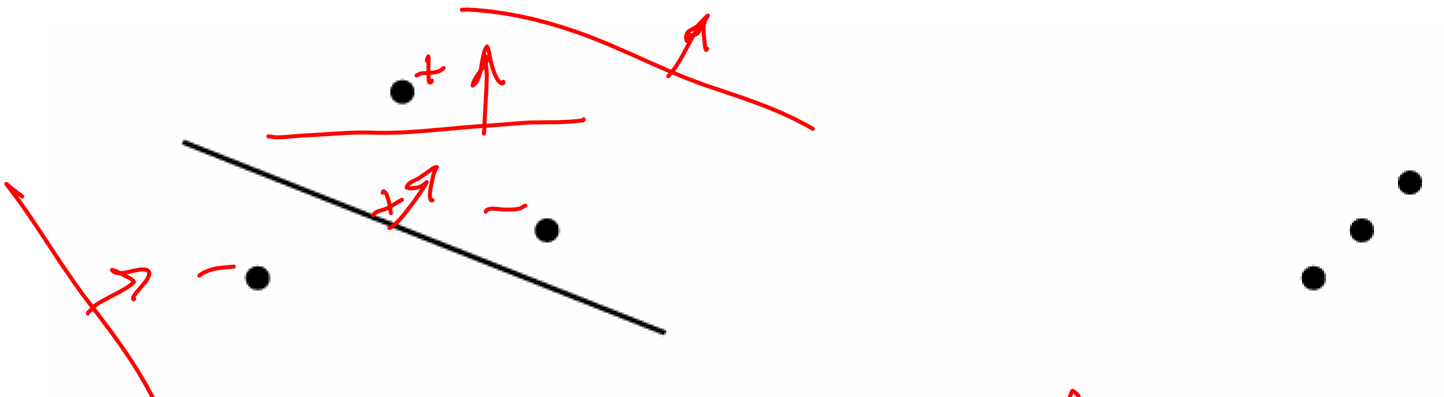
H4: if $a < x < b$ then $y = 1$ else $y = 0$ VC(H4)=3
or, if $a < x < b$ then $y = 0$ else $y = 1$

VC dimension: examples

Consider X = $\Re^2$, want to learn c:X→{0,1}

What is VC dimension of lines in a plane?

- H = { ((wx+b)>0 → y=1) }



H = linear sep. in n dimensions

→ VC(H) = n+1

VC dimension: examples

Consider X = $\Re^2$, want to learn c:X→{0,1}

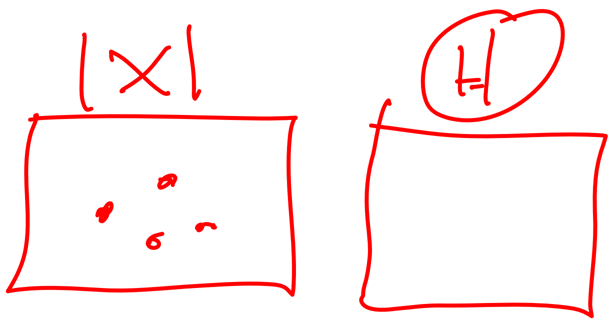
What is VC dimension of

- H = { ((w·x+b)>0 → y=1) }
  - VC(H1)=3
  - For linear separating hyperplanes in n dimensions, VC(H)=n+1

$VC(H) \leq \log_2 |H|$

For any finite hypothesis space H,
give an upper bound on VC(H) in terms of $\boxed{|H|}$

to shatter X requires $2^{|X|}$ hyps.



shattering $\underline{n}$ pts req. 2^n hyps.

More VC Dimension Examples

- Decision trees defined over n boolean features

  F: $<X_1, \ldots X_n> \rightarrow Y$

- Decision trees defined over n continuous features

  Where each internal tree node involves a threshold test $(X_i > c)$

- Decision trees of depth 2 defined over n features

- Logistic regression over n continuous features? Over n boolean features?

- How about 1-nearest neighbor?

Tightness of Bounds on Sample Complexity

How many examples *m* suffice to assure that any hypothesis that fits the training data perfectly is probably (1-δ) approximately (ε) correct?

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

How tight is this bound?

Lower bound on sample complexity (Ehrenfeucht et al., 1989): P(x)

Consider any class C of concepts such that VC(C) ≥ 2, any learner L, any 0 < ε < 1/8, and any 0 < δ < 0.01. Then there exists a distribution $\mathcal{D}$ and target concept in C, such that if L observes fewer examples than

$$\max\left[\frac{1}{\epsilon}\log(1/\delta), \frac{VC(C)-1}{32\epsilon}\right]$$

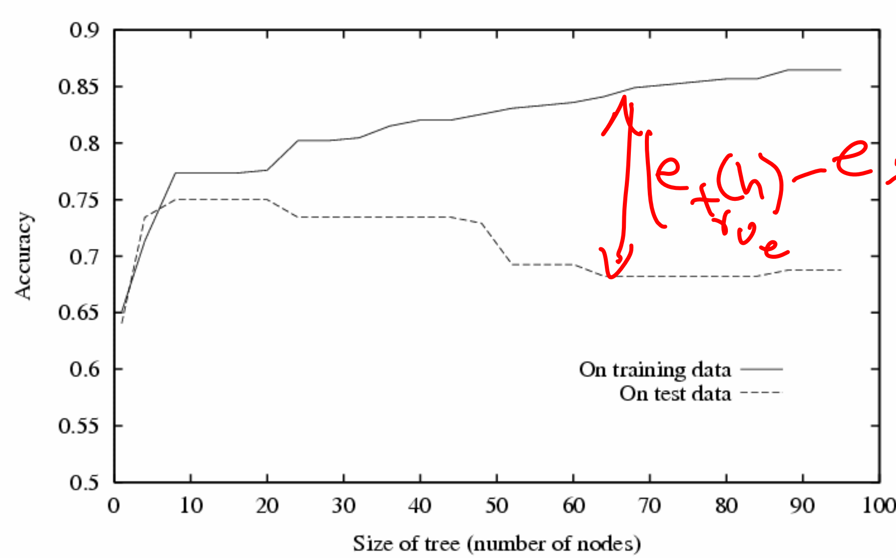
Then with probability at least δ, L outputs a hypothesis with $error_{\mathcal{D}}(h) > \epsilon$

Agnostic Learning: VC Bounds

[Schölkopf and Smola, 2002]

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

$$error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

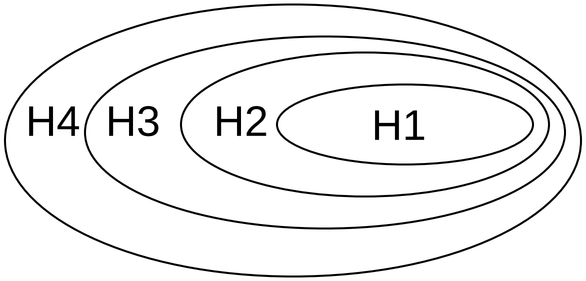


$|e_{true}(h) - e_{train}(h)|$

Structural Risk Minimization [Vapnik]

Which hypothesis space should we choose?

- Bias / variance tradeoff



SRM: choose H to minimize bound on true error!

$$error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}$$

* unfortunately a somewhat loose bound...

What You Should Know

- Sample complexity varies with the learning setting
  - Learner actively queries trainer
  - Examples provided at random
- Within the PAC learning setting, we can bound the probability that learner will output hypothesis with given error
  - For ANY consistent learner (case where c 2 H)
  - For ANY "best fit" hypothesis (agnostic learning, where perhaps c not in H)
- VC dimension as measure of complexity of H
- Quantitative bounds characterizing bias/variance in choice of H
  - but the bounds are quite loose...
- Mistake bounds in learning
- Conference on Learning Theory: http://www.learningtheory.org