CSCE 478/878 Lecture 3: Computational Learning Theory

Stephen D. Scott
(Adapted from Tom Mitchell's slides)

September 8, 2003

Introduction

- Combines machine learning with:
  - Algorithm design and analysis
  - Computational complexity
- Examines the worst-case minimum and maximum data and time requirements for learning
  - Number of examples needed, number of mistakes made before convergence
- Tries to relate:
  - Probability of successful learning
  - Number of training examples
  - Complexity of hypothesis space
  - Accuracy to which target concept is approximated
  - Manner in which training examples presented
- Some average case analyses done as well

Outline

- Probably approximately correct (PAC) learning
- Sample complexity
- Agnostic learning
- Vapnik-Chervonenkis (VC) dimension
- Mistake bound model
- Note: as with previous lecture, we assume no noise, though most of the results can be made to hold in a noisy setting

PAC Learning: The Problem Setting

Given:

- set of instances X
- set of hypotheses H
- set of possible target concepts C (typically, $C \subseteq H$)
- training instances independently generated by a fixed, unknown, arbitrary probability distribution $\mathcal{D}$ over X

Learner observes a sequence D of training examples of form $\langle x, c(x) \rangle$, for some target concept $c \in C$

- instances x are drawn from distribution $\mathcal{D}$
- teacher provides target value $c(x)$ for each

PAC Learning: The Problem Setting

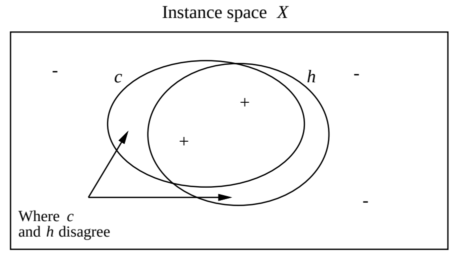
(cont'd)

Learner must output a hypothesis $h \in H$ approximating $c \in C$

- h is evaluated by its performance on subsequent instances drawn according to $\mathcal{D}$

Note: probabilistic instances, noise-free classifications

True Error of a Hypothesis



$c \triangle h =$ symmetric difference between c and h

Definition: The true error (denoted $error_{\mathcal{D}}(h)$) of hypothesis h with respect to target concept c and distribution $\mathcal{D}$ is the probability that h will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

(example $x \in X$ drawn randomly according to $\mathcal{D}$)

Two Notions of Error

Training error of hypothesis h with respect to target concept c

- How often $h(x) \neq c(x)$ over training instances

True error of hypothesis h with respect to c

- How often $h(x) \neq c(x)$ over future random instances

Our concern:

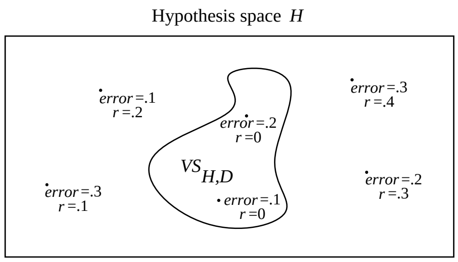
- Can we bound the true error of h given the training error of h?
- First consider when training error of h is zero (i.e., $h \in VS_{H,D}$)

PAC Learning

Consider a class C of possible target concepts defined over a set of instances X of size n, and a learner L using hypothesis space H.

Definition: C is PAC-learnable by L using H if for all $c \in C$, distributions $\mathcal{D}$ over X, ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$, learner L will, with probability at least $(1 - \delta)$, output a hypothesis $h \in H$ such that $error_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $size(c)$.

Exhausting the Version Space



($r =$ training error, $error =$ true error)

Definition: The version space $VS_{H,D}$ is said to be ϵ-exhausted with respect to c and $\mathcal{D}$, if every hypothesis $h \in VS_{H,D}$ has error less than ϵ with respect to c and $\mathcal{D}$.

$$(\forall h \in VS_{H,D}) \ error_{\mathcal{D}}(h) < \epsilon$$

How many examples m will ϵ-exhaust the VS?

- Let $h_1, \ldots, h_k \in H$ be all hyps. with true error $> \epsilon$ w.r.t. c and $\mathcal{D}$ (i.e. the ϵ-bad hyps.)
- VS is not ϵ-exhausted iff at least one of these hyps. is consistent with all m examples
- Prob. that an ϵ-bad hyp consistent with one random example is $\leq (1 - \epsilon)$
- Since random draws are independent, the prob. that a particular ϵ-bad hyp is consistent with m exs. is $\leq (1 - \epsilon)^m$
- So the prob. any ϵ-bad hyp is in VS is
$$\leq k(1-\epsilon)^m \leq |H|(1-\epsilon)^m$$
- Given $(1 - \epsilon) \leq 1/e^{\epsilon}$ for $\epsilon \in [0, 1]$:
$$|H|(1-\epsilon)^m \leq |H|e^{-m\epsilon}$$

How many examples m will ϵ-exhaust the VS? (cont'd)

Theorem: [Haussler, 1988]

If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independent random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that the version space with respect to H and D is not ϵ-exhausted (with respect to c) is
$$\leq |H|e^{-m\epsilon}$$

This bounds the probability that any consistent learner will output a hypothesis h with $error(h) \geq \epsilon$

If we want this probability to be $\leq \delta$ (for PAC):

$$|H|e^{-m\epsilon} \leq \delta$$

then

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

suffices

Learning Conjunctions of Boolean Literals

How many examples are sufficient to assure with probability at least $(1 - \delta)$ that

every h in $VS_{H,D}$ satisfies $error_{\mathcal{D}}(h) \leq \epsilon$

Use the theorem:

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

Suppose H contains conjunctions of constraints on up to n boolean attributes (i.e., n boolean literals). Then $|H| = 3^n$ (why?), and

$$m \geq \frac{1}{\epsilon}(\ln 3^n + \ln(1/\delta)),$$

or

$$m \geq \frac{1}{\epsilon}(n \ln 3 + \ln(1/\delta))$$

Still need to find a hyp. from VS!

How About $EnjoySport$?

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

If H is as given in $EnjoySport$, then $|H| = 973$ and

$$m \geq \frac{1}{\epsilon}(\ln 973 + \ln(1/\delta))$$

... if want to assure that with probability 95%, VS contains only hypotheses with $error_{\mathcal{D}}(h) \leq .1$, then it is sufficient to have m examples, where

$$m \geq \frac{1}{.1}(\ln 973 + \ln(1/.05))$$

$$m \geq 10(\ln 973 + \ln 20)$$

$$m \geq 10(6.88 + 3.00)$$

$$m \geq 98.8$$

Again, how to find a consistent hypothesis?

Unbiased Learners

- Recall the unbiased concept class $C = 2^X$, i.e. set of all subsets of X
- If each instance $x \in X$ is described by n boolean features, $|X| = 2^n$, so $|C| = 2^{2^n}$
- Also, to ensure $c \in H$, need $H = C$, so the theorem gives
  $$m \geq \frac{1}{\epsilon}\left(2^n \ln 2 + \ln(1/\delta)\right),$$
  i.e. exponentially large sample complexity
- Note the above is only sufficient, the theorem does not give necessary sample complexity
- (Necessary sample complexity is still exponential)

$\Rightarrow$ Further evidence for the need of bias (as if we need more)

Agnostic Learning

So far, assumed $c \in H$

Agnostic learning setting: don't assume $c \in H$

- What do we want then?
  - The hypothesis h that makes fewest errors on training data (i.e. the one that minimizes disagreements, which can be harder than finding consistent hyp)
- What is sample complexity in this case?

$$m \geq \frac{1}{2\epsilon^2}(\ln |H| + \ln(1/\delta)),$$

derived from Hoeffding bounds, bounding prob. of large deviation from expected value:

$$Pr[error_{\mathcal{D}}(h) > error_D(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

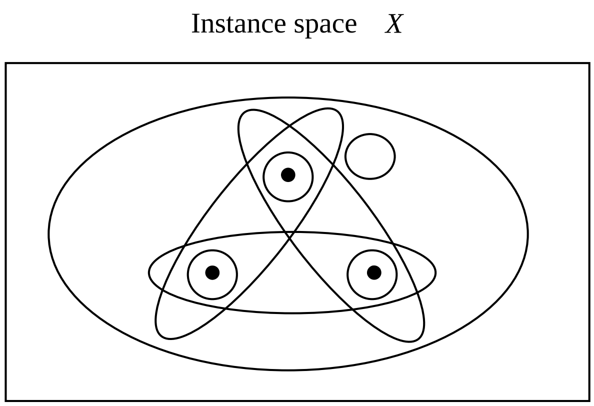
Vapnik-Chervonenkis Dimension

Shattering a Set of Instances

Definition: a dichotomy of a set S is a partition of S into two disjoint subsets, i.e. into a set of + exs. and a set of $-$ exs.

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.

Example: Three Instances Shattered



The Vapnik-Chervonenkis Dimension

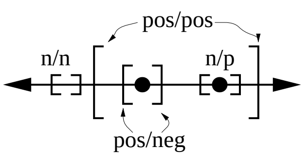
Definition: The Vapnik-Chervonenkis dimension, $VC(H)$, of hypothesis space H defined over instance space X, is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

- So to show that $VC(H) = d$, must show there exists some subset $X' \subset X$ of size d that H can shatter and show that there exists no subset of X of size $> d$ that H can shatter

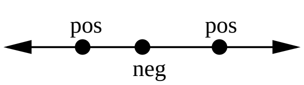
- Note that $VC(H) \leq \log_2 |H|$ (why?)

Example: Intervals on $\Re$

- Let H be the set of closed intervals on the real line (each hyp is a single interval), $X = \Re$, and a point $x \in X$ is positive iff it lies in the target interval c



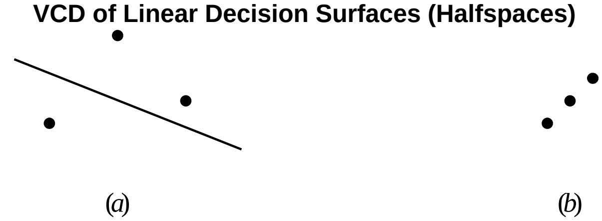
Can shatter 2 pts, so
VCD >= 2



Can't shatter any 3 pts, so
VCD < 3

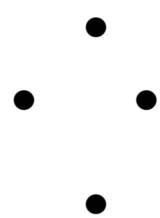
- Thus $VC(H) = 2$ (also note that $|H|$ is infinite)

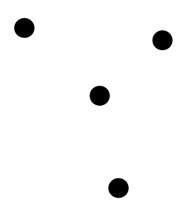
VCD of Linear Decision Surfaces (Halfspaces)



Can't shatter (b), so what is lower bound on VCD?

What about upper bound?





Sample Complexity from VC Dimension

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

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

(compare to finite H case)

- In the worst case, how many are required?

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

i.e. $\exists\,\mathcal{D}$ such that if learner sees fewer than this many examples, with prob. $\geq \delta$, its hyp will have error $> \epsilon$

- Can also get results in the agnostic model and with noisy data (using e.g. statistical queries)

Mistake Bound (On-Line) Learning

- So far only considered how many examples required to learn with high probability
- On-line model: how many mistakes will learner make before convergence (i.e. exactly learning c)?
- Setting:
  - Learning proceeds in trials
  - At each trial t, learner gets example $x_t \in X$ and must predict x_t's label
  - Then teacher informs learner of true value of $c(x_t)$ and learner updates hypothesis if necessary
- Goal: Minimize total number of prediction mistakes (requires exact learning of c)

On-Line vs. PAC Model

- On-line is adversarial (worst-case) model vs. probabilistic of PAC, so assume that adversary presents examples in a way to make learner perform as poorly as possible
- On-line learner that makes $\leq M$ mistakes can PAC learn with sample complexity

$$O\left(\frac{1}{\epsilon}\left(M + \log\frac{1}{\delta}\right)\right) \text{ if } M \text{ known}$$

$$O\left(\frac{M}{\epsilon}\left(M + \log\frac{1}{\delta}\right)\right) \text{ if } M \text{ unknown}$$

- But there exist finite concept classes C that can be efficiently PAC learned but not efficiently learned in on-line model
- So on-line model is harder to learn in!

Mistake Bounds: Find-S

Find-S when $H =$ conjuntion of boolean literals:

- Initialize h to the most specific hypothesis $\ell_1 \wedge \neg\ell_1 \wedge \ell_2 \wedge \neg\ell_2 \wedge \cdots \wedge \ell_n \wedge \neg\ell_n$
- For each positive training instance x, remove from h any literal that is not satisfied by x
- Output hypothesis h

How many mistakes before converging to c? If $c \in H$, Find-S will only misclassify pos. exs., and each mistake results in eliminating literals

- Total number of literals:
- Number of literals eliminated after 1st mistake:
- Number of literals eliminated after each subsequent mistake:
- Total number of mistakes $\leq$ mist. bnd $M =$

Mistake Bounds: Halving Algorithm

The Halving Algorithm:

- Learn concept using version space Candidate-Elimination algorithm (eliminate from VS all inconsistent hyps)
- Classify new instances by majority vote of version space members (classify as $+$ if majority vote $+$, else classify $-$)

How many mistakes before converging to $c \in H$?

- In worst case:
- In best case:

Optimal Mistake Bounds

Let $M_A(C)$ be the max number of mistakes made by algorithm A to learn concepts in C (maximum over all possible $c \in C$, and all possible training sequences)

$$M_A(C) \equiv \max_{c \in C} M_A(c)$$

Definition: Let C be an arbitrary non-empty concept class. The optimal mistake bound for C, denoted $Opt(C)$, is the minimum over all possible learning algorithms A of $M_A(C)$

$$Opt(C) \equiv \min_{A \in \text{all learning algorithms}} M_A(C)$$

I.e. $Opt(C)$ is the number of mistakes made by the best learning algorithm for the hardest target concept in C, using the hardest sequence of training examples

Can show:

$$VC(C) \leq Opt(C) \leq M_{Halving}(C) \leq \log_2(|C|)$$

Topic summary due in 1 week!