Computational Learning Theory

- Notions of interest: efficiency, accuracy, complexity
- Probably, Approximately Correct (PAC) Learning
- Agnostic learning
- VC Dimension and Shattering
- Mistake Bounds

Computational Learning Theory

What general laws constrain inductive learning?

Some potential areas of interest:

- Probability of successful learning
- Number of training examples
- Complexity of hypothesis space
- Accuracy to which target concept is approximated
- Efficiency of learning process
- Manner in which training examples are presented

The Concept Learning Task

Given

- Instance space X – (e.g., possible faces described by attributes Hair, Nose, Eyes, etc.)
- A unknown target function c – (e.g., Smiling : $X \rightarrow$ {yes, no})
- A hypothesis space H: $H = \{ h : X \rightarrow \{\text{yes, no}\} \}$
- A unknown, likely not observable probability distribution D over the instance space X

Determine

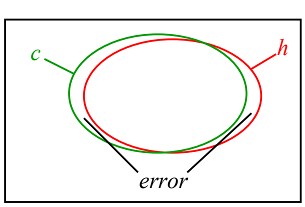
- A hypothesis h in H such that $h(x) = c(x)$ for all x in D?
- A hypothesis h in H such that $h(x) = c(x)$ for all x in X?

Variations on the Task – Data Sample

How many training examples sufficient to learn target concept?

1. Random process (e.g., nature) produces instances
   - Instances x generated randomly, teacher provides $c(x)$
2. Teacher (knows c) provides training examples
   - Teacher provides sequences of form $<x,c(x)>$
3. Learner proposes instances, as queries to teacher
   - Learner proposes instance x, teacher provides $c(x)$

True Error of a Hypothesis



Instance Space X

- *True error* of a hypothesis h with respect to target concept c and distribution D is the probability that h will misclassify an instance drawn at random according to D.

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

Notions of Error

Training error of hypothesis h with respect to target concept c

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

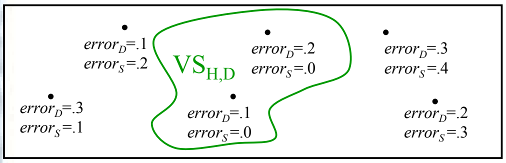
True error of hypothesis h with respect to c

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

Our concern

- Can we bound the true error of h given training error of h?
- Start by assuming training error of h is 0 (i.e., $h \in VS_{H,D}$)

Exhausting the Version Space



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

$$(\forall h \in VS_{H,D})\ error_D(h) < \varepsilon$$

How many examples to ε-exhaust VS?

Theorem:

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

Bounds the probability that any consistent learner will output a hypothesis *h* with $error(h) \geq \varepsilon$

If we want this probability to be below δ

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

Then

$$m \geq (1/\varepsilon)(\ln |H| + \ln(1/\delta))$$

Learning conjunctions of boolean literals

How many examples are sufficient to assure with probability at least (1 – δ) that

every *h* in $VS_{H,D}$ satisfies $error_D(h) \leq \varepsilon$

Use our theorem:

$$m \geq (1/\varepsilon)(\ln |H| + \ln(1/\delta))$$

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

$$m \geq (1/\varepsilon)(\ln 3^n + \ln(1/\delta))$$

or

$$m \geq (1/\varepsilon)(n \ln 3 + \ln(1/\delta))$$

For concept Smiling Face

Concept features:

- Eyes {round,square} → RndEyes, ¬RndEyes
- Nose {triangle,square} → TriNose, ¬TriNose
- Head {round,square} → RndHead, ¬RndHead
- FaceColor {yellow,green,purple} → YelFace, ¬YelFace, GrnFace, ¬GrnFace, PurFace, ¬PurFace
- Hair {yes,no} → Hair, ¬Hair

Size of $|H| = 3^7 = 2187$

If we want to assure that with probability 95%, *VS* contains only hypotheses $error_D(h) \leq .1$, then sufficient to have *m* examples, where

$m \geq (1/.1)(\ln(2187) + \ln(1/.05))$

$m \geq 10(\ln(2187) + \ln(20))$

PAC Learning

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

Definition: *C* is **PAC-learnable** by *L* using *H* if for all c ∈ *C*, distributions *D* over *X*, ε such that 0 < ε < ½, and δ such that 0 < δ < ½, learner L will with prob. at least (1- δ) output a hypothesis $h \in H$ such that $error_D(h) \leq \varepsilon$, in time that is polynomial in 1/ε, 1/δ, *n* and *size(c)*.

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
- What is sample complexity in this case?

  $m \geq (1/2\varepsilon^2)(\ln |H| + \ln(1/\delta))$

Derived from Hoeffding bounds:

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

But what if hypothesis space not finite?

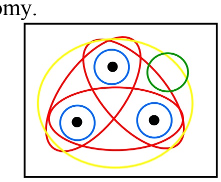
What if |H| can not be determined?

- It is still possible to come up with estimates based not on counting how many hypotheses, but based on how many instances can be completely discriminated by H
- Use the notion of a shattering of a set of instances to measure the complexity of a hypothesis space
- VC Dimension measures this notion and can be used as a stand in for |H|

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* iff for every dichotomy of *S* there exists some hypothesis in *H* consistent with this dichotomy.

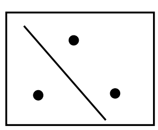
Example: 3 instances shattered



Instance space *X*

The Vapnik-Chervonenkis Dimension

- **Definition**: the **Vapnik-Chervonenkis (VC) 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 $VC(H) = \infty$.
- Example: VC dimension of linear decision surfaces is 3.



Sample Complexity with VC Dimension

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

$$m \geq \frac{1}{\varepsilon}\left(4\log_2\left(\frac{2}{\delta}\right) + 8\,VC(H)\log_2\left(\frac{13}{\varepsilon}\right)\right)$$

Mistake Bounds

So far: how many examples needed to learn?

What about: how many mistakes before convergence?

Consider setting similar to PAC learning:

- Instances drawn at random from *X* according to distribution *D*
- Learner must classify each instance before receiving correct classification from teacher

Can we bound the number of mistakes learner makes before converging?

Mistake Bounds: Find-S

Consider Find-S when *H* = conjunction of boolean literals

Find-S

- Initialize h to the most specific hypothesis:
  $l_1 \wedge \neg l_1 \wedge l_2 \wedge \neg l_2 \wedge l_3 \wedge \neg l_3 \wedge \ldots \wedge l_n \wedge \neg l_n$
- 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 correct *h*?

Mistakes in Find-S

- Assuming $c \in H$
  - Negative examples – can never be mislabeled as positive, the current hypothesis *h* is always at least as specific as target concept *c*
  - Positive examples – can be mislabeled as negative (concept not general enough, consider initial)
  - First positive example, *2n* terms in literal (positive and negative of each feature), *n* will be eliminated
  - Each subsequent mislabeled positive example – will eliminate at least one term
  - Thus at most $n+1$ mistakes

Mistake Bounds: Halving Algorithm

Consider the Halving Algorithm

- Learn concept using version space candidate elimination algorithm
- Classify new instances by majority vote of version space members

- How many mistakes before converging to correct *h*?
- … in worst case?
- … in best case?

Mistakes in Halving

- At each point, predictions are made based on a majority of the remaining hypotheses
- A mistake can be made only when at least half of the hypotheses are wrong
- Thus the size of *H* decreases by half for each mistake
- Thus, worst case bound is related to $log_2 |H|$
- How about best case?
  - Note, prediction of the majority could be correct but number of remaining hypotheses can decrease
  - Possible for the number of hypotheses to reach one with no mistakes