cture 3: Introduction to Computational Learning Theory

- When should learning work?
- Probably approximately correct (PAC) learning

Computational Learning Theory

What general laws constrain inductive learning?

We seek theory 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

Today we assume noise-free, perfect data

There is some theory for particular kinds of noisy data.

Sample Complexity

How many training examples are sufficient to learn the target concept?

1. If learner proposes instances, as queries to teacher
   - Learner proposes instance x, teacher provides $c(x)$
2. If teacher (who knows c) provides training examples
   - teacher provides sequence of examples of form $\langle x, c(x) \rangle$
3. If some random process (e.g., nature) proposes instances
   - instance x generated randomly, teacher provides $c(x)$

Sample Complexity: 1

Learner proposes instance x, teacher provides $c(x)$

(assume c is in learner's hypothesis space H)

Optimal query strategy: play 20 questions

- pick instance x such that half of hypotheses in VS classify x positive, half classify x negative
- When this is possible, need $\lceil \log_2 |H| \rceil$ queries to learn c
- when not possible, need even more

Sample Complexity: 2

Teacher (who knows c) provides training examples

(assume c is in learner's hypothesis space H)

Optimal teaching strategy: depends on H used by learner

Consider the case H = conjunctions of up to n boolean literals and their negations

e.g., $(AirTemp = Warm) \wedge (Wind = Strong)$, where $AirTemp, Wind, \ldots$ each have 2 possible values.

If n possible boolean attributes in H, $n + 1$ examples suffice

Why is that? Can you propose a learning algorithm for doing it?

On-line learning of conjunctions

Learning algorithm:

1. Start with a conjunction of all the literals
2. Whenever there is a mistake, remove all literals that are false in the example

Can you show that this will not make more than $n + 1$ mistakes?

1. Mistakes are only made on positive examples
2. The first mistake will eliminate n literals.
3. Every mistake eliminates at least one literal
4. No literal that is in the concept is ever removed

Mistake-bound model

Seeks a bound on the nuber of mistakes that the learner will make before learning the concept perfectly.

In the previous example, the bound is $n + 1$ mistakes

Does not say when the mistakes will be made (so no guarantees if we see fewer examples than needed).

Sample Complexity: 3

Given:

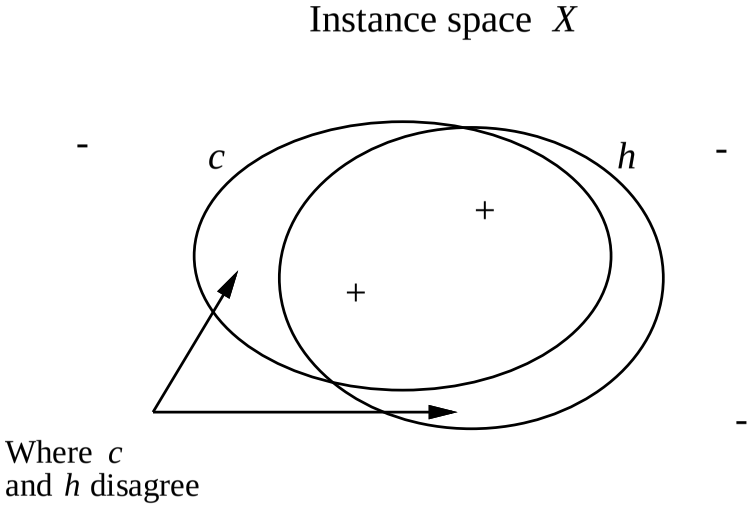
- set of instances X
- set of hypotheses H
- set of possible target concepts C
- training instances generated by a fixed, unknown probability distribution $\mathcal{D}$ over X

Learner observes a seqence D of training examples of form $\langle x, c(x)\rangle$, for some target concept $c \in C$. These instances x are drawn from distribution $\mathcal{D}$ and the teacher provides target value $c(x)$ for each.

Learner must output a hypothesis h estimating c, which will be evaluated by its performance on subsequent instances drawn according to $\mathcal{D}$

Note: probabilistic instances, noise-free classifications

True Error of a Hypothesis



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 $\mathcal{D}$.

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

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?*

Let's play a game

Given: (x,y) points in the plane, labelled $+$ or $-$

Learn: an axis-alligned rectangle that contains all the $+$ points and does not contain any $-$ points

Can you suggest a learning algorithm for this?

The rectangle problem

A simple learning algorithm will give the tightest fit to the training data

What is the error of our learning algorithm?

An error is made if the training points falls within the area between the true rectangle and our guess

So we can bound the error by estimating the probability of a point being in this area according to $\mathcal{D}$!

Now we just need to get enough training examples to attain any error bound ϵ with probability at most δ:

$$m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}$$

PAC Learning

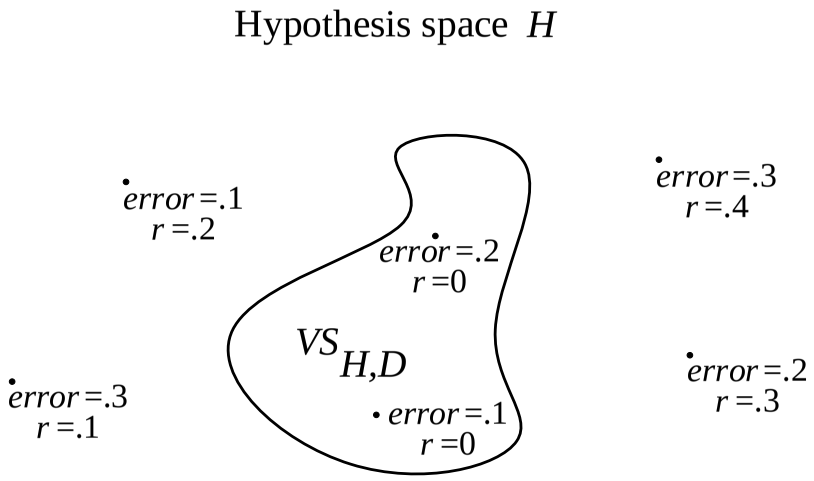
Consider a class C of possible target concepts defined over a set of instances X of length n, and a learner L using hypothesis space H.

> *Definition:* C is **PAC-learnable** by L using H if for all $c \in C$, 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 $error_{\mathcal{D}}(h) \leq \epsilon$.

Polynomial PAC-learning algorithms. The learner will output such a hypothesis in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $size(c)$.

Bounding the true error with zero training error

Version spaces:



(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 will ϵ-exhaust the VS?

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 less than

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

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

If we want to this probability to be below δ

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

then

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

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 our 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 $|H| = 3^n$, and

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

or

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

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

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

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

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

$$m \geq 98.8$$

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?

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

derived from Hoeffding bounds:

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