Computational Learning Theory

[read Chapter 7]
[Suggested exercises: 7.1, 7.2, 7.5, 7.8]

- Computational learning theory
- Setting 1: learner poses queries to teacher
- Setting 2: teacher chooses examples
- Setting 3: randomly generated instances, labeled by teacher
- Probably approximately correct (PAC) learning
- Vapnik-Chervonenkis Dimension

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

Prototypical Concept Learning Task

- Given:
  - Instances $X$: Possible days, each described by the attributes $Sky$, $AirTemp$, $Humidity$, $Wind$, $Water$, $Forecast$
  - Target function $c$: $EnjoySport : X \to \{0, 1\}$
  - Training examples $D$: Positive and negative examples of the target function $\langle x_1, c(x_1) \rangle, \ldots, \langle x_m, c(x_m) \rangle$

- 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 $X$?

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 \( V' \) 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
- \( \text{e.g., } (\text{AirTemp} = \text{Warm}) \land (\text{Wind} = \text{Strong}) \), where \( \text{AirTemp, Wind, ...} \) each have 2 possible values.
- if \( n \) possible boolean attributes in \( H \), \( n + 1 \) examples suffice
- why?

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 \( D \) over \( X \)

Learner observes a sequence \( D \) of training examples of form \( (x,c(x)) \), for some target concept \( c \in C \)
- instances \( x \) are drawn from distribution \( D \)
- teacher provides target value \( c(x) \) for each

Learner must output a hypothesis \( h \) estimating \( c \)
- \( h \) is evaluated by its performance on subsequent instances drawn according to \( D \)

Note: probabilistic instances, noise-free classifications

True Error of a Hypothesis

Definition: The true error (denoted \( \text{error}_D(h) \)) of 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 \( D \).

\[
\text{error}_D(h) = \Pr_{x \sim D}[c(x) \neq h(x)]
\]
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 $h$?
- First consider when training error of $h$ is zero (i.e., $h \in V_{S_{H,D}}$)

---

How many examples will $\epsilon$-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 $\epsilon$-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 $\text{error}(h) \geq \epsilon$

If we want this probability to be below $\delta$

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

then

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

---

Exhausting the Version Space

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

$$\forall h \in V_{S_{H,D}} \text{ error}_D(h) < \epsilon$$

---

Learning Conjunctions of Boolean Literals

How many examples are sufficient to assure with probability at least $(1 - \delta)$ that every $h$ in $V_{S_{H,D}}$ satisfies $\text{error}_D(h) \leq \epsilon$?

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

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

or

$$m \geq \frac{1}{\epsilon}(3^n + \ln(1/\delta))$$
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 we want to assure that with probability 95%, \( VS \) contains only hypotheses with error\(_p(h) \leq 1 \), then it is sufficient to have \( m \) examples, where

\[ m \geq \frac{1}{\epsilon} (\ln 973 + \ln(1/.05)) \]
\[ m \geq 10(\ln 973 + \ln 20) \]
\[ m \geq 10(6.88 + 3.00) \]
\[ m \geq 98.8 \]

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 \( D \) over \( X \), \( \epsilon \) such that \( 0 < \epsilon < 1/2 \), and \( \delta \) such that \( 0 < \delta < 1/2 \), learner \( L \) will with probability at least \((1 - \delta)\) output a hypothesis \( h \in H \) such that error\(_p(h) \leq \epsilon \), in time that is polynomial in \( 1/\epsilon, 1/\delta, n \) and size\(_c(c)\).

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.

Three Instances Shattered

Instance space \( X \)
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 Dim. of Linear Decision Surfaces

Sample Complexity from VC Dimension

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

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