Computational Learning Theory

Reading:
• Mitchell chapter 7

Suggested exercises:
• 7.1, 7.2, 7.5, 7.7

Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

October 12, 2006
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 function is approximated
- Manner in which training examples presented
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 \( c(x) \)
Instances, Hypotheses, and More-General-Than

$\mathbf{Instances \ X}$

$x_1 = \langle \text{Sunny, Warm, High, Strong, Cool, Same} \rangle$

$x_2 = \langle \text{Sunny, Warm, High, Light, Warm, Same} \rangle$

$\mathbf{Hypotheses \ H}$

$h_1 = \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle$

$h_2 = \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle$

$h_3 = \langle \text{Sunny, ?, ?, ?, Cool, ?} \rangle$
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 \( \langle x, c(x) \rangle \), 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: randomly drawn instances, noise-free classifications
True Error of a Hypothesis

Definition: The true error (denoted $error_D(h)$) 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 $\mathcal{D}$.

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

\[
error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)] \equiv \frac{\sum_{x \in \mathcal{D}} \delta(c(x) \neq h(x))}{|\mathcal{D}|}
\]

*True error* of hypothesis $h$ with respect to $c$

- How often $h(x) \neq c(x)$ over future instances drawn at random from $\mathcal{D}$

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

\[
\text{error}_\mathcal{D}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)] \equiv \frac{\sum_{x \in \mathcal{D}} \delta(c(x) \neq h(x))}{|\mathcal{D}|}
\]

**True error** of hypothesis \( h \) with respect to \( c \)

- How often \( h(x) \neq c(x) \) over future instances drawn at random from \( \mathcal{D} \)

\[
\text{error}_\mathcal{D}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]
\]
A hypothesis \( h \) is **consistent** with a set of training examples \( D \) of target concept \( c \) if and only if \( h(x) = c(x) \) for each training example \( \langle x, c(x) \rangle \) in \( D \).

\[
\text{Consistent}(h, D) \equiv (\forall \langle x, c(x) \rangle \in D) \ h(x) = c(x)
\]

The **version space**, \( VS_{H,D} \), with respect to hypothesis space \( H \) and training examples \( D \), is the subset of hypotheses from \( H \) consistent with all training examples in \( D \).

\[
VS_{H,D} \equiv \{ h \in H | \text{Consistent}(h, D) \}
\]
Exhausting the Version Space

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

$$(\forall h \in V S_{H,D}) \text{ error}_D(h) < \varepsilon$$
How many examples will $\varepsilon$-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 $0 \leq \varepsilon \leq 1$, the probability that the version space with respect to $H$ and $D$ is not $\varepsilon$-exhausted (with respect to $c$) is less than

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

Interesting! This bounds the probability that any consistent learner will output a hypothesis $h$ with $\text{error}(h) \geq \varepsilon$.

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

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

then

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

Any(!) learner that outputs a hypothesis consistent with all training examples (i.e., an $h$ contained in $\text{VS}_{H,D}$)
Let \( h_1, h_2, \ldots, h_k \) be the hyps in \( H \) with true error \( \geq \epsilon \).

Prob that one of these \( h_i \), will be consistent w/ one training examp.
\[ \leq (1-\epsilon) \]

Prob that one \( \ldots \) \[ \leq (1-\epsilon)^m \]

Prob at least one of the \( k \) "bad" hyps cons. with \( m \) ex.
\[ < k (1-\epsilon)^m \]

\[ k \leq |H| \]
\[ \leq |H| (1-\epsilon)^m \]

True \( \Rightarrow 0 \leq \epsilon \leq 1 \) then \( (1-\epsilon) \leq e^{-\epsilon} \)
\[ < |H| e^{-\epsilon m} \]
What it means

[Haussler, 1988]: probability that the version space is not $\varepsilon$-exhausted after $m$ training examples is at most $|H|e^{-\varepsilon m}$

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

Suppose we want this probability to be at most $\delta$

1. How many training examples suffice?
   $$m \geq \frac{1}{\varepsilon} \left( \ln |H| + \ln\left(\frac{1}{\delta}\right) \right)$$

2. If $error_{train}(h) = 0$ then with probability at least $(1-\delta)$:
   $$error_{true}(h) \leq \frac{1}{m} \left( \ln |H| + \ln\left(\frac{1}{\delta}\right) \right)$$
Learning Conjunctions of Boolean Literals

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

\[
every \ h \ in \ V S_{H,D} \ satisfies \ error_D(h) \leq \epsilon
\]

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).
Learning Conjunctions of Boolean Literals

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

\[
every \ h \ in \ VS_{H,D} \ satisfies \ error_D(h) \leq \epsilon
\]

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

\[
m \geq \frac{1}{\epsilon}(n \ln 3 + \ln(1/\delta))
\]
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 $X$, $\epsilon$ 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 $\text{error}_D(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, $n$ and $\text{size}(c)$. 
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 $X$, $\epsilon$ 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 $\text{error}_\mathcal{D}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, $n$ and $\text{size}(c)$.

Sufficient condition:
Holds if $L$ requires only a polynomial number of training examples, and processing per example is polynomial.
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?

\[
m \geq \frac{1}{2e^2} (\ln |H| + \ln(1/\delta))
\]

derived from Hoeffding bounds:

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

ture error training error degree of overfitting
Additive Hoeffding Bounds – Agnostic Learning

• Given $m$ independent coin flips of coin with $\Pr(\text{heads}) = \theta$
  bound the error in the estimate $\hat{\theta}$
  \[ \Pr[\theta > \hat{\theta} + \epsilon] \leq e^{-2m\epsilon^2} \]

• Relevance to agnostic learning: for any single hypothesis $h$
  \[ \Pr[error_{true}(h) > error_{train}(h) + \epsilon] \leq e^{-2m\epsilon^2} \]

• But we must consider all hypotheses in $H$
  \[ \Pr[(\exists h \in H) error_{true}(h) > error_{train}(h) + \epsilon] \leq |H| e^{-2m\epsilon^2} \]

• So, with probability at least $(1-\delta)$ every $h$ satisfies
  \[ error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}} \]
General Hoeffding Bounds

• When estimating parameter $\theta \in [a,b]$ from $m$ examples

$$P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

• When estimating a probability $\theta \in [0,1]$, so

$$P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{-2m\epsilon^2}$$

• And if we’re interested in only one-sided error, then

$$P((E[\hat{\theta}] - \hat{\theta}) > \epsilon) \leq e^{-2m\epsilon^2}$$