Computational Learning Theory

Reading:
• Mitchell chapter 7

Suggested exercises:
• 7.1, 7.2, 7.5, 7.7

Machine Learning 10-601

Arvind Rao
Machine Learning Department
Carnegie Mellon University

September 27, 2010
(originally, from slides by Prof. Tom Mitchell)
Announcements

• Problem set 2 is due Wednesday

• Please make sure that your submission directories work because we will not allow code submissions via email

• Monday's recitation is on problem set 2 – you can bring questions

• Problem set 3 will be posted Wednesday and will be due in 2 weeks (Oct 13, 2010)
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

$\begin{align*}
    x_1 &= \langle\text{Sunny, Warm, High, Strong, Cool, Same}\rangle \\
    x_2 &= \langle\text{Sunny, Warm, High, Light, Warm, Same}\rangle \\
    h_1 &= \langle\text{Sunny, ?, ?, Strong, ?, ?}\rangle \\
    h_2 &= \langle\text{Sunny, ?, ?, ?, ?, ?}\rangle \\
    h_3 &= \langle\text{Sunny, ?, ?, ?, Cool, ?}\rangle
\end{align*}$
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 $X$ $\mathcal{D} = P(X)$

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

Learner must output a hypothesis $h$ estimating $c$

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

Note: randomly drawn instances, noise-free classifications
**True Error of a Hypothesis**

Instance space $X$, $P(X) = \mathcal{D}$

Where $c$ and $h$ disagree

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

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

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

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

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

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

$$\text{error}_\mathcal{D}(h) \equiv \Pr_{x \in \mathcal{D}} [c(x) \neq h(x)]$$
Can we bound \( \text{error}_D(h) \) in terms of \( \text{error}_D(h) \)?

if \( D \) was a set of examples drawn from \( \mathcal{D} \) and independent of \( h \), then we could use standard statistical confidence intervals to determine that with 95% probability, \( \text{error}_D(h) \) lies in the interval:

\[
\text{error}_D(h) \pm 1.96 \sqrt{\frac{\text{error}_D(h)(1 - \text{error}_D(h))}{n}}
\]

but \( D \) is the training data for \( h \) ....
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 $D$.

$$Consistent(h, D) \equiv (\forall \langle x, c(x) \rangle \in D) \ h(x) = c(x)$$

The **version space**, $V_{S_{H,D}}$, with respect to hypothesis space $H$ and training examples $D$, is the subset of hypotheses from $H$ consistent with all training examples in $D$.

$$V_{S_{H,D}} \equiv \{h \in H | Consistent(h, D)\}$$
Exhausting the Version Space

Hypothesis space $H$

$$\begin{align*}
\text{error}.1 & \quad r=.2 \\
\text{error}.2 & \quad r=0 \\
\text{error}.3 & \quad r=.1 \\
\text{error}.1 & \quad r=0 \\
\text{error}.3 & \quad r=.4 \\
\text{error}.2 & \quad r=0 \\
\end{align*}$$

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

**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 true error less than $\epsilon$ with respect to $c$ and $D$.

$$(\forall h \in V S_{H,D}) \ error_D(h) < \epsilon$$
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

$$|H|e^{-\epsilon m}$$
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 $\text{error}(h) \geq \epsilon$
What it means

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

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

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

1. How many training examples suffice?

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

2. If $error_{train}(h) = 0$ then with probability at least $(1-\delta)$:

$$error_{true}(h) \leq \frac{1}{m} (\ln |H| + \ln(1/\delta))$$
Learning Conjunctions of Boolean Literals

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

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

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).

E.g.,

\(X = \langle X_1, X_2, \ldots, X_n \rangle\)

Each \(h \in H\) constrains each \(X_i\) to be 1, 0, or “don’t care”

In other words, each \(h\) is a rule such as:

If \(X_2 = 0\) and \(X_5 = 1\)

Then \(Y = 1\), else \(Y = 0\)
How many examples are sufficient to assure with probability at least $(1 - \delta)$ that

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

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

\[ m \geq \frac{1}{\epsilon} (n \ln 3 + \ln(1/\delta)) \]
Example: H is Conjunction of Boolean Literals

Consider classification problem $f: X \rightarrow Y$:

- instances: $<X_1 X_2 X_3 X_4>$ where each $X_i$ is boolean
- learned hypotheses are rules of the form:
  - IF $<X_1 X_2 X_3 X_4> = <0, ?, 1, ?>$, THEN $Y=1$, ELSE $Y=0$
  - i.e., rules constrain any subset of the $X_i$

How many training examples $m$ suffice to assure that with probability at least 0.9, any consistent learner will output a hypothesis with true error at most 0.05?
Example: H is Decision Tree with depth=2
\[ m \geq \frac{1}{\epsilon} (\ln |H| + \ln(1/\delta)) \]

Consider classification problem \( f: X \to Y \):
- instances: \( <X_1, \ldots, X_N> \) where each \( X_i \) is boolean
- learned hypotheses are decision trees of depth 2, using only two variables

How many training examples \( m \) suffice to assure that with probability at least 0.9, any consistent learner will output a hypothesis with true error at most 0.05?
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 $\text{size}(c)$. 
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 $\text{size}(c)$.

**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?

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

  derived from Hoeffding bounds:

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

  note $\epsilon$ here is the difference between the training error and true error

true 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 maximum likelihood estimate $\hat{\theta}$

\[
\Pr[\theta > \hat{\theta} + \epsilon] \leq e^{-2m\epsilon^2}
\]

- Relevance to agnostic learning: for any \textit{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 the mean $\theta$ inside $[a,b]$ from $m$ examples
  \[ P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{\frac{-2m\epsilon^2}{(b-a)^2}} \]

- When estimating a probability $\theta$ is inside $[0,1]$, so
  \[ P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{-2m\epsilon^2} \]

- And if we’re interested in only one-sided error, then
  \[ P((E[\hat{\theta}] - \hat{\theta}) > \epsilon) \leq e^{-2m\epsilon^2} \]
What if H is not finite?

• Can’t use our result for finite H

• Need some other measure of complexity for H
  – Vapnik-Chervonenkis (VC) dimension!
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.
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 $VC(H) \equiv \infty$. 

$VC(H) = 3$
Sample Complexity based on VC dimension

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

ie., to guarantee that any hypothesis that perfectly fits the training data is probably $(1-\delta)$ approximately $(\varepsilon)$ correct

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

Compare to our earlier results based on $|H|$:

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

Consider \( X = <, \) want to learn \( c: X \rightarrow \{0,1\} \)

What is VC dimension of

- Open intervals:
  \[-H1: \text{if } x > a \text{ then } y = 1 \text{ else } y = 0 \quad \text{VC}(H1)=1\]
  \[-H2: \text{if } x > a \text{ then } y = 1 \text{ else } y = 0\]
  \[\text{or, if } x > a \text{ then } y = 0 \text{ else } y = 1 \quad \text{VC}(H2)=2\]

- Closed intervals:
  \[-H3: \text{if } a < x < b \text{ then } y = 1 \text{ else } y = 0 \quad \text{VC}(H3)=2\]
  \[-H4: \text{if } a < x < b \text{ then } y = 1 \text{ else } y = 0\]
  \[\text{or, if } a < x < b \text{ then } y = 0 \text{ else } y = 1 \quad \text{VC}(H4)=3\]
VC dimension: examples

What is VC dimension of lines in a plane?

- $H_2 = \{ ((w_0 + w_1x_1 + w_2x_2)>0 \Rightarrow y=1) \}$
VC dimension: examples

What is VC dimension of

• $H_2 = \{ ((w_0 + w_1x_1 + w_2x_2)>0 \Rightarrow y=1) \}$
  – $VC(H_2)=3$

• For $H_n =$ linear separating hyperplanes in n dimensions, $VC(H_n)=n+1$
For any finite hypothesis space $H$, can you give an upper bound on $\text{VC}(H)$ in terms of $|H|$? (hint: yes)
More VC Dimension Examples to Think About

- Logistic regression over \( n \) continuous features
  - Over \( n \) boolean features?

- Linear SVM over \( n \) continuous features

- Decision trees defined over \( n \) boolean features
  \( F: \langle X_1, \ldots, X_n\rangle \rightarrow Y \)

- Decision trees of depth 2 defined over \( n \) features

- How about 1-nearest neighbor?
Tightness of Bounds on Sample Complexity

How many examples $m$ suffice to assure that any hypothesis that fits the training data perfectly is probably $(1-\delta)$ approximately $(\varepsilon)$ correct?

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

How tight is this bound?
Tightness of Bounds on Sample Complexity

How many examples \( m \) suffice to assure that any hypothesis that fits the training data perfectly is probably \((1-\delta)\) approximately \((\varepsilon)\) correct?

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

How tight is this bound?

**Lower bound on sample complexity** (Ehrenfeucht et al., 1989):

Consider any class \( C \) of concepts such that \( \text{VC}(C) > 1 \), any learner \( L \), any \( 0 < \varepsilon < \frac{1}{8} \), and any \( 0 < \delta < 0.01 \). Then there exists a distribution \( D \) and a target concept in \( C \), such that if \( L \) observes fewer examples than

\[
\max \left[ \frac{1}{\varepsilon} \log \left( \frac{1}{\delta} \right), \frac{\text{VC}(C) - 1}{32\varepsilon} \right]
\]

Then with probability at least \( \delta \), \( L \) outputs a hypothesis with \( \text{error}_D(h) > \varepsilon \).
Agnostic Learning: VC Bounds

With probability at least \((1 - \delta)\) every \(h \in H\) satisfies

\[
\text{error}_{\text{true}}(h) < \text{error}_{\text{train}}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}
\]
Structural Risk Minimization [Vapnik]

Which hypothesis space should we choose?
- Bias / variance tradeoff

SRM: choose $H$ to minimize bound on true error!

$$error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H) \left( \ln \frac{2m}{VC(H)} + 1 \right) + \ln \frac{4}{\delta}}{m}}$$

* unfortunately a somewhat loose bound...