Computational Learning Theory

CS 486/686: Introduction to Artificial Intelligence
Overview

– Introduction to Computational Learning Theory
– PAC Learning Theory

Thanks to T Mitchell
Introduction

• Recall how *inductive learning* works
  
  - Given a training set of examples of the form \((x, f(x))\) return a function \(h\) (a hypothesis) that approximates \(f\)

Decision Trees:
Boolean functions
Computational Learning Theory

- Are there general laws for inductive learning?
- Theory to relate
  - Probability of successful learning
  - Number of training examples
  - Complexity of hypothesis space
  - Accuracy to which \( f \) is approximated
  - Manner in which training examples are presented
Computational Learning Theory

- Sample complexity
  - How many training examples are needed to learn the target function, f?
Computational Learning Theory

- Sample complexity
  - How many training examples are needed to learn the target function, f?

- Computational complexity
  - How much computation effort is needed to learn the target function, f?
Computational Learning Theory

• Sample complexity
  – How many training examples are needed to learn the target function, f?

• Computational complexity
  – How much computation effort is needed to learn the target function, f?

• Mistake bound
  – How many training examples will a learner misclassify before learning the target function, f?
Sample Complexity

• How many examples are sufficient to learn f?
  – If learner proposes instances as queries to a teacher
    • learner proposes x, teacher provides f(x)
  – If the teacher provides training examples
    • teacher provides a sequence of examples of the form (x, f(x))
  – If some random process proposes instances
    • instance x generated randomly, teacher provide f(x)
Function Approximation

\[ H = \{ h : X \rightarrow \{0, 1\} \} \]
\[ |H| = 2^{|X|} = 2^{2^{20}} \]

\[ X = \langle x_1, \ldots, x_{20} \rangle \]
\[ |X| = 2^{20} \]
Function Approximation

How many labelled examples are needed in order to determine which of the hypothesis is the correct one?

All $2^{20}$ instances in $X$ must be labelled!

There is no free lunch!
Sample Complexity

• Given
  – set of instances $X$
  – set of hypothesis $H$
  – set of possible target concepts $F$
  – training instances generated by a fixed unknown probability distribution $D$ over $X$

• Learner observes sequence, $D$, of training examples $(x, f(x))$
  – instances are drawn from distribution $D$
  – teacher provides $f(x)$ for each instance

• Learner must output a hypothesis $h$ estimating $f$
  – $h$ is evaluated by its performance on future instances drawn according to $D$
True Error of a Hypothesis

• The *true error* \( \text{error}_D(h) \) of hypothesis \( h \) with respect to target function \( f \) and distribution \( D \) is the probability that \( h \) will misclassify an instance drawn at random according to \( D \)

\[
\text{error}_D(h) = \Pr_{x \in D}[f(x) \neq h(x)]
\]

Figure from *Machine Learning* by T Mitchell
Note our notation is a bit different:
We use \( f \) instead of \( c \)
Training Error vs True Error

• Training error of h wrt target function f
  – How often $h(x) \neq f(x)$ over training instances $D$
    - $\text{error}_D(h) = \Pr_{x \in D}[f(x) \neq h(x)] = \#(f(x) \neq h(x))/|D|$
    - Note: A consistent h will have $\text{error}_D(h) = 0$

• True error of h wrt to target function f
  – How often $h(x) \neq f(x)$ over future instances drawn at random from $D$
    - $\text{error}_D(h) = \Pr_{x \in D}[f(x) \neq h(x)]$
Version Spaces

• A hypothesis $h$ is consistent with a set of training examples $D$ of target function $f$ if and only if $h(x) = f(x)$.

\[
Consistent(h, D) \equiv (\forall (x, f(x)) \in D) h(x) = f(x)
\]

• A version space, $VS_{H,D}$, is the set of all hypothesis from $H$ that are consistent with all training examples in $D$

\[
VS_{H,D} \equiv \{ h \in H | Consistent(h, D) \}\]
$\varepsilon$-Exhausting $V_{H,D}$

- Version space $V_{H,D}$ is $\varepsilon$-exhausted wrt $f$ and $D$ if every $h$ in $V_{H,D}$ has true error less than $\varepsilon$ wrt $f$ and $D$

\[ (\forall h \in V_{H,D}) \text{error}_D(h) < \varepsilon \]

- $r$: training error
- error: true error

$\varepsilon$-exhausted for $\varepsilon>0.2$
How many examples will $\varepsilon$-exhaust the VS?

- Theorem[Haussler, 1988]. If
  - $H$ is finite and
  - $D$ is a sequence of $m \geq 1$ independent random examples of target function $f$

- Then for any $0 \leq \varepsilon \leq 1$, the probability that $V^S_{H,D}$ is not $\varepsilon$-exhausted (wrt $f$) is less than $|H|e^{-\varepsilon m}$
How many examples will $\varepsilon$-exhaust the VS?

- Theorem[Haussler, 1988]. If
  - $H$ is finite and
  - $D$ is a sequence of $m \geq 1$ independent random examples of target function $f$

- Then for any $0 \leq \varepsilon \leq 1$, the probability that $V_{SH,D}$ is not $\varepsilon$-exhausted (wrt $f$) is less than $|H|e^{-\varepsilon m}$
How many examples will $\varepsilon$-exhaust the VS?

• Theorem [Haussler, 1988]. If $H$ is finite and $D$ is a sequence of $m \geq 1$ independent random examples of target function $f$.

• Then for any $0 \leq \varepsilon \leq 1$, the probability that $V_{SH,D}$ is not $\varepsilon$-exhausted (wrt $f$) is less than $|H|e^{-\varepsilon m}$

Interesting!

This bounds the probability that any consistent learner will output a hypothesis $h$ with $\text{error}(h) \geq \varepsilon$.
Proof
How to interpret this result

\[ Pr[(\exists h \in H) st(error_D(h) = 0) \land (error_D(h) > \epsilon)] \leq |H|e^{-\epsilon m} \]

- Suppose we want this probability to be at most \( \delta \)
  - How many training examples are needed?
    \[ m \geq \frac{1}{\epsilon} (\ln |H| + \ln(1/\delta)) \]
  - If \( error_D(h) = 0 \) then with probability at least \((1-\delta)\)
    \[ error_D(h) \leq \frac{1}{m} (\ln |H| + \ln(1/\delta)) \]
Decision Tree Example
• Probably Approximately Correct (PAC) Learner
  - Given a class $C$ of possible target concepts ($f$) defined over a set of instances $X$ of length $n$, and a learner $L$ using hypothesis space $H$
  - $C$ is **PAC-learnable** by $L$ using $H$ if for all $f$ in $C$, distributions $D$ over $X$, $0<\varepsilon<1/2$ and $0<\delta<1/2$, learner $L$ will with probability at least $(1-\delta)$ output a hypothesis $h$ such that $error_D(h) \leq \varepsilon$ in time polynomial in $1/\varepsilon$, $1/\delta$, $n$, and size($f$)
Agnostic Learners

• So far we have assumed f is in H. What if we do not make this assumption? (Agnostic Learning)

• Goal: Hypothesis h that makes the fewest errors on the training data!
Agnostic Learning Derivation

- **Hoeffding Bound** (additive Chernoff bound): given a coin with \( \Pr(\text{heads}) = \theta \), after \( m \) independent coin flips you observe \( \Pr(\text{heads}|m) = \theta_m \)

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

- In our case, for any single hypothesis

\[
\Pr(\text{error}_D(h) > \text{error}_D(h) + \epsilon) \leq e^{-2m\epsilon^2}
\]
Agnostic Learning Derivation

• To assure that the best hypothesis found by L has an error bounded this way

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

• Sample complexity

\[ m \geq \frac{1}{2\epsilon^2}(\ln |H| + \ln(1/\delta)) \]
Infinite Hypothesis Spaces

• Recall

\[ m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln(1/\delta) \right) \]

– What if \( H = \{ h | h : X \to Y \} \) is infinite? What is the “right” measure of complexity?

– The largest subset of \( X \) for which \( H \) can guarantee zero training error (regardless of target function \( f \))

• Vapnik-Chervonenkis (VC) dimension
Shattering

• A *dichotomy* of a set $S$ is a partition of $S$ into two disjoint subsets
  – Note: Given $S$ there are $2^{|S|}$ dichotomies

• 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
VC Dimension

- The VC(H) dimension of hypothesis space H defined over input space X is the size of the largest finite subset of X shattered by H.
  - If arbitrarily large finite subsets of X can be shattered by H then VC(H) = ∞
VC Examples

• Let $X=\mathbb{R}$, $H=$ all open intervals, $h:a<x<b$ ($h(x)=1$ is $a<x<b$, 0 otherwise)

• Let $S=\{3.1, 4.5\}$, $|S|=2$
  - dichotomies: both are 1, both are 0, one is 1 and the other 0
  - $H$ can shatter $S$
  - $VC(H)$ is at least 2

• What about $S=\{x,y,z\}$ where $x<y<z$?
Sample Complexity

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

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

• Lower bound: For any $0<\varepsilon<1/8$ and $0<\delta<0.01$, there exists some $D$ and a $f$ in $C$ such that if $L$ observes fewer than

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

• Then with prob. at least $\delta$, $L$ outputs a hypothesis with $error_D(h)>\varepsilon$
Changing Directions: Observations about No Free Lunch

- I mentioned “No Free Lunch” earlier in today's lecture
- What do we really mean by “No Free Lunch”?
No Free Lunch Principle

• “No learning is possible without the application of prior domain knowledge”

• For any learning algorithm, there is some distribution that generates data such that when trained over this distribution will produce large error. If ISI is much smaller than IXI then error can be close to 0.5.