Computational Learning Theory – Part 2

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

November 3, 2010
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 $(1-\delta)$:

$$error_{true}(h) \leq \frac{1}{m}(\ln |H| + \ln(1/\delta))$$
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 $\text{size}(c)$.
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.
If $H = \{h \mid h: X \to Y\}$ is infinite, what measure of complexity should we use in place of $|H|$?

$$m \geq \frac{1}{\epsilon} (\ln |H| + \ln(1/\delta))$$
Question: If $H = \{h \mid h: X \rightarrow Y\}$ is infinite, what measure of complexity should we use in place of $|H|$?

Answer: The largest subset of $X$ for which $H$ can guarantee zero training error (regardless of the target function $c$)
If $H = \{h \mid h: X \rightarrow Y\}$ is infinite, what measure of complexity should we use in place of $|H|$?

Answer: The largest subset of $X$ for which $H$ can guarantee zero training error (regardless of the target function $c$)

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

VC dimension of $H$ is the size of this subset
Question: If $H = \{h \mid h: X \to Y\}$ is infinite, what measure of complexity should we use in place of $|H|$?

Answer: The largest subset of $X$ for which $H$ can guarantee zero training error (regardless of the target function $c$)

Informal intuition:
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 $VC(H) \equiv \infty$.

![Diagram showing VC(H) = 3](image)
Sample Complexity based on VC dimension

How many randomly drawn examples suffice to $\varepsilon$-exhaust $V_{S_{H,D}}$ 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: if $x > a$ then $y = 1$ else $y = 0$

  H2: if $x > a$ then $y = 1$ else $y = 0$  
  or, if $x > a$ then $y = 0$ else $y = 1$

- **Closed intervals:**

  H3: if $a < x < b$ then $y = 1$ else $y = 0$

  H4: if $a < x < b$ then $y = 1$ else $y = 0$  
  or, if $a < x < b$ then $y = 0$ else $y = 1$
VC dimension: examples

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

What is VC dimension of

- Open intervals:
  - $H_1$: if $x > a$ then $y = 1$ else $y = 0$ \quad VC($H_1$)=1
  - $H_2$: if $x > a$ then $y = 1$ else $y = 0$ or, if $x > a$ then $y = 0$ else $y = 1$ \quad VC($H_2$)=2

- Closed intervals:
  - $H_3$: if $a < x < b$ then $y = 1$ else $y = 0$ \quad VC($H_3$)=2
  - $H_4$: if $a < x < b$ then $y = 1$ else $y = 0$ or, if $a < x < b$ then $y = 0$ else $y = 1$ \quad VC($H_4$)=3
What is VC dimension of lines in a plane?

- $H_2 = \{ (w_0 + w_1 x_1 + w_2 x_2) > 0 \Rightarrow y=1 \}$
VC dimension: examples

What is VC dimension of

- $H_2 = \{ ((w_0 + w_1 x_1 + w_2 x_2) > 0 \rightarrow y=1) \}$
  - $VC(H_2) = 3$

- 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 $VC(H)$ 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?
How tight is this bound?

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

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

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?

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

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 \( 0 < \varepsilon < 1/8 \), and any \( 0 < \delta < 0.01 \). Then there exists a distribution \( \mathcal{D} \) and a target concept in \( C \), such that if \( L \) observes fewer examples than

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

Then with probability at least \( \delta \), \( L \) outputs a hypothesis with \( error_{\mathcal{D}}(h) > \varepsilon \)
Agnostic Learning: VC Bounds

[Schölkopf and Smola, 2002]

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

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

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

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

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

* unfortunately a somewhat loose bound...
Mistake Bounds

So far: how many examples needed to learn?
What about: how many mistakes before convergence?

Let’s consider similar setting to PAC learning:

- Instances drawn at random from $X$ according to distribution $D$
- Learner must classify each instance before receiving correct classification from teacher
- Can we bound the number of mistakes learner makes before converging?
Mistake Bounds: Find-S

Consider Find-S when $H =$ conjunction of boolean literals

**Find-S:**
- Initialize $h$ to the most specific hypothesis $l_1 \land \neg l_1 \land l_2 \land \neg l_2 \ldots l_n \land \neg l_n$
- For each positive training instance $x$
  - Remove from $h$ any literal that is not satisfied by $x$
- Output hypothesis $h$.

How many mistakes before converging to correct $h$?
Mistake Bounds: Halving Algorithm

Consider the Halving Algorithm:

- Learn concept using version space \textsc{Candidate-Elimination} algorithm
- Classify new instances by majority vote of version space members

How many mistakes before converging to correct $h$?

- ... in worst case?
- ... in best case?

1. Initialize $\text{VS} \leftarrow H$

2. For each training example,
   - remove from $\text{VS}$ every hypothesis that misclassifies this example
Optimal Mistake Bounds

Let $M_A(C)$ be the max number of mistakes made by algorithm $A$ to learn concepts in $C$. (maximum over all possible $c \in C$, and all possible training sequences)

$$M_A(C) \equiv \max_{c \in C} M_A(c)$$

Definition: Let $C$ be an arbitrary non-empty concept class. The optimal mistake bound for $C$, denoted $Opt(C)$, is the minimum over all possible learning algorithms $A$ of $M_A(C)$.

$$Opt(C) \equiv \min_{A \in \text{learning algorithms}} M_A(C)$$

$VC(C) \leq Opt(C) \leq M_{\text{Halving}}(C) \leq \log_2(|C|).$
Weighted Majority Algorithm

\(a_i\) denotes the \(i^{th}\) prediction algorithm in the pool \(A\) of algorithms. \(w_i\) denotes the weight associated with \(a_i\).

- For all \(i\) initialize \(w_i \leftarrow 1\)
- For each training example \(\langle x, c(x) \rangle\)
  * Initialize \(q_0\) and \(q_1\) to 0
  * For each prediction algorithm \(a_i\)
    - If \(a_i(x) = 0\) then \(q_0 \leftarrow q_0 + w_i\)
      If \(a_i(x) = 1\) then \(q_1 \leftarrow q_1 + w_i\)
  * If \(q_1 > q_0\) then predict \(c(x) = 1\)
  * If \(q_0 > q_1\) then predict \(c(x) = 0\)
  * If \(q_1 = q_0\) then predict 0 or 1 at random for \(c(x)\)
  * For each prediction algorithm \(a_i\) in \(A\) do
    - If \(a_i(x) \neq c(x)\) then \(w_i \leftarrow \beta w_i\)

when \(\beta=0\), equivalent to the Halving algorithm...
Weighted Majority

[Relative mistake bound for WEIGHTED-MAJORITY] Let $D$ be any sequence of training examples, let $A$ be any set of $n$ prediction algorithms, and let $k$ be the minimum number of mistakes made by any algorithm in $A$ for the training sequence $D$. Then the number of mistakes over $D$ made by the WEIGHTED-MAJORITY algorithm using $\beta = \frac{1}{2}$ is at most

$$2.4(k + \log_2 n)$$
What You Should Know

• Sample complexity varies with the learning setting
  – Learner actively queries trainer
  – Examples arrive at random
  – ...

• Within the PAC learning setting, we can bound the probability that learner will output hypothesis with given error
  – For ANY consistent learner (case where \( c \in H \))
  – For ANY “best fit” hypothesis (agnostic learning, where perhaps \( c \) not in \( H \))

• VC dimension as measure of complexity of \( H \)

• Mistake bounds

• Conference on Learning Theory: [http://www.learningtheory.org](http://www.learningtheory.org)
• Avrim Blum’s course on Machine Learning Theory: