Computational Learning Theory Part 2
VC dimension, Sample Complexity, Mistake bounds

Required reading:
• Mitchell chapter 7

Optional advanced reading:
• Kearns & Vazirani, ‘Introduction to Computational Learning Theory’
Last time: PAC Learning

1. Finite H, assume target function \( c \in H \)

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

Suppose we want this to be at most \( \delta \). Then \( m \) examples suffice:

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

2. Finite H, agnostic learning: perhaps \( c \) not in H

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

\[
error_{\text{true}}(h) \leq error_{\text{train}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}
\]
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 \( VC(H) \equiv \infty \).
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

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

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

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

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

What is VC dimension of

- Open intervals:
  \[ 1 = \text{Vc}(H_1): \text{if } x > a \text{ then } y = 1 \text{ else } y = 0 \]
  \[ 2 = \text{Vc}(H_2): \text{if } x > a \text{ then } y = 1 \text{ else } y = 0 \]
  or, if \( x > a \) then \( y = 0 \) else \( y = 1 \)

- Closed intervals:
  \[ 2 = \text{Vc}(H_3): \text{if } a < x < b \text{ then } y = 1 \text{ else } y = 0 \]
  \[ H_4: \text{if } a < x < b \text{ then } y = 1 \text{ else } y = 0 \]
  or, if \( a < x < b \) then \( y = 0 \) else \( y = 1 \)
VC dimension: examples

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

What is VC dimension of

- **Open intervals:**
  
  $H_1$: if $x > a$ then $y = 1$ else $y = 0$  \hspace{1cm} \text{VC}(H_1)=1

  $H_2$: if $x > a$ then $y = 1$ else $y = 0$
  or, if $x > a$ then $y = 0$ else $y = 1$  \hspace{1cm} \text{VC}(H_2)=2

- **Closed intervals:**
  
  $H_3$: if $a < x < b$ then $y = 1$ else $y = 0$  \hspace{1cm} \text{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 $y = 1$  \hspace{1cm} \text{VC}(H_4)=3
VC dimension: examples

Consider $X = \mathbb{R}^2$, want to learn $c: X \rightarrow \{0, 1\}$

What is VC dimension of lines in a plane?
- $H = \{ ((wx+b)>0 \rightarrow y=1) \}$

$H =$ linear sep. in $n$ dimensions
$\Rightarrow VC(H) = n+1$
Consider $X = \mathbb{R}^2$, want to learn $c:X \rightarrow \{0,1\}$

What is VC dimension of
- $H = \{ (w \cdot x + b) > 0 \rightarrow y = 1 \}$
  - $VC(H_1) = 3$
  - For linear separating hyperplanes in $n$ dimensions, $VC(H) = n + 1$
For any finite hypothesis space $H$, give an upper bound on $VC(H)$ in terms of $|H|$.

$VC(H) \leq \log_2 |H|$
More VC Dimension Examples

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

• Decision trees defined over n continuous features
  Where each internal tree node involves a threshold test \((X_i > c)\)

• Decision trees of depth 2 defined over n features

• Logistic regression over n continuous features? Over n boolean 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?

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

How tight is this bound?

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

Consider any class \( C \) of concepts such that \( VC(C) \geq 2 \), any learner \( L \), any \( 0 < \varepsilon < 1/8 \), and any \( 0 < \delta < 0.01 \). Then there exists a distribution \( \mathcal{D} \) and target concept in \( C \), such that if \( L \) observes fewer examples than

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

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

[Schölkopf and Smola, 2002]

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

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

\[
|e_{true}(h) - e_{train}(h)|
\]
Structural Risk Minimization [Vapnik]

Which hypothesis space should we choose?

- Bias / variance tradeoff

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

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

* unfortunately a somewhat loose bound...
What You Should Know

• Sample complexity varies with the learning setting
  – Learner actively queries trainer
  – Examples provided 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 \( \notin \) H)
  – For ANY “best fit” hypothesis (agnostic learning, where perhaps c not in H)

• VC dimension as measure of complexity of H

• Quantitative bounds characterizing bias/variance in choice of H
  – but the bounds are quite loose...

• Mistake bounds in learning

• Conference on Learning Theory: http://www.learningtheory.org