Computational Learning Theory
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 $c \in H$

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

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

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

![Instance space $X$](image)

$VC(H) = 3$
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

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

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

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

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

What is VC dimension of

- Open intervals:
  
  $H_1$: if $x > a$ then $y = 1$ else $y = 0$  \quad VC(H1)=1

  $H_2$: if $x > a$ then $y = 1$ else $y = 0$
  or, if $x > a$ then $y = 0$ else $y = 1$  \quad VC(H2)=2

- Closed intervals:
  
  $H_3$: if $a < x < b$ then $y = 1$ else $y = 0$  \quad VC(H3)=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(H4)=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?

- $H = \{ ((w \cdot x + b) > 0 \rightarrow y = 1) \mid w \in \mathbb{R}^2, b \in \mathbb{R} \}$
Consider $X = \mathbb{R}^2$, want to learn $c: X \rightarrow \{0,1\}$

What is VC dimension of

- $H = \{ ((w \cdot x + b) > 0 \rightarrow y = 1) \mid w \in \mathbb{R}^2, b \in \mathbb{R} \}$
  - $VC(H_1) = 3$
  - For linear separating hyperplanes in $n$ dimensions, $VC(H) = n + 1$
For any finite hypothesis space $H$, give an upper bound on $\text{VC}(H)$ in terms of $|H|$.
More VC Dimension Examples

• Decision trees defined over \( n \) boolean features
  \[ F: <X_1, \ldots, X_n> \rightarrow Y \]

• 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 $(\epsilon)$ correct?

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

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 $0 < \epsilon < 1/8$, 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

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

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

[Schölkopf and Smola, 2002]

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

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

Which hypothesis space should we choose?

- Bias / variance tradeoff

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

* unfortunately a somewhat loose bound...
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 $H = \text{conjunction of boolean literals}$

\[
\text{FIND-S:}
\]
\begin{itemize}
  \item Initialize $h$ to the most specific hypothesis \\
  $l_1 \land \neg l_1 \land l_2 \land \neg l_2 \ldots l_n \land \neg l_n$
  \item For each positive training instance $x$
    \begin{itemize}
    \item Remove from $h$ any literal that is not satisfied by $x$
    \end{itemize}
  \item Output hypothesis $h$.
\end{itemize}

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

Consider the Halving Algorithm:

- Learn concept using version space
  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 VS $\leftarrow H$
2. For each training example,
   - remove from 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 $M_A(C)$.

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

$$VC(C) \leq Opt(C) \leq M_{Halving}(C) \leq \log_2(|C|).$$
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 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 $c \in H$)
  - 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

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