Computational Learning Theory – Part 2

04-28-06
What we have seen so far

Finite hypothesis space with consistent hypothesis

• We have shown for consistent-learn
  – Sample complexity required to ensure \((1 - \delta)\) probability of returning a \(\epsilon\)-good hypothesis is
    \[
    m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right)
    \]
  – Given \(m\) samples, with at least \(1 - \delta\) prob., the learned hypothesis will have generalization error
    \[
    \epsilon \leq \frac{1}{m} \left( \ln |H| + \ln \frac{1}{\delta} \right)
    \]

• What if there is no consistent hypothesis?
  – Learner finds the hypothesis \(h\) that minimizes training error
  – Suppose \(h\) has training error \(\epsilon_T\), what can we say about its generalization error?
Additive Hoeffding Bound

- Let $Z$ be a binary r.v. with $P(Z = 1) = p$
- Let $z_1, z_2, ..., z_m$ be $m$ i.i.d. samples of $Z$
- Hoeffding bound gives a bound on the probability that the sample average deviates away from the true mean $p$ by $\gamma$:

\[
P\left(p - \frac{1}{m} \sum z_i > \gamma \right) \leq \exp(-2\gamma^2 m)
\]
\[
P\left(p - \frac{1}{m} \sum z_i < -\gamma \right) \leq \exp(-2\gamma^2 m)
\]
Hoeffding Bound for Generalization Error

• Suppose $h$ has training error $\varepsilon_T$, what can we say about its generalization error?

• To apply Hoeffding bound, Let $Z$ be a Bernoulli r.v. defined by:
  
  - Draw a sample $x$ from $D$, $Z = \begin{cases} 
  1 & \text{if } h(x) \neq f(x) \\
  0 & \text{if } h(x) = f(x) 
  \end{cases}$

• With $m$ training examples, we get $m$ samples of $Z: z_1, \ldots, z_m$ and the training error of $h$ is
  
  $\varepsilon_T = \frac{1}{m} \sum_{i=1}^{m} z_i$

• The true mean of $Z$ is the generalization error of $h$

• Using Hoeffding bounds, we have:
  
  $P(\varepsilon > \varepsilon_T + \gamma) \leq \exp(-2\gamma^2 m)$

• As the training set grows the probability that the training error underestimates the generalization error decreases exponentially fast
Error Bound: Inconsistent Hypothesis

• This suggests that for a random $h \in H$, if the training error of $h$ is $\varepsilon_T$, then we have:

$$P(\varepsilon > \varepsilon_T + \gamma) \leq \exp(-2\gamma^2 m)$$

• Now we would like to bound this for all $h \in H$ simultaneously

$$P(\exists h \in H: \varepsilon(h) > \varepsilon_T(h) + \gamma) \leq |H| \exp(-2\gamma^2 m)$$

• Now bound this probability by $\delta$, and that we have $m$ samples, it is thus guaranteed with probability at $1 - \delta$ that for all $h \in H$:

$$\varepsilon(h) < \varepsilon_T(h) + \gamma = \varepsilon_T(h) + \sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}}$$
Best Possible Hypothesis in H

• Theorem
Consider a learner that always outputs the hypothesis that minimizes the training error, i.e., \( h_L = \arg\min_{h \in H} \epsilon_T(h) \). Let \( m \) be the size of the training set, with probability \( 1-\delta \), we have

\[
\epsilon(h_L) < \epsilon(h^*) + 2 \sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}}
\]

Interpretation: by selecting \( h_L \), we are not too worse off than the best possible hypothesis \( h^* \) and the difference gets smaller as we increase \( m \).
Proof

\[ \epsilon(h_L) \leq \epsilon_T(h_L) + \sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}} \]

\[ \leq \epsilon_T(h^*) + \sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}} \]

\[ \leq \epsilon(h^*) + \sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}} + \sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}} \]
Interpretation

$$\epsilon(h_L) < \epsilon(h^*) + 2\sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}}$$

- **Fundamental tradeoff in selecting Hypothesis space**
  - Bigger hypothesis space causes the 1st term to decrease
  - (this is sometimes called the “bias” of H)
- **However, as |H| increases, the second term increases**
  - (this is related to the “variance” of the learning algorithm)
What we have seen so far

Finite hypothesis space:

• Consistent hypothesis
  – Sample complexity required to ensure \((1 - \delta)\) probability of returning a \(\epsilon\)-good hypothesis is
    \[ m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]
  – Given \(m\) samples, with at least \((1 - \delta)\) prob., the learned hypothesis will have generalization error
    \[ \epsilon \leq \frac{1}{m} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]

• no consistent hypothesis, learner finds the hypothesis \(h\) that minimizes training error
    \[ \epsilon(h_L) < \epsilon(h^*) + 2 \sqrt{\frac{1}{2m} \log \frac{|H|}{\delta}} \]
Case 2: Infinite Hypothesis Space
Case 2: Infinite Hypothesis Space

• Most of our classifiers (LTUs, neural networks, SVMs) have continuous parameters and therefore, have infinite hypothesis spaces

• For some, despite their infinite size, they have limited expressive power, so we should be able to prove something

• We need to characterize the learner’s ability to model complex concepts
  – For finite spaces the complexity of a hypothesis space was characterized roughly by $|H|$
  – For infinite spaces, we will introduce a concept called vc-dimension
**Definition: Shatter**

- **Definition:** Consider $S$, a set of $m$ points in the input space, a hypothesis space $H$ is said to **shatter** $S$ if for every possible way of labeling the points in $S$, there exists an $h \in H$ that correctly classify them.

- **Example:**
  
  $S$: 3 points in 2d, 8 labelings
  
  $H$: space of 2-d linear functions
  
  $H$ shatters $S$!
Definition: VC-dimension

- Definition: The Vapnik-Chervonenkis dimension (VC-dimension) of an hypothesis space H is the size of the largest set S that can be shattered by H
  - As long as we can find one set of size m that can be shattered by H, then $VC(H) \geq m$
  - It does not matter if some other set of size m can not be shattered by H

- So to prove that the VC dimension of H is m, we need to:
  - Show there exists a set of size m that can be shattered by H
  - Show that no set of size m+1 can be shattered by H
VC-dimension Example (1)

• Let $H$ be the set of intervals on the real line such that $h(x) = 1$ iff $x$ is in the interval. $H$ can fit any pair of examples:

![Diagram of intervals fitting two examples]

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• However, $H$ can not always fit any three examples. Therefore the VC-dimension of $H$ is 2

![Diagram of intervals failing to fit three examples]

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
VC-dimension Example (2)

• Let H be the space of linear function in the 2-D plane. We have shown that H can shatter the follow 3 points.

• But H cannot shatter these 3 points

• Since there is at least one set of size 3 that can be shattered, we say $VC(H) \geq 3$
VC-dimension Example (3)

- VC dimension of 2-d linear functions is precisely 3
- To show this, we will need to do some reasoning to show no 4 points could be shattered
- Typically, we show this by reasoning about different possible arrangements of the points

Case 1: all 4 points on a line, label outer 2 points positive, inner 2 negative

Case 2: 3 points on a line, similar to case 1

Case 3: no 3 points on a line, and convex hull involves all four points, label 2 points cross the diagonal to be the same class

Case 4: no 3 points on a line, convex hull involves only 3 points. Label the interior point positive, the rest negative
More about VC-dimension

• In general, the VC-dimension for linear separators in \( n \)-dimensional space is \( n+1 \).

• A good heuristic is that the VC-dimension is equal to the number of tunable parameters in the model (unless the parameters are redundant).

• For finite space \( H \), we have \( VC(H) \leq \log_2 |H| \)

Proof:

To shatter \( m \) points, there needs to be at least \( 2^m \) hypothesis in \( H \). Thus \( |H| \geq 2^m \Rightarrow m \leq \log_2 |H| \)
Bounds for *Consistent* Hypotheses

- The following bound is analogous to the Blumer bound.
- If \( h \in H \) is an hypothesis consistent with a training set of size \( m \), and \( \text{VC}(H) = d \), then with probability at least \( 1 - \delta \), \( h \) will have an error rate less than \( \epsilon \) if
  \[
  m \geq \frac{1}{\epsilon} \left( 4 \log_2 \frac{2}{\delta} + 8d \log_2 \frac{13}{\epsilon} \right)
  \]

- Compared to previous bound based on \( |H| \):
  \[
  m \geq \frac{1}{\epsilon} \left( \ln \frac{1}{\delta} + \ln |H| \right)
  \]
  since \( \text{VC}(H) \leq \log_2(H) \), VC dimension generally gives a tighter upper bound on the number of examples required for PAC learning.
Bound for *Inconsistent* Hypotheses

- **Theorem.** Suppose $\text{VC}(H)=d$ and a learning algorithm finds $h \in H$ with error rate $\epsilon_T$ on a training set of size $m$. Then with probability $1 - \delta$, the true error rate $\epsilon$ of $h$ is

$$
\epsilon \leq \epsilon_T + \sqrt{d \left( \log \frac{2m}{d} + 1 \right) + \log \frac{4}{\delta}}
$$

- **Empirical Risk Minimization Principle**
  – If you have a fixed hypothesis space $H$, then your learning algorithm should minimize $\epsilon_T$: the error on the training data. ($\epsilon_T$ is also called the “empirical risk”
Structural Risk Minimization

- Vapnik’s structural risk minimization (SRM) fits a nested sequence of models of increasing VC dimension and choose the model with the smallest value of the upper bound

\[ \epsilon \leq \epsilon_T^* + \sqrt{\frac{d \left( \log \frac{2m}{d} + 1 \right) + \log \frac{4}{\delta}}{m}} \]

Minimum training error that can be achieve by \( h \in H \)
Structural risk minimization: example

linear

$2^{nd}$ order polynomial

$4^{th}$ order polynomial

$8^{th}$ order polynomial
Example (cont.)

- $m = 50$, probability bound parameter $\delta = 0.05$

<table>
<thead>
<tr>
<th>$H$</th>
<th>$VC$</th>
<th>$\varepsilon_T$</th>
<th>Error bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1$^{st}$ order</td>
<td>3</td>
<td>0.06</td>
<td>0.5501</td>
</tr>
<tr>
<td>2$^{nd}$ order</td>
<td>6</td>
<td>0.06</td>
<td>0.6999</td>
</tr>
<tr>
<td>4$^{th}$ order</td>
<td>15</td>
<td>0.04</td>
<td>0.9494</td>
</tr>
<tr>
<td>8$^{th}$ order</td>
<td>45</td>
<td>0.02</td>
<td>1.2849</td>
</tr>
</tbody>
</table>

Recall that the VC dimension of a linear separator in an $n$-dimensional space is $n+1$. 
Computational Learning Theory: Summary

- Within the PAC learning setting, we can bound the probability that learner will output hypothesis with given error
  - For any consistent learner (where the target concept is in the hypothesis space)
  - For any “best fit” learner (target concept not in the hypothesis space)
  - For finite hypothesis space based on the size of |H|
  - For infinite hypothesis space based on VC dimension of |H|

- VC dimension measures the complexity of H

- Structure Risk Minimization (SRM) provides one principal for model selection. But
  - The generalization bound is highly conservative (loose)
  - Other model selection methods (cross-validation etc.) often work better in practice