Learning as Prediction

Definition: A particular instance of a learning problem is described by a probability distribution $P(X,Y)$.

Definition: A sample $S = ((\tilde{x}_1, y_1), ..., (\tilde{x}_n, y_n))$ is independently identically distributed (i.i.d.) according to $P(X,Y)$ if

$$P(S = ((\tilde{x}_1, y_1), ..., (\tilde{x}_n, y_n))) = \prod_{i=1}^{n} P(X = \tilde{x}_i, Y = y_i)$$
Sample Error and Generalization Error

**Definition:** The error on sample $S$ $\text{Err}_S(h)$ of a hypothesis $h$ is $\text{Err}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \Delta(h(\bar{x}_i), y_i)$.

**Definition:** $\Delta(a, b)$ is the 0/1-loss function

$$\Delta(a, b) = \begin{cases} 0 & \text{if } a == b \\ 1 & \text{else} \end{cases}$$

**Definition:** The prediction/generalization/true error $\text{Err}_P(h)$ of a hypothesis $h$ for a learning task $P(X, Y)$ is

$$\text{Err}_P(h) = \sum_{\bar{x} \in X, y \in Y} \Delta(h(\bar{x}), y) P(X = \bar{x}, Y = y).$$
Learning as Prediction

Overview

• Goal: Find $h$ with small prediction error $\text{Err}_P(h)$ over $P(X,Y)$.
• Strategy: Find (any?) $h$ with small error $\text{Err}_{S_{\text{train}}}(h)$ on training sample $S_{\text{train}}$.

Drawn i.i.d.

Real-world Process $P(X,Y)$

Train Sample $S_{\text{train}}(x_1,y_1), \ldots, (x_n,y_n)$

Test Sample $S_{\text{test}}(x_{n+1},y_{n+1}), \ldots$

Learner

$h$

• Training Error: Error $\text{Err}_{S_{\text{train}}}(h)$ on training sample.
• Test Error: Error $\text{Err}_{S_{\text{test}}}(h)$ on test sample is an estimate of $\text{Err}_P(h)$.
Overfitting

Note: Accuracy = 1.0 - Error
Decision Tree Example: revisited

<table>
<thead>
<tr>
<th>COP</th>
<th>$A^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 = (c, y, c)$</td>
<td>$y_1 = +1$</td>
</tr>
<tr>
<td>$x_2 = (c, n, u)$</td>
<td>$y_2 = -1$</td>
</tr>
<tr>
<td>$x_3 = (c, y, u)$</td>
<td>$y_3 = +1$</td>
</tr>
<tr>
<td>$x_4 = (c, n, c)$</td>
<td>$y_4 = +1$</td>
</tr>
<tr>
<td>$x_5 = (p, y, c)$</td>
<td>$y_5 = -1$</td>
</tr>
<tr>
<td>$x_6 = (g, y, c)$</td>
<td>$y_6 = -1$</td>
</tr>
<tr>
<td>$x_7 = (c, y, c)$</td>
<td>$y_7 = +1$</td>
</tr>
<tr>
<td>$x_8 = (c, y, u)$</td>
<td>$y_8 = +1$</td>
</tr>
<tr>
<td>$x_9 = (p, y, c)$</td>
<td>$y_9 = -1$</td>
</tr>
<tr>
<td>$x_{10} = (c, y, c)$</td>
<td>$y_{10} = +1$</td>
</tr>
</tbody>
</table>
Reduced-Error Pruning

Accuracy vs. Size of tree (number of nodes)

On training data
On test data
On test data (during pruning)
Text Classification Example

Results

• Unpruned Tree:
  – Size: 437 nodes  Training Error: 0.0%  Test Error: 11.0%

• Early Stopping Tree:
  – Size: 299 nodes  Training Error: 2.6%  Test Error: 9.8%

• Post-Pruned Tree:
  – Size: 167 nodes  Training Error: 4.0%  Test Error: 10.8%

• Rule Post-Pruning:
  – Size: 164 tests  Training Error: 3.1%  Test Error: 10.3%
    – Examples of rules
      • IF vs = 1 THEN -  [99.4%]
      • IF vs = 0 & export = 0 & takeover = 1 THEN +  [93.6%]