Supervised Learning

Functional Overview – Learning Stage

Examples

- 
- 
+ + - 
+ - + 
- 

Model/Theory

Age, … income, \{Default | NoDefault\}
Case A. 30, …, $110K, Default
Case B. 50, …, $110K, NoDefault
Case C. 45, …, $90K, NoDefault
Case A. 32, …, $105K, Default
Case B. 49, …, $82K, NoDefault
Case C. 29, …, $50K, NoDefault
Supervised Learning
Functional Overview – Application Stage

Unlabeled Examples

Age, ... income, {Default | NoDefault}
Case zx. 29, ..., $113K, ?
Case zy. 42, ..., $81K, ?
Case zz. 41, ..., $92K, ?

Predictions

zx, Default
zy. NoDefault
zz. NoDefault
Supervised Learning Terminology

- Attribute, instances
- Training set, test set
- Training set accuracy
- Test set accuracy
- Confusion Matrix
- False positive
- False negative
- Cost sensitive learning
Decision Tree For Playing Tennis

ROOT NODE

BRANCH

INTERNAL NODE

LEAF NODE

Disjunction of conjunctions
How and When To Use Decision Trees

• How
  – Classification
    • Gain insights into why customers default on loans
  – Prediction
    • Screen customer loan applications
  – Feature/column/attribute selection
  – To explain other learning techniques

• When
  – Instances are attribute value pairs
  – Discrete target function
  – Noisy training data or missing values
Another Perspective of a Decision Tree Model

Age, … income
Case A. 30, …, $110K, Default
Case B. 50, …, $110K, NoDefault
Case C. 45, …, $90K, NoDefault
Case A. 32, …, $105K, Default
Case B. 49, …, $82K, NoDefault
Case C. 29, …, $50K, NoDefault
Top-Down Tree Induction

Main loop:
1. $A \leftarrow$ the “best” decision attribute for next node
2. Assign $A$ as decision attribute for node
3. For each value of $A$, create new descendant of node
4. Sort training examples to leaf nodes
5. If training examples perfectly classified, Then STOP, Else iterate over new leaf nodes

Which attribute is best?

![Decision Tree Diagram]

CSI535 – Machine Learning Introduction
Which Column and Split Point?

- Multitude of techniques:
  - Entropy/Information gain
  - Chi square test (CHAID)
    - Test of independence
  - GINI index
Information Gain

\[ Gain(S, A) = \text{expected reduction in entropy due to sorting on } A \]

\[ Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \]
Entropy

$Entropy(S) = \text{expected number of bits needed to encode class } (\oplus \text{ or } \ominus) \text{ of randomly drawn member of } S \text{ (under the optimal, shortest-length code)}$

Why?

Information theory: optimal length code assigns $-\log_2 p$ bits to message having probability $p$.

So, expected number of bits to encode $\oplus$ or $\ominus$ of random member of $S$:

$$p_\oplus(-\log_2 p_\oplus) + p_\ominus(-\log_2 p_\ominus)$$

$$Entropy(S) \equiv -p_\oplus \log_2 p_\oplus - p_\ominus \log_2 p_\ominus$$
## Data Set

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Choosing the Next Attribute - 1

Which attribute is the best classifier?

\[ S: [9+5-] \]
\[ E = 0.940 \]

**Humidity**

- **High**
  - \([3+4-]\)
  - \(E = 0.985\)

- **Normal**
  - \([6+1-]\)
  - \(E = 0.592\)

\[ \text{Gain}(S, \text{Humidity}) = 0.940 - \frac{7}{14} \times 0.985 - \frac{7}{14} \times 0.592 \]
\[ = 0.151 \]

\[ S: [9+5-] \]
\[ E = 0.940 \]

**Wind**

- **Weak**
  - \([6+2-]\)
  - \(E = 0.811\)

- **Strong**
  - \([3+3-]\)
  - \(E = 1.00\)

\[ \text{Gain}(S, \text{Wind}) = 0.940 - \frac{8}{14} \times 0.811 - \frac{6}{14} \times 1.0 \]
\[ = 0.048 \]
Choosing the Next Attribute - 2

Which attribute should be tested here?

\[ S_{\text{Sunny}} = \{D1, D2, D8, D9, D11\} \]

\[
\begin{align*}
\text{Gain (} S_{\text{Sunny}}, \text{Humidity}) &= 0.970 - (3/5) 0.0 - (2/5) 0.0 = 0.970 \\
\text{Gain (} S_{\text{Sunny}}, \text{Temperature}) &= 0.970 - (2/5) 0.0 - (2/5) 1.0 - (1/5) 0.0 = 0.570 \\
\text{Gain (} S_{\text{Sunny}}, \text{Wind}) &= 0.970 - (2/5) 1.0 - (3/5) 0.918 = 0.019
\end{align*}
\]
Representational and Search Bias

Note $H$ is the power set of instances $X$

→ Unbiased?

Not really...

- Preference for short trees, and for those with high information gain attributes near the root
- Bias is a *preference* for some hypotheses, rather than a *restriction* of hypothesis space $H$
- Occam’s razor: prefer the shortest hypothesis that fits the data
Occam’s Razor

• 14th Century Franciscan friar; William of Occam.
• The principle states that "Entities should not be multiplied unnecessarily."
• People often reinvented Occam's Razor
  – Newton - "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."
• To most scientist the razor is:
  – "when you have two competing theories which make exactly the same predictions, the one that is simpler is the better."