Decision Tree Learning

[read Chapter 3]
[recommended exercises 3.1, 3.4]

- Decision tree representation
- ID3 learning algorithm
- Entropy, Information gain
- Overfitting

Decision Tree for PayDividend

- Outlook
  - Strong
  - Moderate
  - Weak

- Market
  - Bear
  - Bull

- Earnings
  - Low
  - High

When to Consider Decision Trees

- Instances describable by attribute-value pairs
- Target function is discrete valued
- Disjunctive hypothesis may be required
- Possibly noisy training data

Examples:
- Credit risk analysis
- Stock screening
- Pending threshold events (dividends, stock split, default)
Top-Down Induction of Decision Trees

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]

Entrophy

$Entropy(S) = \text{expected number of bits needed to encode class (\(
\oplus\) or \(
\otimes\) 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 \(
\otimes\) of random member of $S$:

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

Entropy(S) \equiv -p_\oplus \log_2 p_\oplus - p_\otimes \log_2 p_\otimes$

\begin{align*}
\text{Entropy} = & \quad \text{expected reduction in entropy due to sorting on } A \\
& \quad \text{Gain}(S, A) = \sum_{a \in \text{Value}(A)} \frac{|S_a|}{|S|} \text{Entropy}(S_a)
\end{align*}

![Information Gain Diagram]
### Training Examples

<table>
<thead>
<tr>
<th>Stock</th>
<th>Outlook</th>
<th>Price</th>
<th>Market</th>
<th>Earnings</th>
<th>PayDividend</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Strong</td>
<td>Down</td>
<td>Bear</td>
<td>High</td>
<td>No</td>
</tr>
<tr>
<td>S2</td>
<td>Strong</td>
<td>Down</td>
<td>Bear</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>S3</td>
<td>Moderate</td>
<td>Down</td>
<td>Bear</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>S4</td>
<td>Weak</td>
<td>Same</td>
<td>Bear</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>S5</td>
<td>Weak</td>
<td>Up</td>
<td>Bull</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>S6</td>
<td>Weak</td>
<td>Up</td>
<td>Bull</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>S7</td>
<td>Moderate</td>
<td>Up</td>
<td>Bull</td>
<td>Low</td>
<td>Yes</td>
</tr>
<tr>
<td>S8</td>
<td>Strong</td>
<td>Same</td>
<td>Bear</td>
<td>High</td>
<td>No</td>
</tr>
<tr>
<td>S9</td>
<td>Strong</td>
<td>Up</td>
<td>Bull</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>S10</td>
<td>Weak</td>
<td>Same</td>
<td>Bull</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>S11</td>
<td>Strong</td>
<td>Same</td>
<td>Bull</td>
<td>Low</td>
<td>Yes</td>
</tr>
<tr>
<td>S12</td>
<td>Moderate</td>
<td>Same</td>
<td>Bear</td>
<td>Low</td>
<td>Yes</td>
</tr>
<tr>
<td>S13</td>
<td>Moderate</td>
<td>Down</td>
<td>Bull</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>S14</td>
<td>Weak</td>
<td>Same</td>
<td>Bear</td>
<td>Low</td>
<td>No</td>
</tr>
</tbody>
</table>

### Selecting the Next Attribute

#### Which attribute is the best classifier?

- **Market**
  - Gain (S, Market) = $0.940 - (7/14) \cdot 0.985 - (7/14) \cdot 0.592 = 0.048$

- **Earnings**
  - Gain (S, Earnings) = $0.940 - (8/14) \cdot 0.811 - (6/14) \cdot 1.00 = 0.151$

### Hypothesis Space Search by ID3

**Outlook**
- **Strong**
  - Gain (S, Market) = $0.970 - (2/5) \cdot 0.00 - (2/5) \cdot 1.00 = 0.970$
  - Gain (S, Price) = $0.970 - (25) \cdot 0.00 - (25) \cdot 1.00 = 0.570$
  - Gain (S, Earnings) = $0.970 - (25) \cdot 0.10 - (3/5) \cdot 0.918 = 0.019$

- **Moderate**
  - Gain (S, Market) = $0.940 - (7/14) \cdot 0.985 - (7/14) \cdot 0.592 = 0.048$
  - Gain (S, Price) = $0.940 - (8/14) \cdot 0.811 - (6/14) \cdot 1.00 = 0.151$
  - Gain (S, Earnings) = $0.940 - (8/14) \cdot 0.811 - (6/14) \cdot 1.00 = 0.151$

- **Weak**
  - Gain (S, Market) = $0.940 - (7/14) \cdot 0.985 - (7/14) \cdot 0.592 = 0.048$
  - Gain (S, Price) = $0.940 - (8/14) \cdot 0.811 - (6/14) \cdot 1.00 = 0.151$
  - Gain (S, Earnings) = $0.940 - (8/14) \cdot 0.811 - (6/14) \cdot 1.00 = 0.151$
Hypothesis Space Search by ID3

- Hypothesis space is complete!
  - Target function surely in there...
- Outputs a single hypothesis (which one?)
  - Can’t play 20 questions...
- No back tracking
  - Local minima...
- Statistically-based search choices
  - Robust to noisy data...
- Inductive bias: approx “prefer shortest tree”

Inductive Bias in ID3

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

Why prefer short hypotheses?
Argument in favor:
- Fewer short hyps. than long hyps.
  $\rightarrow$ a short hyp that fits data unlikely to be coincidence
  $\rightarrow$ a long hyp that fits data might be coincidence

Argument opposed:
- There are many ways to define small sets of hyps
  - e.g., all trees with a prime number of nodes that use attributes beginning with “Z”
- What’s so special about small sets based on size of hypothesis??

Overfitting in Decision Trees

Consider adding noisy training example #15:
*Strong, Down, Bull, Low, PayDividend = No*

What effect on earlier tree?

```
Outlook
  Strong  Moderate  Weak
Market
  Bear  Bull
No  Yes
  Low  High
Yes
```

**Overfitting**

Consider error of hypothesis \( h \) over

- training data: \( \text{error}_{\text{train}}(h) \)
- entire distribution \( D \) of data: \( \text{error}_D(h) \)

Hypothesis \( h \in H \) **overfits** training data if there is an alternative hypothesis \( h' \in H \) such that

\[
\text{error}_{\text{train}}(h) < \text{error}_{\text{train}}(h')
\]

and

\[
\text{error}_D(h) > \text{error}_D(h')
\]

**Avoiding Overfitting**

How can we avoid overfitting?

- stop growing when data split not statistically significant
- grow full tree, then post-prune

How to select “best” tree:

- Measure performance over training data
- Measure performance over separate validation data set
- MDL: minimize
  \[
  \text{size}(\text{tree}) + \text{size}(\text{miscalifications(tree)})
  \]

**Reduced-Error Pruning**

Split data into \textit{training} and \textit{validation} set

Do until further pruning is harmful:

1. Evaluate impact on \textit{validation} set of pruning each possible node (plus those below it)
2. Greedily remove the one that most improves \textit{validation} set accuracy

- produces smallest version of most accurate subtree
- What if data is limited?
Effect of Reduced-Error Pruning

Rule Post-Pruning

1. Convert tree to equivalent set of rules
2. Prune each rule independently of others
3. Sort final rules into desired sequence for use

Perhaps most frequently used method (e.g., C4.5)

Converting A Tree to Rules

IF \((Outlook = Strong) \land (Market = Bear)\)
THEN \(PayDividend = No\)

IF \((Outlook = Strong) \land (Market = Bull)\)
THEN \(PayDividend = Yes\)

\[\ldots\]
Continuous Valued Attributes

Create a discrete attribute to test continuous

- **Price** = +23%
- **(Price > +15%)** = t, f

<table>
<thead>
<tr>
<th>Price</th>
<th>-23%</th>
<th>-12%</th>
<th>+3%</th>
<th>+15%</th>
<th>+15%</th>
<th>+22%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PayDividend</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Attributes with Many Values

Problem:
- If attribute has many values, **Gain** will select it
- Imagine using **Date = Jun, 3, 1996** as attribute

One approach: use **GainRatio** instead

\[
GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}
\]

\[
SplitInformation(S, A) \equiv - \sum_{i=1}^{v} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}
\]

where \( S_i \) is subset of \( S \) for which \( A \) has value \( v_i \)

Attributes with Costs

Consider
- medical diagnosis, **BloodTest** has cost $150
- finance, some data cost money, other cost time

How to learn a consistent tree with low expected cost?
One approach: replace gain by
- Tan and Schlimmer (1990)

\[
\frac{Gain^2(S, A)}{Cost(A)}
\]

- Nunez (1988)

\[
\frac{2^{Gain(S, A)} - 1}{(Cost(A) + 1)^w}
\]

where \( w \in [0, 1] \) determines importance of cost

Unknown Attribute Values

What if some examples missing values of \( A \)?
Use training example anyway, sort through tree
- If node \( n \) tests \( A \), assign most common value of \( A \) among other examples sorted to node \( n \)
- assign most common value of \( A \) among other examples with same target value
- assign probability \( p_i \) to each possible value \( v_i \) of \( A \)
  - assign fraction \( p_i \) of example to each descendant in tree
Classify new examples in same fashion