Decision Tree Example: BigTip

<table>
<thead>
<tr>
<th>Food</th>
<th>Chat</th>
<th>Speedy</th>
<th>Price</th>
<th>Bar</th>
<th>BigTip</th>
</tr>
</thead>
<tbody>
<tr>
<td>great</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>great</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>mediocre</td>
<td>yes</td>
<td>no</td>
<td>high</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>great</td>
<td>yes</td>
<td>yes</td>
<td>adequate</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

Top-Down Induction of DT (simplified)
Training Data: \( D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \)

\[ \text{TIDIDF}(D, c, def) = \]

- IF (all examples in \( D \) have same class \( c \))
  - Return leaf with class \( c \) (or class \( c, \text{def} \), if \( D = \emptyset \))
- ELSE IF (no attributes left to test)
  - Return leaf with class \( c \) of majority in \( D \)
- ELSE
  - Pick \( A \) as the “best” decision attribute for next node
  - FOR each value \( v_i \) of \( A \) create a new descendent of node
  - Subtree \( t_i \) for \( v_i \) is \( \text{TIDIDT}(D, c, def) \)
  - RETURN tree with \( A \) as root and \( t_i \) as subtrees

Example: Text Classification

- Task: Learn rule that classifies Reuters Business News
  - Class +: “Corporate Acquisitions”
  - Class -: Other articles
  - 2000 training instances
- Representation:
  - Boolean attributes, indicating presence of a keyword in article
  - 9947 such keywords (more accurately, word “stems”)

Example: TDIDT

\[ \text{TIDIDT}(D, c, def) = \]

- IF (all examples in \( D \) have same class \( c \))
  - Return leaf with class \( c \) (or \( c, \text{def} \), if \( D = \emptyset \))
- ELSE IF (no attributes left to test)
  - Return leaf with class \( c \) of majority in \( D \)
- ELSE
  - Pick \( A \) as the “best” decision attribute for next node
  - FOR each value \( v_i \) of \( A \) create a new descendent of node
  - Subtree \( t_i \) for \( v_i \) is \( \text{TIDIDT}(D, c, def) \)
  - RETURN tree with \( A \) as root and \( t_i \) as subtrees

Picking the Best Attribute to Split

- Ockham’s Razor:
  - All other things being equal, choose the simplest explanation
- Decision Tree Induction:
  - Find the smallest tree that classifies the training data correctly
- Problem:
  - Finding the smallest tree is computationally hard
- Approach:
  - Use heuristic search (greedy search)
Which Attribute is “Best”?

- **Heuristics**
  - Pick split that decreases training error the most
  - Pick split that maximizes information (Information Gain)
  - Other statistical tests

**Information Gain**
- Idea: Measure how much information an attribute conveys
- Entropy: Number of bits to transmit one label (~disorder)
  \[
  \text{Entropy}(D) = -p \log_2 p - n \log_2 n
  \]
- Information Gain: Reduction in entropy, if attribute value known
  \[
  \text{Gain}(D, A) = \text{Entropy}(D) - \sum_{v \in \text{values}(A)} \frac{|D_v|}{|D|} \text{Entropy}(D_v)
  \]

Decision Tree for “Corporate Acq.”

- \(v_s = 1: -
- \| v_s = 0:
  - export = 1: ...
  - export = 0:
    - rate = 1: ...
    - stake = 1: +
    - stake = 0: ...
    - debenture = 1: +
    - debenture = 0: ...
    - takeover = 1: +
    - takeover = 0:
      - file = 0: -
      - file = 1: ...
      - share = 1: +
      - share = 0: -
    ...
    and many more

Total size of tree:
- 299 nodes

Note: word stems expanded for improved readability.

How Expressive are Decision Trees?

- **What functions** \(h: X \to Y\) **can a decision tree represent?**
  - Assume that \(X\) is finite (only finite number of instances)
  - Decision trees can represent any function over a finite instance space \(X\).
  - What if \(X\) is not finite (e.g. integer-valued attributes)?
  - What if \(X\) is not discrete (e.g. real-valued attributes)?
  - What if the data contains noise?
    - In the most extreme case, examples can have the same attribute values, but different labels.

TDIDT Extensions

- **Numerical (continuous) attributes**
  - Use > and < in attribute tests
- **Finite attributes with many values**
  - Example:
    - Target concept is “brakes defect”
    - Instances: all cars in the US
    - Attributes: Manufacturer (3 values), VIN (100,000,000 values)
    - Which attribute will Information Gain select? \(\text{GainRatio}\)
- **Numerical (continuous) target attribute (regression)**
  - E.g. pick attribute test so that target values become more similar
  - E.g. predict mean value of examples in each leaf
- **Early stopping and Pruning**