Induction of Decision Trees

Blaž Zupan and Ivan Bratko

magix.fri.uni-lj.si/predavanja UISP
An Example Data Set and Decision Tree

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Outlook</td>
<td>Company</td>
</tr>
<tr>
<td>1</td>
<td>sunny</td>
<td>big</td>
</tr>
<tr>
<td>2</td>
<td>sunny</td>
<td>med</td>
</tr>
<tr>
<td>3</td>
<td>sunny</td>
<td>med</td>
</tr>
<tr>
<td>4</td>
<td>sunny</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>sunny</td>
<td>big</td>
</tr>
<tr>
<td>6</td>
<td>rainy</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>rainy</td>
<td>med</td>
</tr>
<tr>
<td>8</td>
<td>rainy</td>
<td>big</td>
</tr>
<tr>
<td>9</td>
<td>rainy</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>rainy</td>
<td>med</td>
</tr>
</tbody>
</table>

Decision Tree:

- **outlook**
  - sunny
    - company
      - med
        - yes
      - big
        - yes
    - rainy
      - no
        - sailboat
          - small
            - no
          - big
            - yes
      - big
        - no
## Classification

### Table

<table>
<thead>
<tr>
<th>#</th>
<th>Outlook</th>
<th>Company</th>
<th>Sailboat</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sunny</td>
<td>no</td>
<td>big</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>rainy</td>
<td>big</td>
<td>small</td>
<td>?</td>
</tr>
</tbody>
</table>

### Decision Tree

```
            outlook
               /    
          sunny  rainy
            /      /
       yes  company
            /      /
      no    med    big
         /  
    no   sailboat
          /  
       yes  no
          /  
    yes  no
```
Induction of Decision Trees

• Data Set (Learning Set)
  - Each example = Attributes + Class

• Induced description = Decision tree

• TDIDT
  - Top Down Induction of Decision Trees

• Recursive Partitioning
Some TDIDT Systems

• ID3 (Quinlan 79)
• CART (Brieman et al. 84)
• C4.5 (Quinlan 93)
• See5 (Quinlan 97)
• ...
• Orange (Demšar, Zupan 98-03)
Breast Cancer Recurrence

Tree induced by Assistant Professional

Interesting: Accuracy of this tree compared to medical specialists
Prostate cancer recurrence

Secondary Gleason Grade

PSA Level

Stage

Primary Gleason Grade

No

Yes

≤14.9

>14.9

T1c,T2a,
T2b,T2c

T1ab,T3
TDIDT Algorithm

• Also known as ID3 (Quinlan)

• To construct decision tree $T$ from learning set $S$:
  
  - **If** all examples in $S$ belong to some class $C$ **Then** make leaf labeled $C$
  
  - **Otherwise**
    
    • select the “most informative” attribute $A$
    • partition $S$ according to $A$’s values
    • recursively construct subtrees $T_1$, $T_2$, ..., for the subsets of $S$
TDIDT Algorithm

- Resulting tree $T$ is:

```
    A
  /   \
 v1   v2
 / \   / \  \
 T1 T2 Tn
```

- Attribute $A$
- $A$'s values
- Subtrees
## Another Example

<table>
<thead>
<tr>
<th>#</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>no</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>yes</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>no</td>
<td>P</td>
</tr>
<tr>
<td>4</td>
<td>rainy</td>
<td>moderate</td>
<td>high</td>
<td>no</td>
<td>P</td>
</tr>
<tr>
<td>5</td>
<td>rainy</td>
<td>cold</td>
<td>normal</td>
<td>no</td>
<td>P</td>
</tr>
<tr>
<td>6</td>
<td>rainy</td>
<td>cold</td>
<td>normal</td>
<td>yes</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>overcast</td>
<td>cold</td>
<td>normal</td>
<td>yes</td>
<td>P</td>
</tr>
<tr>
<td>8</td>
<td>sunny</td>
<td>moderate</td>
<td>high</td>
<td>no</td>
<td>N</td>
</tr>
<tr>
<td>9</td>
<td>sunny</td>
<td>cold</td>
<td>normal</td>
<td>no</td>
<td>P</td>
</tr>
<tr>
<td>10</td>
<td>rainy</td>
<td>moderate</td>
<td>normal</td>
<td>no</td>
<td>P</td>
</tr>
<tr>
<td>11</td>
<td>sunny</td>
<td>moderate</td>
<td>normal</td>
<td>yes</td>
<td>P</td>
</tr>
<tr>
<td>12</td>
<td>overcast</td>
<td>moderate</td>
<td>high</td>
<td>yes</td>
<td>P</td>
</tr>
<tr>
<td>13</td>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>no</td>
<td>P</td>
</tr>
<tr>
<td>14</td>
<td>rainy</td>
<td>moderate</td>
<td>high</td>
<td>yes</td>
<td>N</td>
</tr>
</tbody>
</table>
Attribute Selection Criteria

- **Main principle**
  - Select attribute which partitions the learning set into subsets as “pure” as possible

- **Various measures of purity**
  - Information-theoretic (entropy)
  - Gini index
  - $X^2$ (entropy)
  - ReliefF
  - ...
**Information-Theoretic Approach**

- To classify an object, a certain information is needed
  - $I$, information

- After we have learned the value of attribute $A$, we only need some remaining amount of information to classify the object
  - $I_{res}$, residual information

- **Gain**
  - $\text{Gain}(A) = I - I_{res}(A)$

- The most 'informative' attribute is the one that minimizes $I_{res}$, i.e., maximizes $\text{Gain}$
The average amount of information $I$ needed to classify an object is given by the entropy measure

$$I = - \sum_c p(c) \log_2 p(c)$$

For a two-class problem:
Residual Information

- After applying attribute $A$, $S$ is partitioned into subsets according to values $v$ of $A$
- $I_{res}$ is equal to weighted sum of the amounts of information for the subsets

$$I_{res} = - \sum_v p(v) \sum_c p(c|v) \log_2 p(c|v)$$
# Triangles and Squares

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Color</td>
<td>Outline</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>1</td>
<td>green</td>
<td>dashed</td>
</tr>
<tr>
<td>2</td>
<td>green</td>
<td>dashed</td>
</tr>
<tr>
<td>3</td>
<td>yellow</td>
<td>dashed</td>
</tr>
<tr>
<td>4</td>
<td>red</td>
<td>dashed</td>
</tr>
<tr>
<td>5</td>
<td>red</td>
<td>solid</td>
</tr>
<tr>
<td>6</td>
<td>red</td>
<td>solid</td>
</tr>
<tr>
<td>7</td>
<td>green</td>
<td>solid</td>
</tr>
<tr>
<td>8</td>
<td>green</td>
<td>dashed</td>
</tr>
<tr>
<td>9</td>
<td>yellow</td>
<td>solid</td>
</tr>
<tr>
<td>10</td>
<td>red</td>
<td>solid</td>
</tr>
<tr>
<td>11</td>
<td>green</td>
<td>solid</td>
</tr>
<tr>
<td>12</td>
<td>yellow</td>
<td>dashed</td>
</tr>
<tr>
<td>13</td>
<td>yellow</td>
<td>solid</td>
</tr>
<tr>
<td>14</td>
<td>red</td>
<td>dashed</td>
</tr>
</tbody>
</table>
# Triangles and Squares

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Color Outline</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>green dashed no</td>
<td>triangle</td>
</tr>
<tr>
<td>2</td>
<td>green dashed yes</td>
<td>triangle</td>
</tr>
<tr>
<td>3</td>
<td>yellow dashed no</td>
<td>square</td>
</tr>
<tr>
<td>4</td>
<td>red dashed no</td>
<td>square</td>
</tr>
<tr>
<td>5</td>
<td>red solid no</td>
<td>square</td>
</tr>
<tr>
<td>6</td>
<td>red solid yes</td>
<td>triangle</td>
</tr>
<tr>
<td>7</td>
<td>green solid no</td>
<td>square</td>
</tr>
<tr>
<td>8</td>
<td>green dashed no</td>
<td>triangle</td>
</tr>
<tr>
<td>9</td>
<td>yellow solid yes</td>
<td>square</td>
</tr>
<tr>
<td>10</td>
<td>red solid no</td>
<td>square</td>
</tr>
<tr>
<td>11</td>
<td>green solid yes</td>
<td>square</td>
</tr>
<tr>
<td>12</td>
<td>yellow dashed yes</td>
<td>square</td>
</tr>
<tr>
<td>13</td>
<td>yellow solid no</td>
<td>square</td>
</tr>
<tr>
<td>14</td>
<td>red dashed yes</td>
<td>triangle</td>
</tr>
</tbody>
</table>

Data Set:
A set of classified objects
Entropy

- 5 triangles
- 9 squares
- class probabilities

\[ p(\square) = \frac{9}{14} \]
\[ p(\triangle) = \frac{5}{14} \]

- entropy

\[ I = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940 \text{ bits} \]
Entropy reduction by data set partitioning

$I(\text{red}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971 \text{ bits}$

$I(\text{green}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971 \text{ bits}$

$I(\text{yellow}) = -\frac{4}{4} \log_2 \frac{4}{4} - 0 \log_2 \frac{0}{4} = 0.0 \text{ bits}$
Entropija vrednosti atributa

\[ I = 0.940 \]

\[ I(\text{red}) = 0.971 \text{ bits} \]

\[ p(\text{red}) = \frac{5}{14} \]

\[ p(\text{green}) = \frac{5}{14} \]

\[ p(\text{yellow}) = \frac{4}{14} \]

\[ I(\text{green}) = 0.971 \text{ bits} \]

\[ I(\text{yellow}) = 0.0 \text{ bits} \]

\[ I_{res}(\text{Color}) = \sum p(v) I(v) = \frac{5}{14} \cdot 0.971 + \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0.0 = 0.694 \text{ bits} \]
Information Gain

$I = 0.940$

$I(red) = 0.971$ bits

$I(green) = 0.971$ bits

$I(yellow) = 0.0$ bits

$Gain(Color) = I - I_{res}(Color) = 0.940 - 0.694 = 0.246$ bits
Information Gain of The Attribute

• Attributes
  - Gain(Color) = 0.246
  - Gain(Outline) = 0.151
  - Gain(Dot) = 0.048

• Heuristics: attribute with the highest gain is chosen
• This heuristics is local (local minimization of impurity)
Gain(Outline) = 0.971 - 0 = 0.971 bits
Gain(Dot) = 0.971 - 0.951 = 0.020 bits
Color?

- red
- green
- yellow

Outline?

- dashed
- solid

Gain(Outline) = 0.971 - 0.951 = 0.020 bits

Gain(Dot) = 0.971 - 0 = 0.971 bits
Gini Index

- Another sensible measure of impurity (i and j are classes)

\[ Gini = \sum_{i \neq j} p(i)p(j) \]

- After applying attribute A, the resulting Gini index is

\[ Gini(A) = \sum_{v} p(v) \sum_{i \neq j} p(i|v)p(j|v) \]

- Gini can be interpreted as expected error rate
Gini Index

\[ p(\square) = \frac{9}{14} \]
\[ p(\triangle) = \frac{5}{14} \]

\[ \text{Gini} = \sum_{i \neq j} p(i)p(j) \]
\[ \text{Gini} = \frac{9}{14} \times \frac{5}{14} = 0.230 \]
Gini Index for Color

\[ Gini(\text{Color}) = \frac{5}{14} \times \left( \frac{3}{5} \times \frac{2}{5} \right) + \frac{5}{14} \times \left( \frac{2}{5} \times \frac{3}{5} \right) + \frac{4}{14} \times \left( \frac{4}{4} \times \frac{0}{4} \right) = 0.171 \]
Gain of Gini Index

\[ Gini = \frac{9}{14} \times \frac{5}{14} = 0.230 \]

\[ Gini(\text{Color}) = \frac{5}{14} \times \left( \frac{3}{5} \times \frac{2}{5} \right) + \frac{5}{14} \times \left( \frac{2}{5} \times \frac{3}{5} \right) + \frac{4}{14} \times \left( \frac{4}{4} \times \frac{0}{4} \right) = 0.171 \]

\[ GiniGain(\text{Color}) = 0.230 - 0.171 = 0.058 \]
Three Impurity Measures

<table>
<thead>
<tr>
<th>A</th>
<th>Gain(A)</th>
<th>GainRatio(A)</th>
<th>GiniGain(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>0.247</td>
<td>0.156</td>
<td>0.058</td>
</tr>
<tr>
<td>Outline</td>
<td>0.152</td>
<td>0.152</td>
<td>0.046</td>
</tr>
<tr>
<td>Dot</td>
<td>0.048</td>
<td>0.049</td>
<td>0.015</td>
</tr>
</tbody>
</table>

- These impurity measures assess the effect of a single attribute.
- Criterion “most informative” that they define is local (and “myopic”).
- It does not reliably predict the effect of several attributes applied jointly.