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>
<tr>
<td>#</td>
<td>Attribute</td>
<td>Class</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>Outlook</td>
<td>Company</td>
</tr>
<tr>
<td>1</td>
<td>sunny</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>rainy</td>
<td>big</td>
</tr>
</tbody>
</table>

Classification

```
outlook
  sunny
    yes
  rainy
    company
      no
        med
          small
            yes
            no
      big
        yes
```

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)
- Assistant (Cestnik et al. 87)
- C4.5 (Quinlan 93)
- See5 (Quinlan 97)
- ...
- Orange (Demšar, Zupan 98-03)
The worst pH value at ICU

The worst active partial thromboplastin time

PH_ICU and APPT_WORST are exactly the two factors (theoretically) advocated to be the most important ones in the study by Rotondo et al., 1997.
Tree induced by Assistant Professional

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

Secondary Gleason Grade

1,2

3

4

5

Yes

PSA Level

≤14.9

>14.9

No

Primary Gleason Grade

2,3

4

No

Yes

Stage

T1c,T2a, T2b,T2c

T1ab,T3

No

Yes
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 T1, T2, ..., for the subsets of S
TDIDT Algorithm

- Resulting tree $T$ is:

```
     A
    /|
   v1 v2 vn
  /|
 T1 T2 Tn
```

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

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Outlook</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Humidity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Windy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Play</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>sunny</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>sunny</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>overcast</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>rainy</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>rainy</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>rainy</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>overcast</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>sunny</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>sunny</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>rainy</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>sunny</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>overcast</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>overcast</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>rainy</td>
<td></td>
</tr>
</tbody>
</table>

- Outlook: sunny, overcast, rainy
- Temperature: hot, moderate, cold
- Humidity: high, normal
- Windy: no, yes
- Play: N, P
Complicated Tree

- **Temperature**
  - cold
  - moderate
  - hot

- **Outlook**
  - sunny
  - overcast
  - rainy

- **Windy**
  - yes
  - no

- **Humidity**
  - high
  - normal

- **Predicted Value**
  - P (yes)
  - N (no)
Attribute Selection Criteria

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

- **Various measures of purity**
  - Information-theoretic
  - Gini index
  - $X^2$
  - ReliefF
  - ...

- **Various improvements**
  - probability estimates
  - normalization
  - binarization, subsetting
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
  - $Gain(A) = I - I_{res}(A)$

• The most ‘informative’ attribute is the one that minimizes $I_{res}$, i.e., maximizes $Gain$
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>Color</td>
<td>Outline</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</td>
<td>Outline</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>

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

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

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

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

$I = 0.940$

$I(red) = 0.971$ bits

$p(red) = \frac{5}{14}$

$p(yellow) = \frac{4}{14}$

$p(green) = \frac{5}{14}$

$I(yellow) = 0.0$ bits

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

\[ I = 0.940 \]

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

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

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

\[ Gain(\text{Color}) = I - I_{res}(\text{Color}) = 0.940 - 0.694 = 0.246 \text{ bits} \]
Information Gain of The Attribute

- Attributes
  - \( \text{Gain(Color)} = 0.246 \)
  - \( \text{Gain(Outline)} = 0.151 \)
  - \( \text{Gain(Dot)} = 0.048 \)
- Heuristics: attribute with the highest gain is chosen
- This heuristics is local (local minimization of impurity)
Color?

Gain(Outline) = 0.971 - 0 = 0.971 bits
Gain(Dot) = 0.971 - 0.951 = 0.020 bits
$\text{Gain(Outline)} = 0.971 - 0.951 = 0.020 \text{ bits}$

$\text{Gain(Dot)} = 0.971 - 0 = 0.971 \text{ bits}$
Decision Tree

- Color
  - red
  - yellow
  - green
- Dot
  - yes: triangle
  - no: square
- Outline
  - dashed: triangle
  - solid: square
A Defect of *Ires*

- *Ires* favors attributes with many values
- Such attribute splits $S$ to many subsets, and if these are small, they will tend to be pure anyway
- One way to rectify this is through a corrected measure of information gain ratio.
Information Gain Ratio

• $I(A)$ is amount of information needed to determine the value of an attribute $A$

$$I(A) = - \sum_v p(v) \log_2(p(v))$$

• Information gain ratio

$$GainRatio(A) = \frac{Gain(A)}{I(A)} = \frac{I - I_{res}(A)}{I(A)}$$
Information Gain Ratio

$I(A) = - \sum_v p(v) \log_2(p(v))$

$I(\text{Color}) = - \frac{5}{14} \log_2 \frac{5}{14} - \frac{5}{14} \log_2 \frac{5}{14} - \frac{4}{14} \log_2 \frac{4}{14} = 1.58 \text{ bits}$

$GainRatio(\text{Color}) = \frac{Gain(\text{Color})}{I(\text{Color})} = \frac{0.940 - 0.694}{1.58} = 0.156$
Information Gain and Information Gain Ratio

| A          | \(|v(A)|\) | Gain(A) | GainRatio(A) |
|------------|------------|---------|--------------|
| Color      | 3          | 0.247   | 0.156        |
| Outline    | 2          | 0.152   | 0.152        |
| Dot        | 2          | 0.048   | 0.049        |
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

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

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

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

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

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