Pattern Recognition
Chapter 5: Decision Trees

Asst. Prof. Dr. Chumphol Bunkhumpornpat
Department of Computer Science
Faculty of Science
Chiang Mai University
Learning Objectives

• How decision trees are used to choose the course of action
• How decision trees are used for classification
• The strength and weakness of decision trees
• The splitting criterion used at the nodes
• What is the meant by induction of decision trees
• Why pruning of decision tree is sometimes neccesasy
Decision Tree

- The most popularly used data structure
- Highly Transparent
- User-Friendly Characteristics
- Internal Node: Decision
- Leaf Node: Class Label
- Link: Possible Value of Decision
# Descriptions of a Set of Animals

<table>
<thead>
<tr>
<th></th>
<th>Legs</th>
<th>Horns</th>
<th>Size</th>
<th>Colour</th>
<th>Beak</th>
<th>Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cow</td>
<td>4</td>
<td>yes</td>
<td>medium</td>
<td>white</td>
<td>no</td>
<td>moo</td>
</tr>
<tr>
<td>Crow</td>
<td>2</td>
<td>no</td>
<td>small</td>
<td>black</td>
<td>yes</td>
<td>caw</td>
</tr>
<tr>
<td>Elephant</td>
<td>4</td>
<td>no</td>
<td>large</td>
<td>black</td>
<td>no</td>
<td>trumpet</td>
</tr>
<tr>
<td>Goat</td>
<td>4</td>
<td>yes</td>
<td>small</td>
<td>brown</td>
<td>no</td>
<td>bleat</td>
</tr>
<tr>
<td>Mouse</td>
<td>4</td>
<td>no</td>
<td>small</td>
<td>black</td>
<td>no</td>
<td>squeak</td>
</tr>
<tr>
<td>Parrot</td>
<td>2</td>
<td>no</td>
<td>small</td>
<td>green</td>
<td>yes</td>
<td>squawk</td>
</tr>
<tr>
<td>Sparrow</td>
<td>2</td>
<td>no</td>
<td>small</td>
<td>brown</td>
<td>yes</td>
<td>chirp</td>
</tr>
<tr>
<td>Tiger</td>
<td>4</td>
<td>no</td>
<td>medium</td>
<td>orange</td>
<td>no</td>
<td>roar</td>
</tr>
</tbody>
</table>
Decision Tree for Classification of Animals

No. of legs

Beak?

Horns?

Colour?

Size

- Yes
  - Small
    - Crow
  - Medium
    - Sparrow
  - Large
    - Parrot

- No
  - Brown

Size

- Yes
  - Small
    - Goat
  - Medium
    - Cow
  - Large
    - Mouse

- No
  - Human
    - Black
    - Green
    - Brown

204453: Pattern Recognition
The Observations of the Decision Tree

• Class labels are associated with leaf nodes.
  – The leaf nodes are associated with animal names.

• Root to leaf represents a rule.
  – if (no. of legs = 4) and (has horns = false) and (size = small) then (mouse)

• Classification involves making a decision at every node.
  – Classification moves down the appropriate branch till we reach a leaf node.
The Observations of the Decision Tree (cont.)

• Irrelevant features do not occur in the decision tree.
  – Sound is not needed for the discrimination of the patterns.

• Numerical and categorical features can be used.
  – Colour is a categorical feature.
  – The number of legs is a numerical feature.

• The tree can be binary or non-binary.
  – A many-way decision can be converted into a number of yes/no decisions.
The Observations of the Decision Tree (cont.)

• A set of patterns is associated with each node.
  – This set is larger at the top nodes and keeps reducing in size at subsequent levels.
  – Decision 2 legs contains only birds and human.

• The rule are simple and easy to understand.
  – Conjunction of a number of Antecedents
  – Single Outcome
  – Simple Boolean Function
Axis-parallel Decision Tree

- Rectangular Classification
- Decision Boundary
- Hyper-plane
- Region
Non-Rectangular Decision Boundary
Decision Tree Performed on a Non-Rectangular Region
Construction of Decision Trees

• The most discriminative attribute is used at the earliest level in the decision tree.
• At each node, the set of examples is split up and each outcome is a new decision tree learning problem by itself.
Classification into two classes: An illustration
Classification into two classes: An illustration (cont.)

• The decision $f_1 \geq a$ divides both class 1 and class 2 so that all the patterns of each class is on one side of the boundary.

• The decision $f_2 \geq b$ divides both class 1 and class 2 so that there are two patterns on one side and two patterns on the other side.
Classification into two classes: An illustration (cont.)

• The decision $f_1 \geq a$ is a better option as it directly classifies the patterns as belonging to class 1 or class 2.

• At each node, the query which makes data to the subsequent nodes as pure as possible is chosen.
Entropy Impurity

• The entropy impurity at a node $N$ is $i(N)$ and is given by $i(N) = -\sum_j P(w_j) \log_2 P(w_j)$.

• $P(w_j)$ is the fraction of patterns at node $N$ of category $w_j$. 
Information Gain

• The attribute to be chosen should decrease the impurity as much as possible.

$$\Delta i(N) = i(N) - \sum_j P_j \times i(N_j)$$

• $j$ takes on the value for each of the outcomes of the decision made at the node.

• $\Delta i(N)$ can also be called the gain in information at the node.

• The attribute which maximises $\Delta i(N)$ is to be chosen.
Example Training Data Set for Induction of a Decision Tree

<table>
<thead>
<tr>
<th>Cook</th>
<th>Mood</th>
<th>Cuisine</th>
<th>Tasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sita</td>
<td>Bad</td>
<td>Indian</td>
<td>Yes</td>
</tr>
<tr>
<td>Sita</td>
<td>Good</td>
<td>Continental</td>
<td>Yes</td>
</tr>
<tr>
<td>Asha</td>
<td>Bad</td>
<td>Indian</td>
<td>No</td>
</tr>
<tr>
<td>Asha</td>
<td>Good</td>
<td>Indian</td>
<td>Yes</td>
</tr>
<tr>
<td>Usha</td>
<td>Bad</td>
<td>Indian</td>
<td>Yes</td>
</tr>
<tr>
<td>Usha</td>
<td>Bad</td>
<td>Continental</td>
<td>No</td>
</tr>
<tr>
<td>Asha</td>
<td>Bad</td>
<td>Continental</td>
<td>No</td>
</tr>
<tr>
<td>Asha</td>
<td>Good</td>
<td>Continental</td>
<td>Yes</td>
</tr>
<tr>
<td>Usha</td>
<td>Good</td>
<td>Indian</td>
<td>Yes</td>
</tr>
<tr>
<td>Usha</td>
<td>Good</td>
<td>Continental</td>
<td>No</td>
</tr>
<tr>
<td>Sita</td>
<td>Good</td>
<td>Indian</td>
<td>Yes</td>
</tr>
<tr>
<td>Sita</td>
<td>Bad</td>
<td>Continental</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Decision Tree Induced from Training Examples

Cook

Asha
Mood
Bad
No
Good
Yes

Sita
Yes

Usha
Cuisine
Continental
No
Indian
Yes
Over-fitting

• Whenever there is a large set of possible hypotheses, the tree can keep growing—thus becoming too specific.
• One unique path through the three for every pattern in the training set.
Pruning

- It prevents recursive splitting using irrelevant attributes.
- Irrelevant Attribute
  - Zero Information Gain
  - Sub-sets may be roughly the same proportion of each class as the original dataset.
labor.arff: unpruned
labor.arff: pruned
Logarithm Formula

- \( \log_a 1 = 0 \)
- \( \log_a a = 1 \)
- \( \log_a a^b = b \)
- \( \log_a (b/c) = \log_a b - \log_a c \)
- \( \log_a b = \frac{\log_c b}{\log_c a} \)
Reference

Cook = Sita : \( i(N_s) = -(4/4) \log(4/4) = 0.0 \)
Cook = Asha : \( i(N_a) = -(2/4) \log(2/4) - (2/4) \log(2/4) = 1.0 \)
Cook = Usha : \( i(N_u) = -(2/4) \log(2/4) - (2/4) \log(2/4) = 1.0 \)
\[ \Delta i(N) = 0.9183 - (4/12)(1.0) - (4/12)(1.0) = 0.2516 = \text{MAX} \]

Mood = Bad : \( i(N_b) = -(3/6) \log(3/6) - (3/6) \log(3/6) = 1.0 \)
Mood = Good : \( i(N_g) = -(1/6) \log(1/6) - (5/6) \log(5/6) = 0.65 \)
\[ \Delta i(N) = 0.9183 - (6/12)(0.65) - (6/12)(1.0) = 0.0933 \]

Cuisine = Indian : \( i(N_i) = -(1/6) \log(1/6) - (5/6) \log(5/6) = 0.65 \)
Cuisine = Continental : \( i(N_c) = -(3/6) \log(3/6) - (3/6) \log(3/6) = 1.0 \)
\[ \Delta i(N) = 0.9183 - (6/12)(0.65) - (6/12)(1.0) = 0.0933 \]