Topic 2.4:
Decision Trees

Decision Tree Induction

Example: to determine if a given film will be a success or not

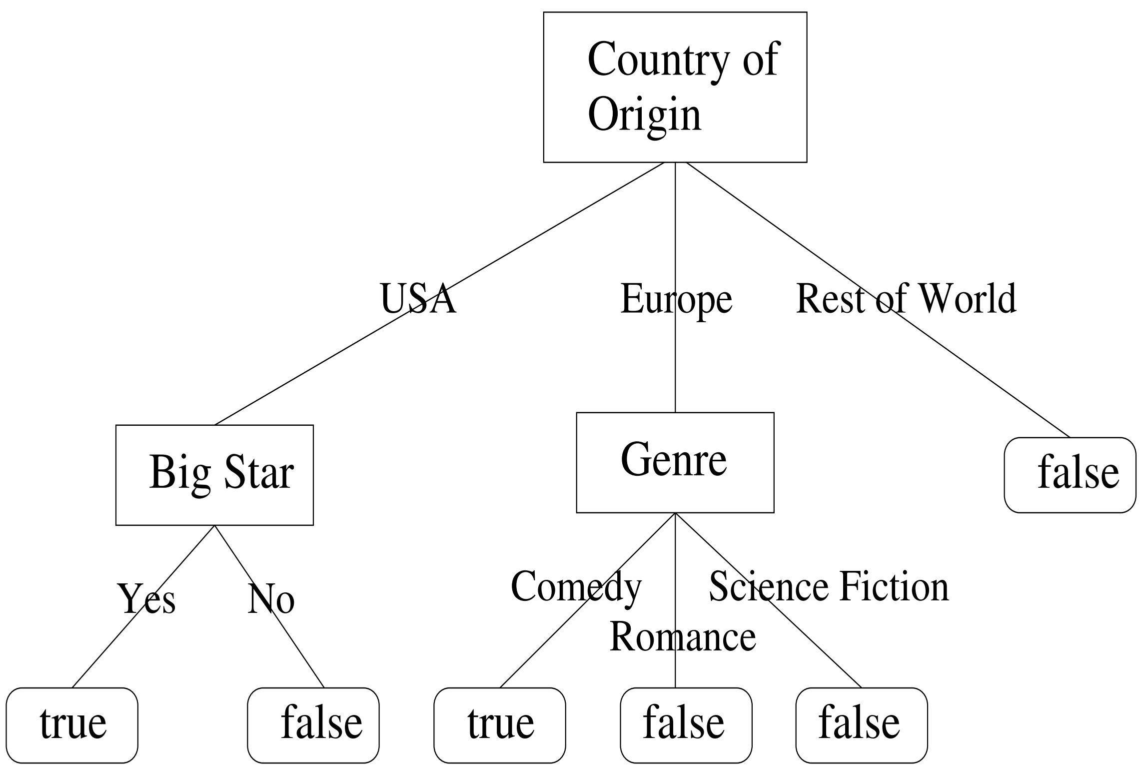

Figure 1: Decision tree example

Decision tree algorithms: ID3, C4.5/C5

Types of decision trees

- Classification tree – leaf nodes represent different discreet classes
- Regression tree – leaf nodes represent numerical values
- Model tree – leaf nodes represent multi-variate linear/nonlinear models

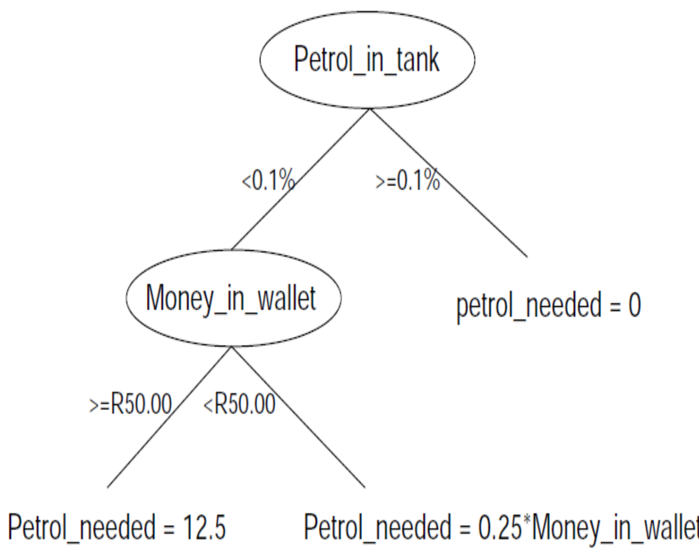

Figure 2: Model tree example

How to induce a tree? Divide and conquer

- given a set T of training cases
- classes: $\{C_1, C_2, \cdots, C_k\}$
- three possibilities:

1. T contains one or more cases of a single class $C_j \rightarrow$ decision tree for T is a leaf identifying class C_j
2. T contains no cases $\rightarrow$ leaf constructed from domain knowledge
3. T contains cases of different classes $\rightarrow$ refine T into subsets of cases that are less inhomogeneous collections of cases:
   - choose a test based on single attribute
   - to produce mutually exclusive outcomes $\{O_1, O_2, \cdots, O_n\}$
   - $T \rightarrow T_1, T_2, \cdots, T_n$
     T_i contains all cases of outcome O_i
4. Apply recursively to each subset T_i until subset represents a specific class – i.e. overfits

How to use it for classification?

Information gain

- select the feature that provides the greatest information gain
- information gain is defined as the reduction in entropy
- What is Entropy?
  - a measure from information theory
  - is the average amount of information needed to identify the class of a case
  - characterizes the (im)purity, or homogeneity, of an arbitrary collection of cases
- Define a message as:
  "case p belongs to class C_j"
- the information conveyed by a message on its probability can be measured in bits as

$$-\log_2(P_{C_j})$$

where

$$P_{C_j} = \frac{freq(C_j, S)}{|S|}$$

with S the sample set

- the entropy of a set S is defined as

$$H(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} log_2(\frac{freq(C_j, S)}{|S|})$$

- $H(S) = 0$ if all examples are positive or all are negative, i.e. perfectly homogeneous
- $H(S) = 1$ when S is perfectly inhomogeneous
- information gain of a particular feature indicates how closely that feature represents the entire target function

Suppose T is split according to test X into n subsets (branches)

- Average entropy:

$$H_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

- information gain by partitioning T in accordance with test X:

$$gain(X) = H(T) - H_X(T)$$

- Objective: maximize $gain(X)$, thus select a test with minimum $H_X(T)$

Example

Film	**Country**	**Big Star**	**Genre**	**Success**
Film 1	USA	yes	Science Fiction	true
Film 2	USA	no	Comedy	false
Film 3	USA	yes	Comedy	true
Film 4	Europe	no	Comedy	true
Film 5	Europe	yes	Science Fiction	false
Film 6	Europe	yes	Romance	false
Film 7	Other	yes	Comedy	false
Film 8	Other	no	Science Finction	false
Film 9	Europe	yes	Comedy	true
Film 10	USA	yes	Comedy	true

- $H(S) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{1}{5}\log_2\frac{1}{5} = 1$
- To calculate information gain of an attribute, calculate entropy of each attribute value:
  - $H_{Country}(USA) = -3/4\log_2 3/4 - 1/4\log_2 1/4 = 0.811$
  - $H_{Country}(Europe) = 1$
  - $H_{Country}(Other) = 0$
- $Gain(Country) = 1-(0.4\times 0.811)-(0.4\times 1)-(0.2\times 0) = 0.2756$
  - $H_{BigStar}(yes) = 0.9852$
  - $H_{BigStar}(no) = 1$
- $Gain(BigStar) = 1-(0.7\times 0.9852)-(0.3\times 1) = 0.01$
- $Gain(Genre) = 0.17$

- Country provides maximum information gain, so it is selected to split the training data

What now?

Gain ratio criterion – C4.5/C5/See5

- Gain criterion biased in favor of tests with many outcomes - what is the consequence?
- Normalize: Consider info content of message pertaining to a case that indicates outcome of the test, not the class

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} log_2(\frac{|T_i|}{|T|})$$

- split info(X) = potential information generated by dividing T into n subsets
- gain ratio = the proportion of information generated by the split that appears helpful for classification

  $$gain\ ratio(X) = gain(X)/split\ info(X)$$

  where $gain(X)$ measures information relevant to classification
- Objective: maximize $gain\ ratio(X)$

What to do with continuous attributes?

- Sort cases in T on values of attribute A: $\{v_1, v_2, \cdots, v_m\}$
- perform $m-1$ splits between v_i and v_{i+1}
- examine each split
- threshold value $= \frac{v_i + v_{i+1}}{2}$

What to do with missing values?

Overfitting

Memorization

Happens when

- there is noise in the training data
- the model has too many free parameters
- training is too long

When does a decision tree overfit?

How can overfitting in decision trees be avoided?

Note:

- A decistion tree is induced to overfit the training data
- This produces a specialized tree
- Then pruning is applied to generalize the tree

Rule extraction from decision trees