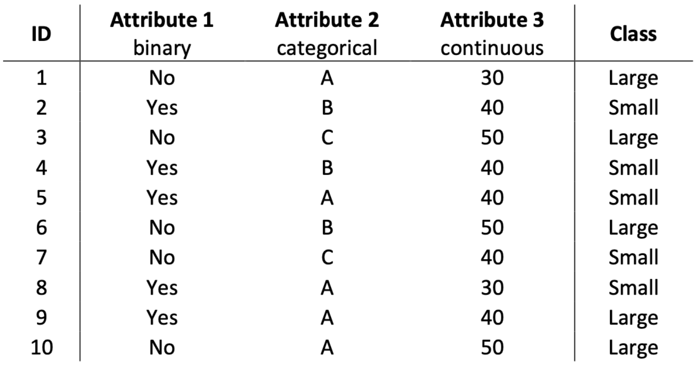
CI603 Data Mining

Classification

Tutorial 5

1. The table below shows a data sample where each item has three attributes and there are two classes ‘Small’ and ‘Large’. Attribute 1 is binary with values ‘yes’ or ‘no’; attribute 2 is categorical with values ‘A’, ‘B’ or ‘C’; attribute 3 is continuous.



ID	Attribute 1	Attribute 2	Attribute 3	Class
	binary	categorical	continuous	
1	No	A	30	Large
2	Yes	B	40	Small
3	No	C	50	Large
4	Yes	B	40	Small
5	Yes	A	40	Small
6	No	B	50	Large
7	No	C	40	Small
8	Yes	A	30	Small
9	Yes	A	40	Large
10	No	A	50	Large

The aim here is to construct a **binary decision tree** using **entropy** to measure impurity (in a binary tree each non-leaf node has two children)

a) Calculate the entropy of the parent node, using the entropy formula

$$Entropy = -\,P(small)\,log_2 P(small) - P(large)\,log_2 P(large)$$

b) Calculate the information gain for each of the following four possible 2-way splits:

- Attribute 1: ‘no’ or ‘yes’;
- Attribute 2: ‘A’ or ‘B/C’;
- Attribute 3: ‘≤ 35’ or ‘> 35’;
- Attribute 3: ‘≤ 45’ or ‘> 45’.

Hence, draw level 1 of the decision tree. Are either of the nodes leaf nodes?

c) Complete level 2 of the decision tree. That is, for each non-leaf node at level 1, consider the possible 2-way splits as identified in part (b) and choose the split with the largest **information gain**.

d) Complete the decision tree.