CI603 Data Mining

Classification

Tutorial 5

(Solution)

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 binary	Attribute 2 categorical	Attribute 3 continuous	Class
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

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

Answer:

Root node: Small: 5; Large: 5

$$Entropy = -\frac{5}{10}\,log_2\frac{5}{10} - \frac{5}{10}\,log_2\frac{5}{10} = 1$$

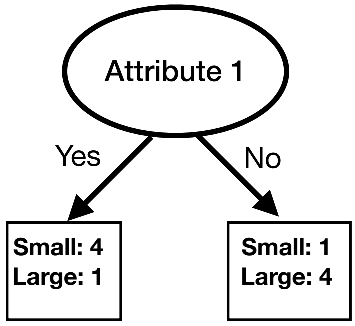
b) Calculate the information gain for each of the following four possible 2-way splits:

- Attribute 1: ‘Yes’ or ‘No’;
- 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?

Answer:

1) Splitting on **Attribute 1**:



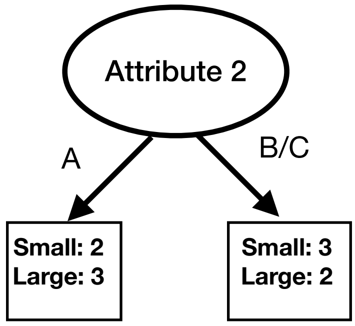
$$I(Attribute\ 1 = Yes) = -\frac{4}{5}\,log_2\frac{4}{5} - \frac{1}{5}\,log_2\frac{1}{5} = 0.7219$$

$$I(Attribute\ 1 = No) = -\frac{1}{5}\,log_2\frac{1}{5} - \frac{4}{5}\,log_2\frac{4}{5} = 0.7219$$

$$I(Attribute) = \frac{5}{10} \times 0.7219 + \frac{5}{10} \times 0.7219 = 0.7219$$ ((**Weighted entropy**)

$$Information\ gain = 1 - 0.7219 = 0.2781$$

2) Splitting on **Attribute 2**: binary split {A, B/C}



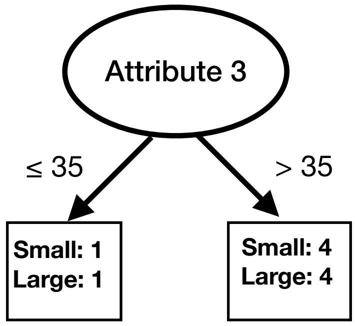
$$I(Attribute\ 2 = A) = -\frac{2}{5}\,log_2\frac{2}{5} - \frac{3}{5}\,log_2\frac{3}{5} = 0.9710$$

$$I(Attribute\ 2 = B/C) = -\frac{3}{5}\,log_2\frac{3}{5} - \frac{2}{5}\,log_2\frac{2}{5} = 0.9710$$

$$I(Attribute) = \frac{5}{10} \times 0.9710 + \frac{5}{10} \times 0.9710 = 0.9710$$ ((**Weighted entropy**)

$$Information\ gain = 1 - 0.9710 = 0.0290$$

3) Splitting on **Attribute 2**: binary split {A, B/C}



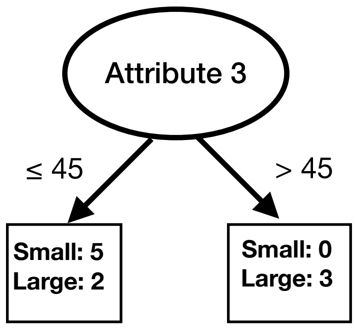
$$I(Attribute\,3 \le 35) = -\frac{1}{2}\,log_2\frac{1}{2} - \frac{1}{2}\,log_2\frac{1}{2} = 1$$

$$I(Attribute\,3 > 35) = -\frac{4}{8}\,log_2\frac{4}{8} - \frac{4}{8}\,log_2\frac{4}{8} = 1$$

$$I(Attribute\,35) = \frac{2}{10} \times 1 + \frac{8}{10} \times 1 = 1$$ ((**Weighted entropy**)

$$Information\,gain = 1 - 1 = 0$$

4) Splitting on **Attribute 2**: binary split {A, B/C}



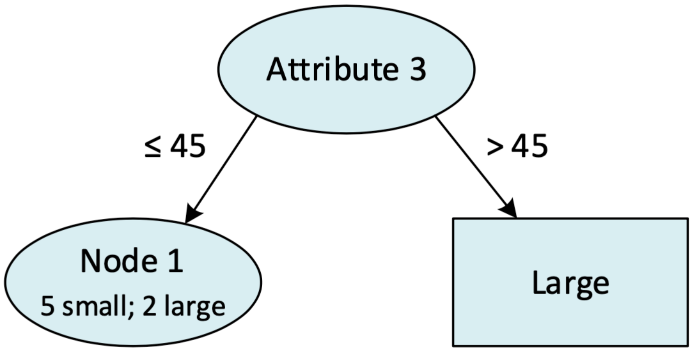
$$I(Attribute\,3 \le 45) = -\frac{5}{7}\,log_2\frac{5}{7} - \frac{2}{7}\,log_2\frac{2}{7} = 0.8631$$

$$I(Attribute\,3 > 45) = 0$$

$$I(Attribute\,45) = \frac{7}{10} \times 0.8631 + \frac{3}{10} \times 0 = 0.6042$$ ((**Weighted entropy**)

$$Information\,gain = 1 - 0.6042 = 0.3958$$

The largest **information gain** is when splitting on **Attribute 3: ≤ 45, >45**. This gives the following level 1 tree where the node following ‘> 45’ is a leaf node.



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**.

Answer:

For the node 1 we have the following data:

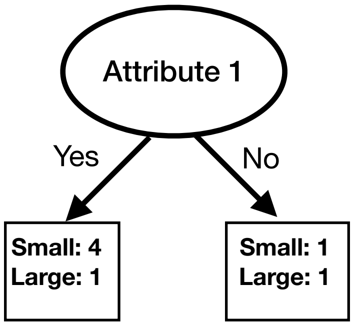
ID	Attribute 1	Attribute 2	Class
1	No	A	Large
2	Yes	B	Small
4	Yes	B	Small
5	Yes	A	Small
7	No	C	Small
8	Yes	A	Small
9	Yes	A	Large

The entropy of this as a parent node is:

Node 1: Small: 5; Large: 2

$$Entropy = -\frac{5}{7}\,log_2\frac{5}{7} - \frac{2}{7}\,log_2\frac{2}{7} = 0.8631$$

1) Splitting on **Attribute 1**:



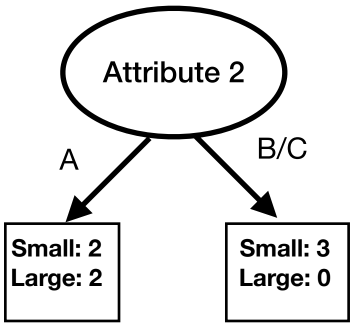
$$I(Attribute\ 1 = Yes) = -\frac{4}{5}\,log_2\frac{4}{5} - \frac{1}{5}\,log_2\frac{1}{5} = 0.7219$$

$$I(Attribute\ 1 = No) = 1$$

$I(Attribute) = \frac{5}{7} \times 0.7219 + \frac{2}{7} \times 1 = 0.8014$ ((**Weighted entropy**)

$$Information\ gain = 0.8631 - 0.8014 = 0.0617$$

2) Splitting on **Attribute 2**: binary split {A, B/C}



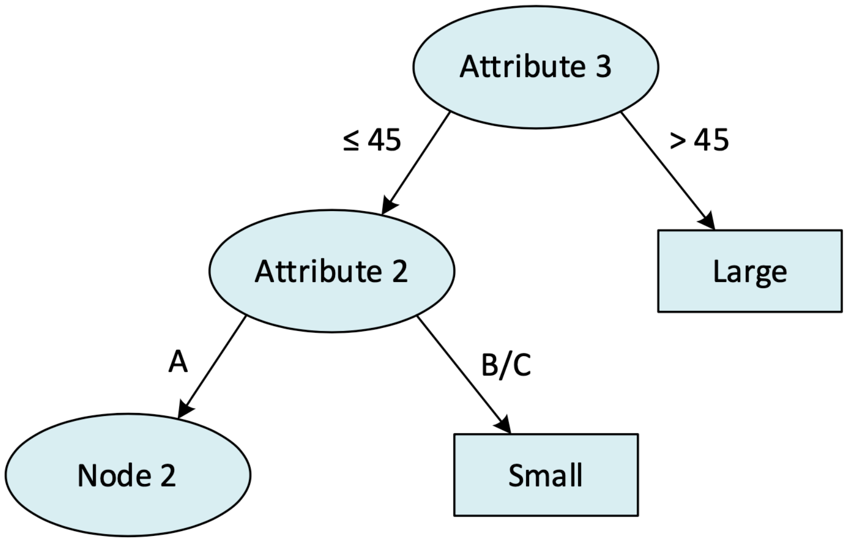
$$I(Attribute\ 2 = A) = 1$$

$$I(Attribute\ 2 = B/C) = 0$$

$I(Attribute) = \frac{4}{7} \times 1 + \frac{3}{7} \times 0 = 0.5714$ ((**Weighted entropy**)

$$Information\ gain = 0.8631 - 0.5714 = 0.2917$$

The largest **information gain** is when splitting on Attribute 2: {A, B/C}. This gives the following level 2 tree where the node following 'B/C' is a leaf node.



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

Answer:

For node 2 we have the following data:

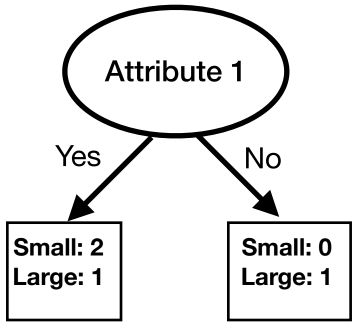
ID	Attribute 1	Class
1	No	Large
5	Yes	Small
8	Yes	Small
9	Yes	Large

The entropy of this as a parent node is:

Node 2: Small: 2; Large: 2

$$Entropy = -\frac{2}{4}\,log_2\frac{2}{4} - \frac{2}{4}\,log_2\frac{2}{4} = 1$$

1) Splitting on **Attribute 1**:



Splitting on **Attribute 1** gives the following completed decision tree.

- The node following 'No' is a **leaf node.**
- For the node following 'Yes' all the attribute values are **identical** so the **splitting stops** and node 3 becomes a **leaf node.** Since 2/3 of the data at **node 3** are in the 'Small' class, we would classify data at this node as '**Small**'.

e) Complete the decision tree.

Answer: