CZ4032 Data Analytics and Mining

Week 9 Tutorial: Classification

1. Consider the following training data of building materials, train a decision tree model using the entropy-based impurity measure at a node t, i.e.,

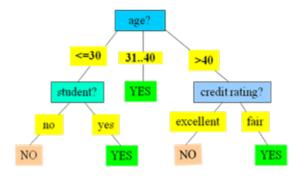
$$Entropy(t) = -\sum_{j} p(j|t) \log_2 p(j|t)$$

where p(j | t) is the relative frequency (or probability) of class j at node t.

Use the trained decision tree model to determine if a material instance (Size="small", Color="green", Shape= "pillar") is likely to be used for construction? Here we assume that each distinct value of a categorical value will become a child node at splitting.

Id	Size	Color	Shape	Can be used?
1	medium	blue	brick	Yes
2	small	red	sphere	Yes
3	large	green	pillar	Yes
4	large	green	sphere	Yes
5	small	red	wedge	No
6	large	red	wedge	No
7	large	red	pillar	No

- 2. The tree growth phase in the construction of a tree classifier is computationally expensive and also data-intensive. Briefly describe why this is so.
- 3. Extract rules from the decision given below.



4. Explain the following pseudocode of CBA for mining CARs.

```
F_1 = \{\text{large 1-ruleitems}\};
     CAR_1 = genRules(F_1);
     prCAR_1 = pruneRules(CAR_1);

for (k = 2; F_{k,1} \neq \emptyset; k++) do
         C_k = candidateGen(F_{k-1});
for each data case d \in D do
5
6
              C_d = \text{ruleSubset}(C_k, d);
7
8
              for each candidate c \in C_d do
9
                  c.condsupCount++;
10
                  if d.class = c.class then c.rulesupCount++
11
              end
12
          end
13
          F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\};
          CAR_k = genRules(F_k);
15
          prCAR_k = pruneRules(CAR_k);
16 end
17 CARs = \bigcup_{k} CAR_{k};
18 prCARs = \bigcup_{k} prCAR_{k};
```

Figure 1: The CBA-RG algorithm

5. apply the following CBA classifier to classify test data:

CBA classifier: <r5, r1, r6, r7, default class n >

$$r_5$$
: $B = w \rightarrow n$
 r_1 : $A = e \rightarrow y$
 r_6 : $A = g, B = q \rightarrow y$
 r_7 : $A = g \rightarrow n$

Test data:

Attribute A	Attribute B	Class C
е	р	??
g	q	??
g	m	??
k	p	??