Data Mining

Problem Set 3

Chapter 3 (pg 198)

2. Consider the training examples shown in Table 4.7 for a binary classification problem.

Table 4.7. Data set for Exercise 2.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

(a) Compute the Gini index for the overall collection of training examples.

(b) Compute the Gini index for the ***Customer ID*** attribute.

(c) Compute the Gini index for the ***Gender*** attribute.

(d) Compute the Gini index for the ***Car Type*** attribute using multiway split.

(e) Compute the Gini index for the ***Shirt Size*** attribute using multiway split.

(f) Which attribute is better, ***Gender***, ***Car Type***, or ***Shirt Size***?

(g) Explain why ***Customer ID*** should not be used as the attribute test condition even though it has the lowest Gini.

3. Consider the training examples shown in Table 4.8 for a binary classification problem.

Table 4.8. Data set for Exercise 3.

Instance	α_1	α_2	α_3	Target Class
1	T	T	1	+
2	T	T	6	+
3	T	F	5	-
4	F	F	4	+
5	F	T	7	-
6	F	T	3	-
7	F	F	8	-
8	T	F	7	+
9	F	T	5	-

(a) What is the entropy of this collection of training examples with respect to the positive class?

(b) What are the information gains of α_1 and α_2 relative to these training examples?

(c) For α_3 , which is a continuous attribute, compute the information gain for every possible split.

(d) What is the best split (among α_1 , α_2 , and α_3) according to the information gain?

(e) What is the best split (between α_1 and α_2) according to the classification error rate?

(f) What is the best split (between α_1 and α_2) according to the Gini index?

5. Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

(a) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(c) Figure 4.13 shows that entropy and the Gini index are both monotonously increasing on the range [0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.

7. The following table summarizes a data set with three attributes A, B, C and two class labels +, −. Build a two-level decision tree.

			Number of Instances	
A	B	C	+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0

F	F	F	0	25

(a) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

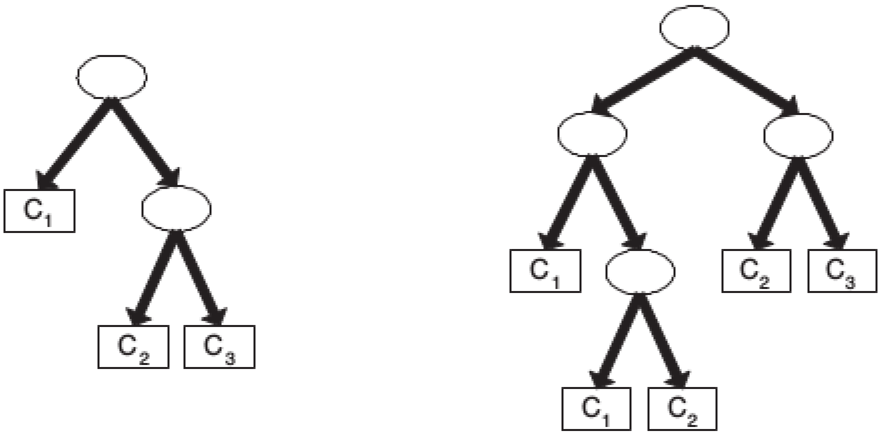
(b) Repeat for the two children of the root node.

(c) How many instances are misclassified by the resulting decision tree?

(d) Repeat parts (a), (b), and (c) using C as the splitting attribute.

(e) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm. The greedy heuristic does not necessarily lead to the best tree.

9. Consider the decision trees shown in Figure 4.3. Assume they are generated from a data set that contains 16 binary attributes and 3 classes, C_1, C_2, and C_3. Compute the total description length of each decision tree according to the minimum description length principle.



(a) Decision tree with 7 errors

(b) Decision tree with 4 errors

Figure 4.3. Decision trees for Exercise 9.

- The total description length of a tree is given by:

  $$Cost(tree, data) = Cost(tree) + Cost(data|tree).$$

- Each internal node of the tree is encoded by the ID of the splitting attribute. If there are m attributes, the cost of encoding each attribute is $log_2(m)$ bits.
- Each leaf is encoded using the ID of the class it is associated with. If there are k classes, the cost of encoding a class is $log_2(k)$ bits.

- $Cost(tree)$ is the cost of encoding all the nodes in the tree. To simplify the computation, you can assume that the total cost of the tree is obtained by adding up the costs of encoding each internal node and each leaf node.
- $Cost(data|tree)$ is encoded using the classification errors the tree commits on the training set. Each error is encoded by $log_2(n)$ bits, where n is the total number of training instances.

Which decision tree is better, according to the MDL principle?