CSci 8980: Data Mining (Fall 2000) Homework #1

Due date:

1. Give decision trees to represent the following boolean functions:

   (a) A and B and not C

   (b) ((not A) and B) or (C and D)

   (c) (A or B) xor (B and C)

2. Consider the following set of training examples:

Instance	a_1	a_2	a_3	classification
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	T	1.0	+
9	T	F	10.0	+
10	F	T	8.0	-

   (a) What is the entropy of this collection of training examples with respect to the target function classification?

   (b) What are the information gains of a_1 and a_2 relative to these training examples?

   (c) For a_3, which is a continuous attribute, compute the information gains of every possible split.

   (d) What is the best split (among a1, a2, and a3) according to the information gain?

   (e) What is the best split (among a1 and a2) according to classification error rate?

   (f) What is the best split (among a1 and a2) according to gini index?

3. Solve the following questions for the training data tabulated below. (Note: There are total 200 training examples.)

X	Y	Z	No. of Class C1 Examples	No. of Class C2 Examples
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

(a) Compute a *two-level* decision tree using the greedy approach taken by C4.5 algorithm. You can use the classification error rate as the criterion used for splitting. Compute the overall error rate of the tree on training data.

(b) Use variable X as the first splitting attribute, then choose the best available splitting attribute at each of the two successor nodes. Compute the overall error rate of the resultant tree on training data.

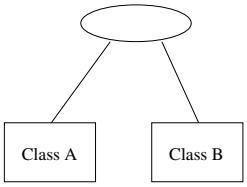
(c) Discuss the results obtained in parts (a) and (b) above. Comment on the suitability of the greedy heuristic used for splitting attribute selection.

4. Compute the description length of two decision trees shown below according to the MDL principle. Which decision tree is better?

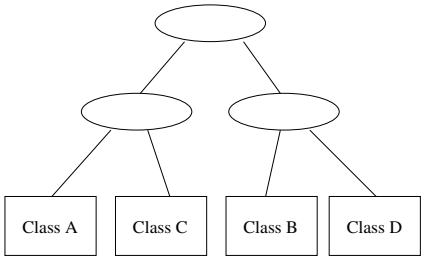
   - For this data, there are 16 samples, 8 attributes, and 4 classes.
   - The description length is given by:

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

   - Each internal node is encoded by the id of the splitting attribute. If there are m attributes, the cost of encoding each attribute is $log_2 m$.
   - 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 is $log_2 k$
   - $Cost(tree)$ is the cost of encoding all the nodes in the tree.
   - $Cost(data|tree)$ is encoded using the classification errors the tree commits on training set. Each error is encoded by $log_2 n$, where n is the total number of training instances.

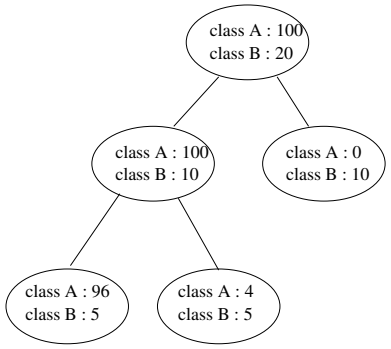


(a) Decision tree with 4 errors

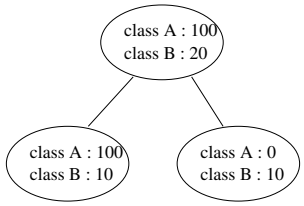


(b) Decision tree with 2 errors

5. Consider the following two trees.



(a) Original tree



(b) Pruned tree

(a) Which tree is better in terms of classification error rate?

(b) Which tree is better in terms of precision, recall, and F1 of class B?

(c) Assume that the penalty of misclassifying class B data as class A is 1 and the penalty of misclassifying class A data as class B is 2. Which tree is better in terms of minimizing penalty?