Assignment#3 Solutions (Chapter 4)

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

A	B	C	Number of Instances	
			+	−
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.

Answer: The error rate for the data without partitioning on any attribute is

$$E_{orig} = 1 - \max(\frac{50}{100}, \frac{50}{100}) = \frac{50}{100}.$$

After splitting on attribute *A*, the gain in error rate is:

	$A = T$	$A = F$
+	25	25
−	0	50

$$E_{A=T} = 1 - \max(\frac{25}{25}, \frac{0}{25}) = \frac{0}{25} = 0$$

$$E_{A=F} = 1 - \max(\frac{25}{75}, \frac{50}{75}) = \frac{25}{75}$$

$$\Delta_A = E_{orig} - \frac{25}{100}E_{A=T} - \frac{75}{100}E_{A=F} = \frac{25}{100}$$

After splitting on attribute *B*, the gain in error rate is:

	$B = T$	$B = F$
+	30	20
−	20	30

$$E_{B=T} = \frac{20}{50}$$

$$E_{B=F} = \frac{20}{50}$$

$$\Delta_B = E_{orig} - \frac{50}{100}E_{B=T} - \frac{50}{100}E_{B=F} = \frac{10}{100}$$

After splitting on attribute *C*, the gain in error rate is:

	$C = T$	$C = F$
+	25	25
−	25	25

$$E_{C=T} = \frac{25}{50}$$

$$E_{C=F} = \frac{25}{50}$$

$$\Delta_C = E_{orig} - \frac{50}{100}E_{C=T} - \frac{50}{100}E_{C=F} = \frac{0}{100} = 0$$

The algorithm chooses attribute *A* because it has the highest gain.

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

Answer: Because the *A* = *T* child node is pure, no further splitting is needed. For the *A* = *F* child node, the distribution of training instances is:

B	C	Class label	
		+	−
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

The classification error of the *A* = *F* child node is:

$$E_{orig} = \frac{25}{75}$$

After splitting on attribute *B*, the gain in error rate is:

	$B = T$	$B = F$
+	25	0
−	20	30

$$E_{B=T} = \frac{20}{45}$$
$$E_{B=F} = 0$$
$$\Delta_B = E_{orig} - \frac{45}{75}E_{B=T} - \frac{20}{75}E_{B=F} = \frac{5}{75}$$

After splitting on attribute *C*, the gain in error rate is:

	$C = T$	$C = F$
+	0	25
−	25	25

$$E_{C=T} = \frac{0}{25}$$
$$E_{C=F} = \frac{25}{50}$$
$$\Delta_C = E_{orig} - \frac{25}{75}E_{C=T} - \frac{50}{75}E_{C=F} = 0$$

The split will be made on attribute *B*.

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

Answer: 20 instances are misclassified. (The error rate is 20/100.)

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

Answer: For the *C* = *T* child node, the error rate before splitting is:
E_{orig} = 25/50.
After splitting on attribute *A*, the gain in error rate is:

	$A = T$	$A = F$
+	25	0
−	0	25

$$E_{A=T} = 0$$
$$E_{A=F} = 0$$
$$\Delta_A = \frac{25}{50}$$

After splitting on attribute *B*, the gain in error rate is:

	$B = T$	$B = F$
+	5	20
−	20	5

$$E_{B=T} = \frac{5}{25}$$
$$E_{B=F} = \frac{5}{25}$$
$$\Delta_B = \frac{15}{50}$$

Therefore, *A* is chosen as the splitting attribute.
For the C = F child, the error rate before splitting is: E_{orig} = 25/50.
After splitting on attribute *A*, the error rate is:

	$A = T$	$A = F$
+	0	25
−	0	25

$$E_{A=T} = 0$$
$$E_{A=F} = \frac{25}{50}$$
$$\Delta_A = 0$$

After splitting on attribute *B*, the error rate is:

	$B=T$	$B=F$
+	25	0
−	0	25

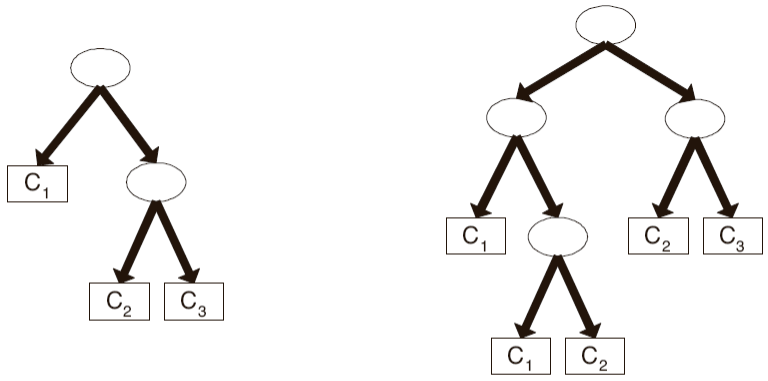
$$E_{B=T} = 0$$
$$E_{B=F} = 0$$
$$\Delta_B = \frac{25}{50}$$

Therefore, *B* is used as the splitting attribute. The overall error rate of the induced tree is 0.

(e) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm.

Answer: 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?

Answer: Because there are 16 attributes, the cost for each internal node in the decision tree is:
$\log_2(m) = \log_2(16) = 4$
Furthermore, because there are 3 classes, the cost for each leaf node is:
$\lceil \log_2(k) \rceil = \lceil \log_2(3) \rceil = 2$
The cost for each misclassification error is $\log_2(n)$.

The overall cost for the decision tree (a) is $2\times4+3\times2+7\times\log_2 n = 14+7\log_2 n$ and the overall cost for the decision tree (b) is $4\times4+5\times2+4\times5 = 26+4\log_2 n$. According to the MDL principle, tree (a) is better than (b) if $n < 16$ and is worse than (b) if $n > 16$.

10. While the .632 bootstrap approach is useful for obtaining a reliable estimate of model accuracy, it has a known limitation. Consider a two-class problem, where there are equal number of positive and negative examples in the data. Suppose the class labels for the examples are generated randomly. The classifier used is an unpruned decision tree (i.e., a perfect memorizer). Determine the accuracy of the classifier using each of the following methods.

(a) The holdout method, where two-thirds of the data are used for training and the remaining one-third are used for testing.

Answer: Assuming that the training and test samples are equally representative, the test error rate will be close to 50% since each example of the test sample is equally likely to be positive or negative as in the Training sample.

(b) Ten-fold cross-validation.

Answer: Assuming that the training and test samples for each fold are equally representative, the test error rate will be close to 50% using the same argument as in part (a).

(c) The .632 bootstrap method.

Answer: The training error for a perfect memorizer is 100% while the error rate for each bootstrap sample is close to 50%. Substituting this information into the formula for .632 bootstrap method, the accurarcy estimate is:

$$\frac{1}{b}\sum_{i=1}^{b}\left[0.632\times0.5+0.368\times1\right]=0.684.$$

(d) From the results in parts (a), (b), and (c), which method provides a more reliable evaluation of the classifier's accuracy?

Answer: The ten-fold cross-validation and holdout method provides a better error estimate than the .632 bootstrap method, which produced a more optimistic estimate of the accuracy.