Name: Arpit Jain Roll No: 18ucs085

Assignment 1

Q1. Draw the full decision tree for the parity function of four Boolean attributes, A, B, C, and D. Is it possible to simplify the tree?

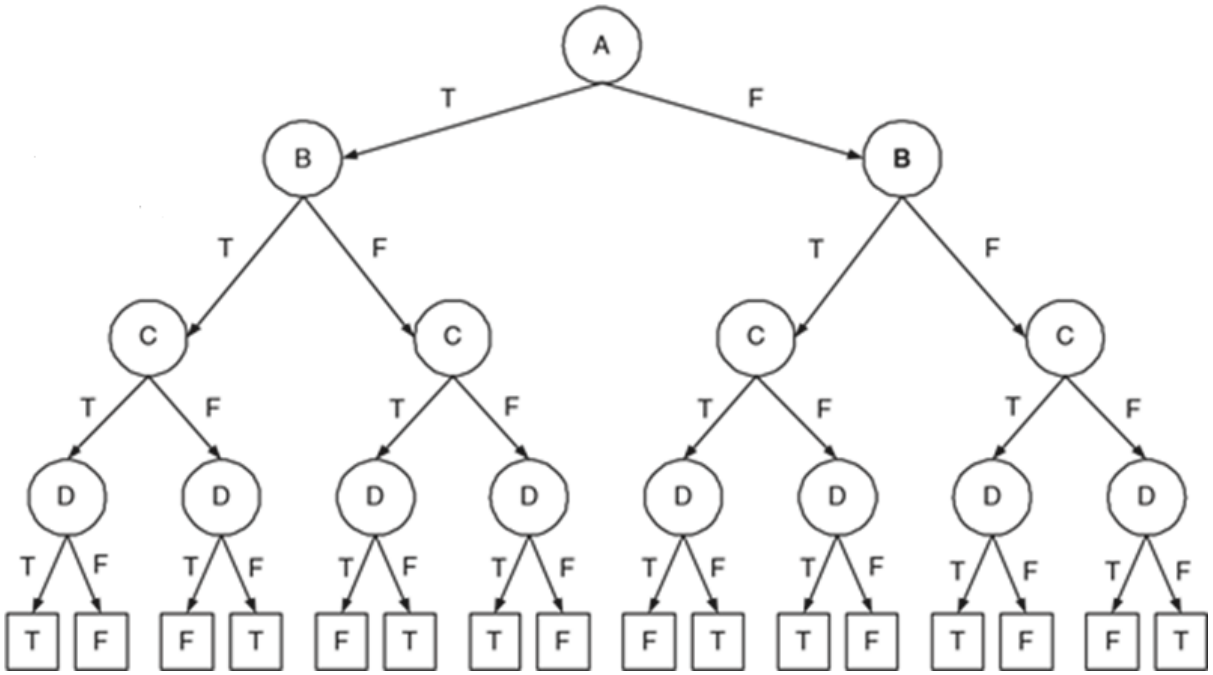
Ans1.

Below is the Decision Table and the Decision Tree for the Boolean attributes. The tree cannot be further simplified.

Decision Table:

A	B	C	D	Class
T	T	T	T	T
T	T	T	F	F
T	T	F	T	F
T	T	F	F	T
T	F	T	T	F
T	F	T	F	T
T	F	F	T	T
T	F	F	F	F
F	T	T	T	F
F	T	T	F	T
F	T	F	T	T
F	T	F	F	F
F	F	T	T	T
F	F	T	F	F
F	F	F	T	F
F	F	F	F	T

Decision Tree:



Q2. 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.
Ans.
Gini = $1 - 2*0.5^2 = 0.5$.

(b) Compute the Gini index for the Customer ID attribute.
Ans.
The Gini index value for each Customer ID is 0. Therefore, the overall Gini index value for Customer ID is 0.

(c) Compute the Gini index for the Gender attribute.
Ans.
The Gini index value for male is $1 - 2 * 05^2 = 0.5$. The Gini index value for female is also 0.5. Therefore, the overall Gini index value for Gender is 0.5 * 0.5 + 0.5 * 0.5 = 0.25 + 0.25 = 0.5.

(d) Compute the Gini index for the Car Type attribute using multiway split.
Ans.
The Gini index value for the Family Car is 0.375, Sports Car is 0, and Luxury Car is 0.2188. The overall Gini index value for Car Type is 0.1625.

(e) Compute the Gini index for the Shirt Size attribute using multiway split.
Ans.
The Gini index value for Small Shirt Size is 0.48, Medium Shirt Size is 0.4898, Large Shirt Size is 0.5, and Extra-Large Shirt Size is 0.5. The overall Gini index value for Shirt Size is 0.4914.

(f) Which attribute is better, Gender, Car Type, or Shirt Size?
Ans.
The better attribute can be decided based on the Gini index value. Car Type has the smallest Gini index value out of given attributes.

(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.
Ans.
Each customer has a unique Customer ID, which eliminates using it as a predictive attribute because each new customer has a new Customer ID.

Q3. Consider the training examples shown in Table 4.8 for a binary classification problem.

Table 4.8. Data set for Exercise 3.

Instance	a1	a2	a3	Target Class
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	F	7.0	+
9	F	T	5.0	-

(a) What is the entropy of this collection of training examples with respect to the positive class?
Ans.
Four training examples support the (+) class, and the (-) class is supported by five training examples. P(+) = 4/9 and P(-) = 5/9.The total entropy of the collection of training examples is $-(4/9)\log_2(4/9) - (5/9)\log_2(5/9) = 0.9911$.

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

The count of training examples supporting an attribute are:

	a1		a2	
	+	-	+	-
T	3	1	2	3
F	1	4	2	2

The entropy for a_1 is:

$$\frac{4}{9}\left[-\left(\frac{3}{4}\right)log_2(3/4)-\left(\frac{1}{4}\right)log_2(1/4)\right]+\frac{5}{9}\left[-\left(\frac{1}{5}\right)log_2\left(1/5\right)-\left(\frac{4}{5}\right)log_2\left(4/5\right)\right]=0.7616$$

The total information gain for a_1 is: 0.9911 – 0.7616 = 0.2294

The entropy for a_2 is:

$$\frac{5}{9}\left[-\left(\frac{2}{5}\right)log_2\left(2/5\right)-\left(\frac{3}{5}\right)log_2\left(3/5\right)\right]+\frac{4}{9}\left[-\left(\frac{2}{4}\right)log_2(2/4)-\left(\frac{2}{4}\right)log_2(2/4)\right]=0.9839$$

The total information gain for a_1 is: 0.9911 – 0.9839 = 0.0072

(c) For a_3, which is a continuous attribute, compute the information gain for every possible split.
Ans.
The information gain for every possible split based on attribute a_3 is:

a_3	Class label	Split point	Entropy	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0 5.0	- -	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

(d) What is the best split (among a_1, a_2 and a_3) according to the information gain?
Ans.
As per the information gain computed previously, a_1 produces the best split.

(e) What is the best split (between a_1 and a_2) according to the classification error rate?
Ans.
The error rate for attribute a_1 is 2/9.
The error rate for attribute a_2 is 4/9.
Therefore, a1 produces the best split.

(f) What is the best split (between a_1 and a_2) according to the Gini index?
Ans.
The Gini index value for attribute a_1 is:

$$\frac{4}{9}\left[1-\left(3/4\right)^2-\left(1/4\right)^2\right]+\frac{5}{9}\left[1-\left(1/5\right)^2-\left(4/5\right)^2\right]=0.3444$$

The Gini index value for attribute a_2 is:

$$\frac{5}{9}\left[1-\left(2/5\right)^2-\left(3/5\right)^2\right]+\frac{4}{9}\left[1-\left(2/4\right)^2-\left(2/4\right)^2\right]=0.4889$$

Therefore, a_1 produces the best split.

Q4. Show that the entropy of a node never increases after splitting it into smaller successor nodes.

Ans4.

Let $Y = \{y_1, y_2, \cdots, y_c\}$ denote the c classes and $X = \{x_1, x_2, \cdots, x_k\}$ denote the k attribute values of an attribute X. Before a node is split on X, the entropy is:

$$E(y) = -\sum_{j=1}^{c} P(y_j)\, log_2\, P(y_j) = \sum_{i=1}^{c} \sum_{i=1}^{k} P(x_1, y_j)\, log_2\, P(y_j)$$

where we have used the fact that $P(y_j) = \sum_{i=1}^{k} P(x_1, y_j)$ from the law of total probability.

After splitting on X, the entropy for each child node $X = x_i$ is:

$$E(Y|x_i) = -\sum_{j=1}^{c} P(y_j|x_i)\, log_2\, P(y_j|x_i)$$

where $P(y_j|x_i)$ is the fraction of examples with $X = x_i$ that belong to class y_j. The entropy after splitting on X is given by the weighted entropy of the children nodes:

$$\begin{aligned} E(Y|X) &= \sum_{i=1}^{k} P(x_i) E(Y|x_i) \\ &= -\sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i) P(y_j|x_i) \log_2 P(y_j|x_i) \\ &= -\sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j|x_i), \end{aligned}$$

where we have used a known fact from probability theory that $P(x_i, y_j) = P(y_i|x_i) * P(x_i)$. Note that $E(Y|X)$ is also known as the conditional entropy of Y given X.

To answer this question, we need to show that $E(Y|X) \leq E(Y)$. Let us compute the difference between the entropies after splitting and before splitting, i.e., $E(Y|X) - E(Y)$:

$$\begin{aligned} & E(Y|X) - E(Y) \\ = & -\sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j|x_i) + \sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j) \\ = & \sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(y_j)}{P(y_j|x_i)} \\ = & \sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(x_i) P(y_j)}{P(x_i, y_j)} \end{aligned}$$

To prove that Equation 4.4 is non-positive, we use the following property of a logarithmic function:

$$\sum_{k=1}^{d} a_k \log(z_k) \leq \log \left(\sum_{k=1}^{d} a_k z_k \right)$$

, subject to the condition that $\sum_{k=1}^{d} a_k = 1$. This property is a particular case of a more general theorem involving convex functions (the logarithmic function) known as Jensen's inequality.

By applying Jensen's inequality, Equation 4.4 can be bounded as follows:

$$\begin{aligned} E(Y|X) - E(Y) &\leq \log_2 \left[\sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i, y_j) \frac{P(x_i)P(y_j)}{P(x_i, y_j)} \right] \\ &= \log_2 \left[\sum_{i=1}^{k} P(x_i) \sum_{j=1}^{c} P(y_j) \right] \\ &= \log_2(1) \\ &= 0 \end{aligned}$$

Because $E(Y|X) - E(Y) \leq 0$, it follows that entropy never increases after splitting on an attribute.

Q5. 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?

Ans.

The count of training examples supporting an attribute are:

	A		B	
	+	-	+	-
T	4	3	3	1
F	0	3	1	5

The overall entropy before splitting is

$$\left[-\left(\frac{4}{10}\right) log_2(4/10) - \left(\frac{6}{10}\right) log_2(6/10) \right] = 0.9710$$

The information gain after splitting on A is

For True, $\left[-\left(\frac{4}{7}\right) log_2(4/7) - \left(\frac{3}{7}\right) log_2(3/7) \right] = 0.9852$

For False, $\left[-\left(\frac{3}{3}\right) log_2(3/3) - \left(\frac{0}{7}\right) log_2(0/7) \right] = 0$

Therefore, information gain after splitting on A is

0.9710 – (7/10)(0.9852) – (3/10)(0) = 0.2813.

The information gain after splitting on B is

For True, $\left[-\left(\frac{3}{4}\right) log_2\left({}^3/_4\right) - \left(\frac{1}{4}\right) log_2\left({}^1/_4\right)\right] = 0.8113$

For False, $\left[-\left(\frac{1}{6}\right) log_2\left({}^1/_6\right) - \left(\frac{5}{6}\right) log_2\left({}^5/_6\right)\right] = 0.6500$

Therefore, information gain after splitting on B is

0.9710 – (4/10)(0.8113) – (6/10)(0.6500) = 0.2565.

Therefore, attribute A would be chosen to split the node.

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

Ans.

The overall Gini index value before splitting is,

$1 – (0.4)^2 – (0.6)^2 = 0.48$

The gain in Gini index value after splitting on A is,

For True, $1 – (4/7)^2 – (3/7)^2 = 0.4898$

For False, $1 – (3/3)^2 – (0/3)^2 = 0$

Total = 0.48 – (7/10)(0.4898) – (3/10)(0) = 0.1371

The gain in Gini index value after splitting on B is,

For True, $1 – (1/4)^2 – (3/4)^2 = 0.3750$

For False, $1 – (1/6)^2 – (5/6)^2 = 0.2778$

Total = 0.48 – (4/10)(0.3750) – (6/10)(0.2778) = 0.1633

Therefore, attribute B would be chosen to split the node.

(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.

Ans.

Yes. We can see that information gain prefers to attribute A, whereas the increase in Gini index value prefers attribute B for split despite these measures having similar range and monotonous behaviour.

Q6. Consider the following set of training examples.

X	Y	Z	No. of C1 Examples	No. of 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 described in this chapter. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

Ans.
Let's first split the attributes based on Level 1.
The count of training examples supporting an attribute are:

	X		Y		Z	
	0	1	0	1	0	1
C1	60	40	40	60	30	70
C2	60	40	60	40	70	30

The error rates of the attributes based on splitting on Level1 are:
X: (60+40)/200 = 0.5
Y: (40+40)/200 = 0.4
Z: (30+30)/200 = 0.3
Here, Z has the lowest error rate. Hence it is chosen for splitting at Level 1.

Now let's split the attributed based on Level 2. After splitting on Z, we have X and Y left further to be splitting upon. The distribution of training examples on split based on X and Y,
When splitting upon Z=0 is as follows:

	X		Y	
	0	1	0	1
C1	15	15	15	15
C2	45	25	45	25

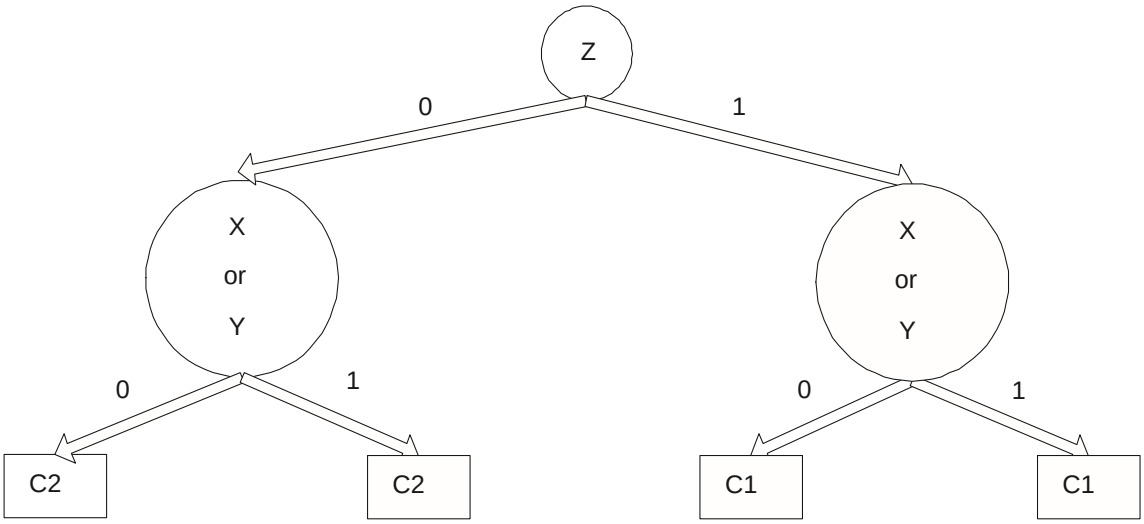
The error rates of the attributes, when followed upon the split on Z=0 are:
X: (15+15)/100 = 0.3
Y: (15+15)/100 = 0.3
Both of these have an equal error rate.

When splitting upon Z=1 is as follows:

	X		Y	
	0	1	0	1
C1	45	25	25	45
C2	15	15	15	15

The error rates of the attributes, when followed upon the split on Z=1 are:
X: (15+15)/100 = 0.3
Y: (15+15)/100 = 0.3
Both of these have an equal error rate.

Hence, the decision tree at level 2 can be constructed by either splitting on X or Y.



The overall rate of error of the tree is (15+15+15+15)/200 = 0.3.

(b) Repeat part (a) using X as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?

Ans. Let's choose X as the Level 1 split. After splitting on X, we have Y and Z left further to be splitting upon. The distribution of training examples on split based on Y and Z,

When splitting upon X=0 is as follows:

	Y		Z	
	0	1	0	1
C1	5	55	15	45
C2	55	5	45	15

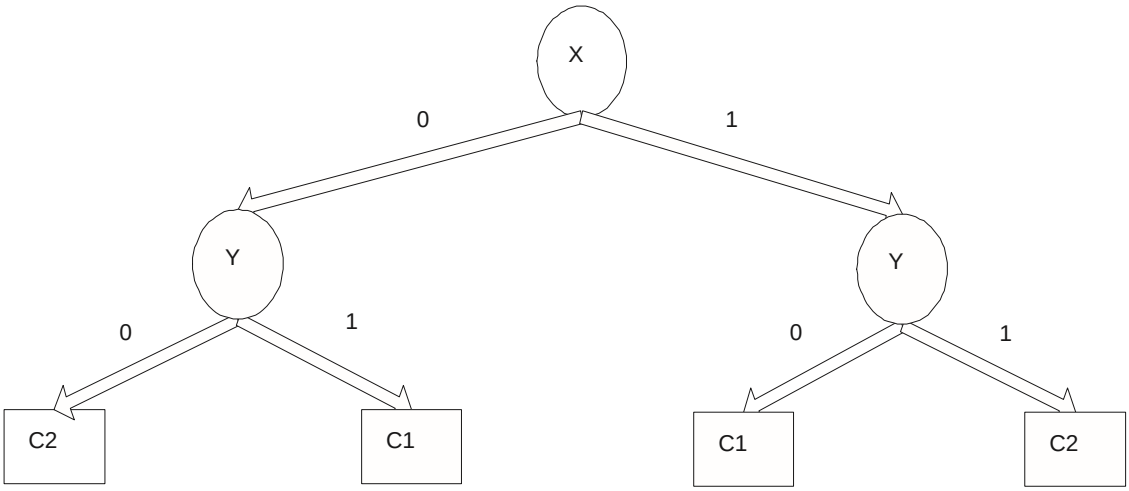
The error rates of the attributes, when followed upon the split on X=0 are:
Y: (5+5)/120 = 10/120
Z: (15+15)/120 = 30/120
Here, the error rate of Y is lower. Hence it can provide a better split.

When splitting upon X=1 is as follows:

	Y		Z	
	0	1	0	1
C1	35	5	15	25
C2	5	35	25	15

The error rates of the attributes, when followed upon the split on X=1 are:
Y: (5+5)/80 = 10/80
Z: (15+15)/80 = 30/80
Here, the error rate of Y is lower. Hence it can provide a better split.

Here, for both X=0 and X=1, Y provides a better split. Hence Y would be used as a second level split.



The overall rate of error of the tree is (10+10)/200 = 0.1.

(c) Compare the results of parts (a) and (b). Comment on the suitability of the greedy heuristic used for splitting attribute selection.
Ans. As we can see that approach (b) has a lower overall error rate when compared with (a), This suggests that the greedy heuristic is not the best suitable method for splitting attribute selection.

Q7. 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	No. 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.
Ans.
The initial error rate for the data without partitioning any attribute is
1 – max(50/100, 50/100) = 0.5

The count of training examples supporting an attribute are:

	A		B		C	
	+	-	+	-	+	-

T	25	0	30	20	25	25
F	25	50	20	30	25	25

The error rates of the attributes based on splitting are:
Gain in error rate for X:
For True, 1 – max(25/25, 0/25) = 0
For False, 1 – max(25/75, 50/75) = 25/75
For X, (50/100) – (25/100)(0) – (75/100)*(25/75) = 25/100

Gain in error rate for Y:
For True, 1 – max(30/50, 20/50) = 20/50
For False, 1 – max(20/50, 30/50) = 20/50
For Y, (50/100) – (50/100)(20/50) – (50/100)*(20/50) = 10/100

Gain in error rate for Z:
For True, 1 – max(25/50, 25/50) = 25/50
For False, 1 – max(25/50, 25/50) = 25/50
For Z, (50/100) – (50/100)(25/50) – (50/100)*(25/50) = 0

The maximum gain in error is found in A. Hence it would be used for splitting.

(b) Repeat for the two children of the root node.
Ans.
Since for True, in the case of A, it is a pure child. Hence, it does not need further splitting.
For False, in the case of A, the count distribution of training data is as follows:

B	C	Class Label	
		+	-
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

The initial error rate for the data split on attribute A on False = 25/75

On splitting on attribute B,
For True, 20/45
For False, 0
Gain in error rate on B = (25/75) – (45/75)*(20/45) – (30/75)*(0) = 5/75

On splitting on attribute C,
For True, 0/25
For False, 25/50
Gain in error rate on B = (25/75) – (25/75)*(0/25) – (50/75)*(25/50) = 0

Here, the gain in error rate is maximum in B. Hence the split would be done on B.

(c) How many instances are misclassified by the resulting decision tree?
Ans.
The error rate is (20/100). Hence 20 instances are misclassified.

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

Ans.

Considering C as the attribute and T as the value, the error rate before splitting is 25/50.

	A		B	
	+	-	+	-
T	25	0	5	20
F	0	25	20	5

On further splitting on A,
For True, 0
For False, 0
The gain in error rate on A, 25/50

On further splitting on B,
For True, 5/25
For False, 5/25
The gain in error rate on A, 15/50

Since A has a higher gain in error rate, hence it is chosen.

Considering C as the attribute and F as the value, the error rate before splitting is 25/50.

	A		B	
	+	-	+	-
T	0	0	25	0
F	25	25	0	25

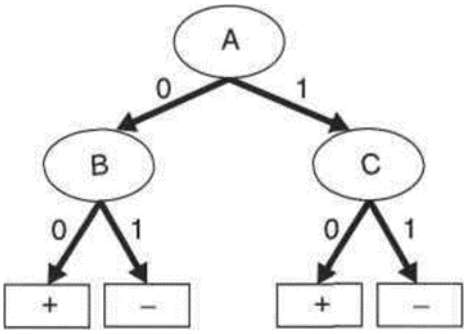
On further splitting on A,
For True, 0
For False, 25/50
The gain in error rate on A, 0

On further splitting on B,
For True, 0
For False, 0
The gain in error rate on A, 25/50

Since B has a higher gain in error rate, hence it is chosen.

(e) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm.
Ans. Since both approaches lead to an equal final error rate, it can be said that the greedy heuristic does not necessarily lead to the best tree.

Q8. Consider the decision tree shown in Figure 4.30.



Training:

Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	-
5	1	0	0	+
6	1	0	0	+
7	1	1	0	-
8	1	0	1	+
9	1	1	0	-
10	1	1	0	-

Validation:

Instance	A	B	C	Class
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	-
15	1	0	0	+

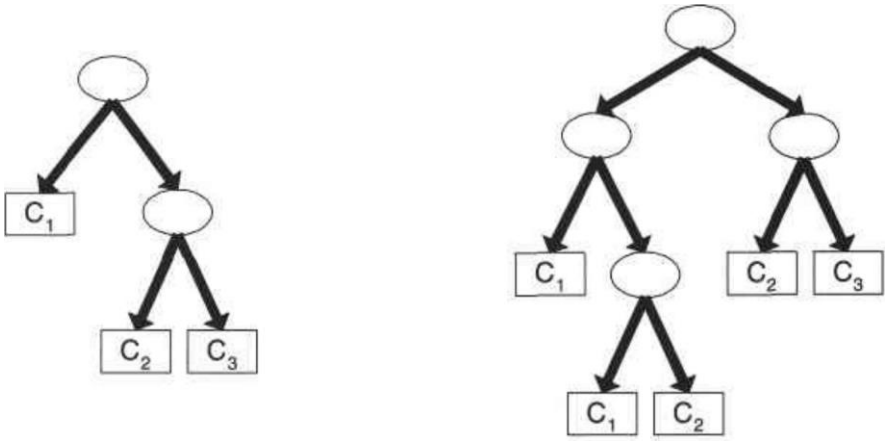
Table 4.30. Decision tree and data sets for Exercise 8.

(a) Compute the generalization error rate of the tree using the optimistic approach.
Ans.
3/10 = 0.3.

(b) Compute the generalization error rate of the tree using the pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)
Ans.
(3+4*0.5)/10 = 0.5.

(c) Compute the generalization error rate of the tree using the validation set shown above. This approach is known as reduced error pruning.
Ans.
4/5 = 0.8.

Q9. Consider the decision trees shown in Figure 4.31. Assume they are generated from a data set that contains 16 binary attributes and 3 classes, C1, C2, and C3.



(a) Decision tree with 7 errors. (b) Decision tree with 4 errors

Table 4.31. Decision trees for Exercise 9.

Compute the total description length of each decision tree according to the minimum description length principle.

- 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 log2 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 log2 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 log2 n bits, where n is the total number of training instances.

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

Ans 9.

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 $n > 16$.

Q10. While the .632 bootstrap approach is useful for obtaining a reliable estimate of model accuracy, it has a known limitation 1127]. 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.
Ans.
Assuming the equivalent distribution of training and test samples, 50%(approx.).

(b) Ten-fold cross-validation.
Ans.
Assuming the equivalent distribution of training and test samples, 50%(approx.).

(c) The .632 bootstrap method.
Ans.
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 the .632 bootstrap method, the error 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 (u), (b), and (c), which method provides a more reliable evaluation of the classifier's accuracy?
Ans.
The ten-fold cross-validation and holdout method provides a better error estimate than the .632 bootstrap method.

Q11. Consider the following approach for testing whether a classifier A beats another classifier B. Let N be the size of a given data set, p_A be the accuracy of classifier A, p_B be the accuracy of classifier B, and $p = (p_A + p_B)/2$ be the average accuracy for both classifiers. To test whether classifier A is significantly better than B, the following Z-statistic is used:

$$z=\frac{pA-pB}{\sqrt{\frac{2p(1-p)}{N}}}$$

Classifier A is assumed to be better than classifier B if Z > 1.96.

Table 4.9 compares the accuracies of three different classifiers, decision tree classifiers, naive Bayes classifiers, and support vector machines, on various data sets. (The latter two classifiers are described in Chapter 5.)

Table 4.9. Comparing the accuracy of various classification methods.

Data Set	Size (N)	Decision Tree (%)	naive Bayes (%)	Support vector machine (%)

Anneal	898	92.09	79.62	87.19
Australia	690	85.51	76.81	84.78
Auto	205	81.95	58.05	70.73
Breast	699	95.14	95.99	96.42
Cleve	303	76.24	83.50	84.49
Credit	690	85.80	77.54	85.07
Diabetes	768	72.40	75.91	76.82
German	1000	70.90	74.70	74.40
Glass	214	67.29	48.59	59.81
Heart	270	80.00	84.07	83.70
Hepatitis	155	81.94	83.23	87.10
Horse	368	85.33	78.80	82.61
Ionosphere	351	89.17	82.34	88.89
Iris	150	94.67	95.33	96.00
Labor	57	78.95	94.74	92.98
Led7	3200	73.34	73.16	73.56
Lymphography	148	77.03	83.11	86.49
Pima	768	74.35	76.04	76.95
Sonar	208	78.85	69.71	76.92
Tic-tac-toe	958	83.72	70.04	98.33
Vehicle	846	71.04	45.04	74.94
Wine	178	94.38	96.63	98.88
	101	93.07	93.07	96.04

Summarize the performance of the classifiers given in Table 4.9 using the following3 x 3table:

win-loss-draw	Decision tree	Naive Bayes	Support vector machine
Decision tree	0 - 0 - 23		
Naive Bayes		0 - 0 – 23	
Support vector machine			0 - 0 - 23

Each cell in the table contains the number of wins. losses, and draws when comparing the classifier in a given row to the classifier in a given column.

Ans 11.

win-loss-draw	Decision tree	Naive Bayes	Support vector machine
Decision tree	0 - 0 - 23	9 - 3 – 11	2 - 7 – 14
Naive Bayes	3 - 9 – 11	0 - 0 – 23	0 - 8 – 15
Support vector machine	7 - 2 – 14	8 - 0 - 15	0 - 0 - 23

Q12. Let X be a binomial random variable with mean N_p and variance N_p (l -p). Show that the ratio *X/N* also has a binomial distribution with mean *p* and variance *p(1-p)/N*.

Ans 12.

Let $r = X/N$. Since X has a binomial distribution, r also has the same distribution. The mean and variance for r can be computed as follows:

Mean, $E[r] = E[X/N] = E[X]/N = (Np)/N = p$;

Variance, $E[(r-E[r])^2] = E[(X/N - E[X]/N)^2] = E[(X - E[X])^2]/N^2 = (Np)/(1-p) = p(1-p)/N$