Assignment#2 Solutions (Chapter 4)

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

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

Answer: There are four positive examples and five negative examples. Thus, $P(+) = 4/9$ and $P(-) = 5/9$. The entropy of the 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?

Answer: For attribute a_1, the corresponding counts and probabilities are:

a_1	+	-
T	3	1
F	1	4

The entropy for a_1 is given by:
4/9 [-(3/4)log(3/4) –(1/4)log(1/4)] + 5/9[-(1/5)log(1/5) – (4/5)log(4/5)] = 0.7616. Therefore, the information gain for a_1 is $0.9911 - 0.7616 = 0.2294$.

For attribute a_2, the corresponding counts and probabilities are:

A_2	+	-
T	2	3
F	2	2

The entropy for a_2 is given by:
5/9[-(2/5)log(2/5) – (3/5)log(3/5)] + 4/9 [-(2/4)log(2/4) – (2/4)log(2/4)] = 0.9839. Therefore, the information gain for a_2 is $0.9911 - 0.9839 = 0.0072$.

c) For a_3, which is a continuous attribute, compute the information gain for every possible split.

Answer:

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

The best split for a_3 occurs at split point equals to 2.

d) What is the best split (among a_1, a_2, and a_3) according to the infor-mation gain?

Answer: According to information gain, a_1 produces the best split.

e) What is the best split (between a_1 and a_2) according to the classification error rate?

Answer: For attribute a_1, the error rate = 2/9. For attribute a_2, the error rate = 4/9. Therefore, according to error rate, a_1 produces the best split.

f) What is the best split (between a_1 and a_2) according to the Gini index?

Answer: For attribute a_1 the Gini index is given by:

$4/9 [1-(3/4)^2-(1/4)^2] + 5/9 [1-(1/5)^2 -(4/5)^2] = 0.3444$.

For attribute a_2, the Gini index is given by:

$5/9[1-(2/5)^2-(3/5)^2] + 4/9[1-(2/4)^2-(2/4)^2] = 0.4889$.

Therefore a_1 produces a better split.

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?

Answer: The contingency tables after splitting on attributes *A* and *B* are:

	A = T	*A = F*
+	4	0
−	3	3

	B = T	*B = F*
+	3	1
−	1	5

The overall entropy before splitting is:

$E_{orig} = -0.4 \log 0.4 - 0.6 \log 0.6 = 0.9710$

The information gain after splitting on A is:

$E_{A=T} = -(4/7)\log (4/7) - (3/7)\log (3/7) = 0.9852$.

$E_{A=F} = -(3/3)\log (3/3) - (0/3) \log (0/3) = 0$

$\Delta = E_{orig} - (7/10) E_{A=T} - (3/10) E_{A=F} = 0.2813$.
Similarly, the information gain after splitting on B is given by:

$\Delta = E_{orig} - 4/10 E_{B=T} - 6/10 E_{B=F} = 0.2565$
Therefore, attribute *A* will 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?

Answer: The overall Gini index before splitting is: $G_{orig} = 1 - 0.4^2 - 0.6^2 = 0.48$
The gain in the Gini index after splitting on A is:

$G_{A=T} = 1 - (4/7)^2 - (3/7)^2 = 0.4898$
$G_{A=F} = 1 - (3/3)^2 - (0/3)^2 = 0$
Hence the corresponding gain is equal to $G_{orig} - (7/10)G_{A=T} - (3/10)G_{A=F} = 0.1371$.

Similarly, we can compute the gain after splitting on B, which will be given by:
$G_{orig} - (4/10)G_{B=T} - (6/10) G_{B=F} = 0.1633$.

Therefore, attribute *B* will 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.

Answer: Yes, even though these measures have similar range and monotonous behavior, their respective gains, Δ, which are scaled differences of the measures, do not necessarily behave in the same way, as illustrated by the results in parts (a) and (b).

6. Consider the following set of training

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 described in this chapter. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

Answer:
Splitting Attribute at Level 1.
To determine the test condition at the root node, we need to com-pute the error rates for attributes *X*, *Y* , and *Z*. For attribute *X*, the corresponding counts are:

X	C1	C2
0	60	60
1	40	40

Therefore, the error rate using attribute *X* is (60 + 40)/200 = 0.5. For attribute *Y*, the corresponding counts are:

Y	C1	C2
0	40	60
1	60	40

Therefore, the error rate using attribute *Y* is (40 + 40)/200 = 0.4. For attribute *Z*, the corresponding counts are:

Z	C1	C2
0	30	70
1	70	30

Therefore, the error rate using attribute *Y* is (30 + 30)/200 = 0.3. Since *Z* gives the lowest error rate, it is chosen as the splitting attribute at level 1.

Splitting Attribute at Level 2.
After splitting on attribute *Z*, the subsequent test condition may in-volve either attribute *X* or *Y*. This depends on the training examples distributed to the *Z* = 0 and *Z* = 1 child nodes.

For *Z* = 0, the corresponding counts for attributes *X* and *Y* are the same, as shown in the table below.

X	C1	C2
0	15	45
1	15	25

Y	C1	C2
0	15	45
1	15	25

The error rate in both cases (*X* and *Y*) are (15 + 15)/100 = 0.3.

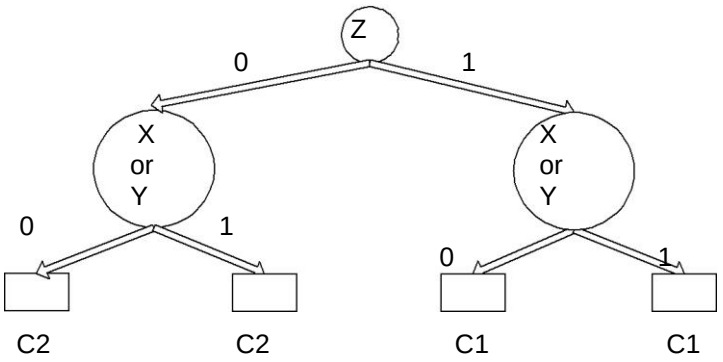
For *Z* = 1, the corresponding counts for attributes *X* and *Y* are shown in the tables below.

X	C1	C2
0	45	15
1	25	15

Y	C1	C2
0	25	15
1	45	15

Although the counts are somewhat different, their error rates remain the same, (15 + 15)/100 = 0.3.

The corresponding two-level decision tree is shown below.



The overall error rate of the induced 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?

Answer: After choosing attribute *X* to be the first splitting attribute, the sub-sequent test condition may involve either attribute *Y* or attribute *Z*.

For *X* = 0, the corresponding counts for attributes *Y* and *Z* are shown in the table below.

Y	C1	C2
0	5	55
1	55	5

Z	C1	C2
0	15	45
1	45	15

The error rate using attributes *Y* and *Z* are 10/120 and 30/120, respectively. Since attribute *Y* leads to a smaller error rate, it provides a better split.

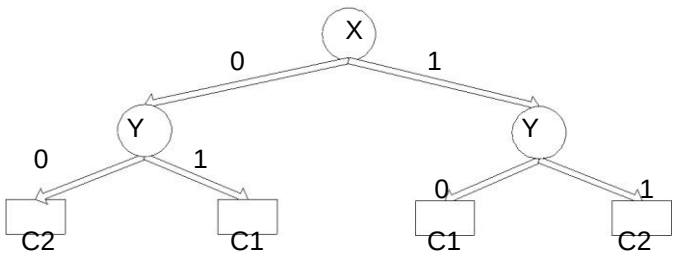
For *X* = 1, the corresponding counts for attributes *Y* and *Z* are shown in the tables below.

Y	C1	C2
0	35	5
1	5	35

Z	C1	C2
0	15	25
1	25	15

The error rate using attributes *Y* and *Z* are 10/80 and 30/80, respectively. Since attribute *Y* leads to a smaller error rate, it provides a better split.

The corresponding two-level decision tree is shown below.



The overall error rate of the induced 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.

Answer:
From the preceding results, the error rate for part (a) is significantly larger than that for part (b). This example shows that a greedy heuristic does not always produce an optimal solution.