2. Consider the training examples shown in Table 4.1 for a binary classificatio problem.

Table 4.1. 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.

Answer: The Gini index for the overall examples is $1 - (5/10)^2 - (5/10)^2 = 0.5$.

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

Answer: The Gini index for the Customer ID attribute is 0.

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

Answer: The gini for Male (of Female) is $1 - 0.4^2 - 0.6^2 = 0.48$. Therefore, the Gini index for the Gender attribute is $0.5 \times 0.48 + 0.5 \times 0.48 = 0.48$.

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

Answer: The gini for Family car is $1-(1/4)^2-(3/4)^2=0.375$, Sports car is 0, and Luxury car is 0.2188. Therefore, the Gini index is 0.1625.

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

Answer: The gini 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. Therefore, the Gini index for Shirt Size attribute is 0.4914.

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

Answer: Car Type because it has the lowest Gini index.

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

Answer: The attribute cannot be used for prediction (it has no predictive power) since new customers are assigned to new Customer IDs.

3. Consider the training examples shown in Table 4.2 for a binary classification problem.

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

Answer: The entropy of the training examples is −4/9 log2(4/9) − 5/9 log2(5/9) = 0.9911.

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

Answer:

The entropy for a_1 is

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

Therefore, the information gain for a_1 is $0.9911 - 0.7616 = 0.2294$.

The entropy for a_2 is

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

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 information gain?

Answer: a1

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

Answer: The error rate for a1 is 2/9 and that for a2 is 4/9 so that a1 is the best split attribute.

(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

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

For attribute a_2, the gini index is

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

Since the gini index for a_1 is smaller, it produces the better split.

Table 4.2. Data set for Exercise 3.

Instance	a_1	a_2	a_3	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	−

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:

$$\begin{aligned} E_{A=T} &= -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.9852 \\ E_{A=F} &= -\frac{3}{3}\log\frac{3}{3} - \frac{0}{3}\log\frac{0}{3} = 0 \\ \Delta &= E_{orig} - 7/10 E_{A=T} - 3/10 E_{A=F} = 0.2813 \end{aligned}$$

The information gain after splitting on B is:

$$\begin{aligned} E_{B=T} &= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113 \\ E_{B=F} &= -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} = 0.6500 \\ \Delta &= E_{orig} - 4/10 E_{B=T} - 6/10 E_{B=F} = 0.2565 \end{aligned}$$

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 before splitting is:

$$G_{orig} = 1 - 0.4^2 - 0.6^2 = 0.48$$

The gain in gini after splitting on A is:

$$
\begin{aligned}
G_{A=T} &= 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898 \\
G_{A=F} &= 1 = \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0 \\
\Delta &= G_{orig} - 7/10 G_{A=T} - 3/10 G_{A=F} = 0.1371
\end{aligned}
$$

The gain in gini after splitting on B is:

$$
\begin{aligned}
G_{B=T} &= 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.3750 \\
G_{B=F} &= 1 = \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.2778 \\
\Delta &= G_{orig} - 4/10 G_{B=T} - 6/10 G_{B=F} = 0.1633
\end{aligned}
$$

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.

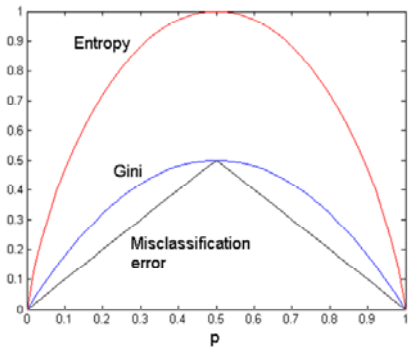


Figure 4.13 Comparison among the impurity measures for binary classification problems.

Answer: Yes, because their respective gains, Δ, do not have the same property.

6. Consider the following set of 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 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:

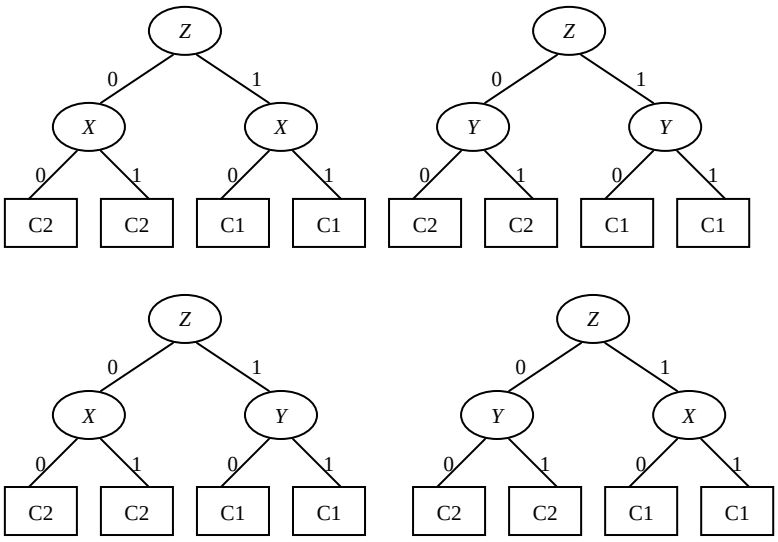
At level 1

The error rate using attribute *X* is (60 + 40)/200 = 0.5; the error rate using attribute *Y* is (40 + 40)/200 = 0.4; the error rate using attribute *Z* is (30 + 30)/200 = 0.3. Since *Z* gives the lowest error rate, it is chosen as the splitting attribute at level 1.

At level 2

For *Z* = 0, the error rate in both cases (*X* and *Y*) are (15 + 15)/100 = 0.3. For *Z* = 1, their error rates remain the same, (15 + 15)/100 = 0.3.

Therefore, the corresponding two-level decision tree can be one of the four possibilities shown below and 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?

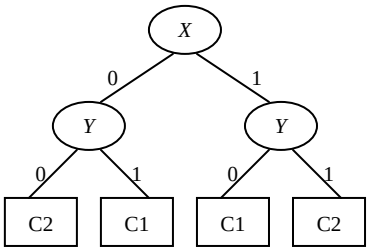
Answer:

After choosing attribute *X* to be the first splitting attribute, the subsequent test condition may involve either attribute *Y* or attribute *Z*.

For *X* = 0, 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 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.

Therefore, the corresponding two-level decision tree is shown below and 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 examples show that a greedy heuristic does not always produce an optimal solution.

7. Consider the data set shown in Table 5.1

Table 5.1. Data set for Exercise 7.

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

(a) Estimate the conditional probabilities for $P(A|+)$, $P(B|+)$, $P(C|+)$, $P(A|-)$, $P(B|-)$, and $P(C|-)$.

Answer:

$P(A = 1|-) = 2/5 = 0.4$, $P(B = 1|-) = 2/5 = 0.4$,
$P(C = 1|-) = 1$, $P(A = 0|-) = 3/5 = 0.6$,
$P(B = 0|-) = 3/5 = 0.6$, $P(C = 0|-) = 0$; $P(A = 1|+) = 3/5 = 0.6$,
$P(B = 1|+) = 1/5 = 0.2$, $P(C = 1|+) = 2/5 = 0.4$, =4/5=0.8
$P(A = 0|+) = 2/5 = 0.4$, $P(B = 0|+) = 4/5 = 0.8$,
$P(C = 0|+) = 3/5 = 0.6$. =1/5=0.2

(b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample $(A = 0, B = 1, C = 0)$ using the naïve Bayes approach.

Answer:

$$= \frac{P(+|A = 0, B = 1, C = 0)}{\frac{P(A = 0, B = 1, C = 0|+) \times P(+)}{P(A = 0, B = 1, C = 0)}}$$

Since $P(A = 0, B = 1, C = 0|+) = P(A = 0|+)P(B = 1|+)P(C = 0|+)$,

$P(+|A = 0, B = 1, C = 0) = 0.4 \times 0.2 \times 0.6 \times 0.5 / P(A = 0, B = 1, C = 0)$. 0.2

Similarly,

$$= \frac{P(-|A = 0, B = 1, C = 0)}{\frac{P(A = 0, B = 1, C = 0|-) \times P(-)}{P(A = 0, B = 1, C = 0)}}$$

Since $P(A = 0, B = 1, C = 0|-) = P(A = 0|-) \times P(B = 1|-) \times P(C = 0|-)$,

$P(-|A = 0, B = 1, C = 0) = 0 / P(A = 0, B = 1, C = 0)$.

Therefore, the class label should be '+'.

8. Consider the data set shown in Table 5.2.

Table 5.2. Data set for Exercise 8.

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

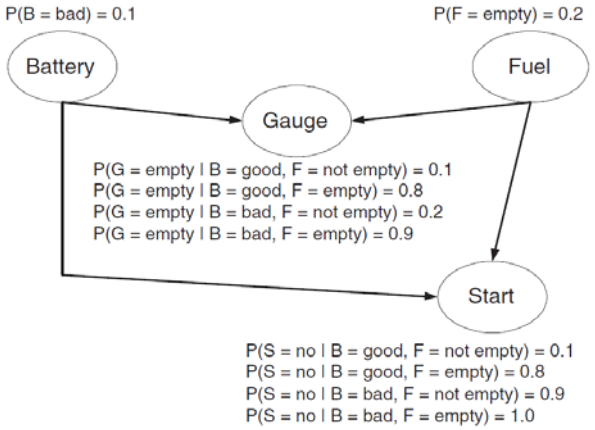


Figure 5.4. Bayesian belief network for Exercise 12.

(a) Estimate the conditional probabilities for $P(A = 1|+)$, $P(B = 1|+)$, $P(C = 1|+)$, $P(A = 1|-)$, $P(B = 1|-)$, and $P(C = 1|-)$ using the same approach as in the previous problem.

Answer:

$$P(A = 1|+) = 0.6$$
$$P(B = 1|+) = 0.4$$
$$P(C = 1|+) = 0.8$$
$$P(A = 1|-) = 0.4$$
$$P(B = 1|-) = 0.4$$
$$P(C = 1|-) = 0.2$$

(b) Use the conditional probabilities in part (a) to predict the class label for a test sample $(A = 1, B = 1, C = 1)$ using the naïve Bayes approach.

Answer: Similar to 7(b), the record is assigned to (+) class.

(c) Compare $P(A = 1)$, $P(B = 1)$, and $P(A = 1, B = 1)$. State the relationships between A and B.

Answer: $P(A = 1) = 0.5$, $P(B = 1) = 0.4$ and $P(A = 1, B = 1) = P(A) \times P(B) = 0.2$.

Therefore, A and B are independent.

12. Given the Bayesian network shown in Figure 5.4, compute the following probabilities:

(a) $P(\text{B} = \text{good}, \text{F} = \text{empty}, \text{G} = \text{empty}, \text{S} = \text{yes})$.

Answer:

$$\begin{aligned} & P(B = good, F = empty, G = empty, S = yes) \\ = \; & P(B = good) \times P(F = empty) \times P(G = empty|B = good, F = empty) \\ & \times P(S = yes|B = good, F = empty) \\ = \; & 0.9 \times 0.2 \times 0.8 \times 0.2 = 0.0288. \end{aligned}$$

(b) $P(\text{B} = \text{bad}, \text{F} = \text{empty}, \text{G} = \text{not empty}, \text{S} = \text{no})$.

Answer:

$$
\begin{aligned}
& P(B = bad, F = empty, G = not\ empty, S = no) \\
= \ & P(B = bad) \times P(F = empty) \times P(G = not\ empty | B = bad, F = empty) \\
& \times P(S = no | B = bad, F = empty) \\
= \ & 0.1 \times 0.2 \times 0.1 \times 1.0 = 0.002.
\end{aligned}
$$

(c) Given that the battery is bad, compute the probability that the car will start.

Answer:

$$
\begin{aligned}
& p(S = yes \mid B = bad) \\
&= p(S = yes, B = bad) \Big/ p(B = bad) \\
&= \sum_{\alpha} p(S = yes, B = bad, F = \alpha) \Big/ p(B = bad) \\
&= \sum_{\alpha} p(S = yes | B = bad, F = \alpha) \times p(B = bad, F = \alpha) \Big/ p(B = bad) \\
&= \sum_{\alpha} p(S = yes | B = bad, F = \alpha) \times p(B = bad) \times p(| F = \alpha) \Big/ p(B = bad) \\
&= \sum_{\alpha} p(S = yes \mid B = bad, F = \alpha) \times p(| F = \alpha) \\
&= (0.1 + 0) \times 0.8 = 0.08
\end{aligned}
$$

18. Following is a data set that contains two attributes, X and Y, and two class labels, "+" and "−". Each attribute can take three different values: 0, 1, or 2.

X	Y	Number of Instances	
		+	−
0	0	0	100
1	0	0	0
2	0	0	100
0	1	10	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

The concept for the "+" class is $Y = 1$ and the concept for the "−" class is $X = 0 \vee X = 2$.

(a) Build a decision tree on the data set. Does the tree capture the "+" and "−" concepts?

Answer:

There are 30 positive and 600 negative examples in the data. Therefore, at the root node, the error rate is

$$E_{orig} = 1 - \max(30/630, 600/630) = 30/630.$$

If we split on X, the gain in error rate is:

	$X=0$	$X=1$	$X=2$
+	10	10	10
−	300	0	300

$$E_{X=0} = 10/310$$
$$E_{X=1} = 0$$
$$E_{X=2} = 10/310$$

$$\Delta_X = E_{orig} - \frac{310}{630}\frac{10}{310} - \frac{10}{630}0 - \frac{310}{630}\frac{10}{310} = 10/630.$$

If we split on Y, the gain in error rate is:

	$Y=0$	$Y=1$	$Y=2$
+	0	30	0
−	200	200	200

$$E_{Y=0} = 0$$
$$E_{Y=1} = 30/230$$
$$E_{Y=2} = 0$$

$$\Delta_Y = E_{orig} - \frac{230}{630}\frac{30}{230} = 0.$$

Therefore, X is chosen to be the first splitting attribute. Since the $X = 1$ child node is pure, it does not require further splitting. We may use attribute Y to split the impure nodes, $X = 0$ and $X = 2$, as follows:

- The $Y = 0$ and $Y = 2$ nodes contain 100 − instances.
- The $Y = 1$ node contains 100 − and 10 + instances.

In all three cases for Y, the child nodes are labeled as −. The resulting concept is

$$\text{class} = \begin{cases} +, & \text{X} = 1; \\ -, & \text{otherwise.} \end{cases}$$

(b) What are the accuracy, precision, recall, and F_1-measure of the decision tree? (Note that precision, recall, and F_1-measure are defined with respect to the "+" class.)

Answer: The confusion matrix on the training data:

		Predicted +	Predicted −
Actual	+	10	20
	−	0	600

$$\text{accuracy} : \frac{610}{630} = 0.9683$$
$$\text{precision} : \frac{10}{10} = 1.0$$
$$\text{recall} : \frac{10}{30} = 0.3333$$
$$\text{F} - \text{measure} : \frac{2 * 0.3333 * 1.0}{1.0 + 0.3333} = 0.5$$

(c) Build a new decision tree with the following cost function:

$$C(i,j) = \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{if } i = +, j = -; \\ \frac{\text{Number of } - \text{ instances}}{\text{Number of } + \text{ instances}}, & \text{if } i = -, j = +. \end{cases}$$

(Hint: only the leaves of the old decision tree need to be changed.) Does the decision tree capture the "+" concept?

Answer:

The cost matrix can be summarized as follows:

		Predicted +	Predicted −
Actual	+	0	600/30=20
	−	1	0

The decision tree in part (a) has 7 leaf nodes, $X = 1$, $X = 0 \wedge Y = 0$, $X = 0 \wedge Y = 1$, $X = 0 \wedge Y = 2$, $X = 2 \wedge Y = 0$, $X = 2 \wedge Y = 1$, and $X = 2 \wedge Y = 2$. Only $X = 0 \wedge Y = 1$ and $X = 2 \wedge Y = 1$ are impure nodes. The cost of misclassifying these impure nodes as positive class is:

$$10 * 0 + 1 * 100 = 100$$

while the cost of misclassifying them as negative class is:

$$10 * 20 + 0 * 100 = 200.$$

These nodes are therefore labeled as $+$.

The resulting concept is

$$\text{class} = \begin{cases} +, & X = 1 \vee (X = 0 \wedge Y = 1) \vee (X = 2 \wedge Y = 2); \\ -, & \text{otherwise.} \end{cases}$$

(d) What are the accuracy, precision, recall, and F_1-measure of the new decision tree?

Answer:

The confusion matrix of the new tree

		Predicted $+$	Predicted $-$
Actual	$+$	30	0
	$-$	200	400

$$\text{accuracy} : \frac{430}{630} = 0.6825$$

$$\text{precision} : \frac{30}{230} = 0.1304$$

$$\text{recall} : \frac{30}{30} = 1.0$$

$$\text{F} - \text{measure} : \frac{2 * 0.1304 * 1.0}{1.0 + 0.1304} = 0.2307$$