The preceding tree cannot be simplified.

2. Consider the training examples shown in Table 4.1 for a binary classification 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:

Gini $= 1 - 2 \times 0.5^2 = 0.5$.

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

Answer:

The gini for each `Customer ID` value is 0. Therefore, the overall gini for `Customer ID` is 0.

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

Answer:

The gini for `Male` is $1 - 2 \times 0.5^2 = 0.5$. The gini for `Female` is also 0.5. Therefore, the overall gini for `Gender` is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

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	−

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

Answer:

The gini for `Family` car is 0.375, `Sports` car is 0, and `Luxury` car is 0.2188. The overall gini 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. The overall gini 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 among the three attributes.

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

Answer:

The attribute has no predictive power since new customers are assigned to new `Customer ID`s.

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:

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

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

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

For attribute a_2, the corresponding counts and probabilities are:

a_2	+	-
T	2	3
F	2	2

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:

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: error rate = 2/9.

For attribute a_2: 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

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

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.

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

Answer:

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:

$$E(Y)=-\sum_{j=1}^{c}P(y_j)\log_2 P(y_j)=\sum_{j=1}^{c}\sum_{i=1}^{k}P(x_i,y_j)\log_2 P(y_j), \tag{4.1}$$

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

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

$$E(Y|x_i)=-\sum_{j=1}^{c}P(y_j|x_i)\log_2 P(y_j|x_i) \tag{4.2}$$

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:

$$\begin{aligned} E(Y|X) &= \sum_{i=1}^{k} P(x_i)E(Y|x_i) \\ &= -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i)P(y_j|x_i)\log_2 P(y_j|x_i) \\ &= -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i,y_j)\log_2 P(y_j|x_i), \end{aligned} \tag{4.3}$$

where we have used a known fact from probability theory that $P(x_i, y_j) = P(y_j|x_i) \times P(x_i)$. 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., $E(Y|X) - E(Y)$, using Equations 4.1 and 4.3:

$$\begin{aligned} & E(Y|X) - E(Y) \\ = & -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i,y_j)\log_2 P(y_j|x_i) + \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i,y_j)\log_2 P(y_j) \\ = & \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i,y_j)\log_2 \frac{P(y_j)}{P(y_j|x_i)} \\ = & \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i,y_j)\log_2 \frac{P(x_i)P(y_j)}{P(x_i,y_j)} \end{aligned} \tag{4.4}$$

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

$$\sum_{k=1}^{d} a_k \log(z_k) \leq \log\big(\sum_{k=1}^{d} a_k z_k\big), \tag{4.5}$$

subject to the condition that $\sum_{k=1}^{d} a_k = 1$. This property is a special case of a more general theorem involving convex functions (which include 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.

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:

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

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:

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

The gain in gini after splitting on B is:

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

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.

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 examples.