# 4

# Classification: Basic Concepts, Decision Trees, and Model Evaluation

1. 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?

| 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 |

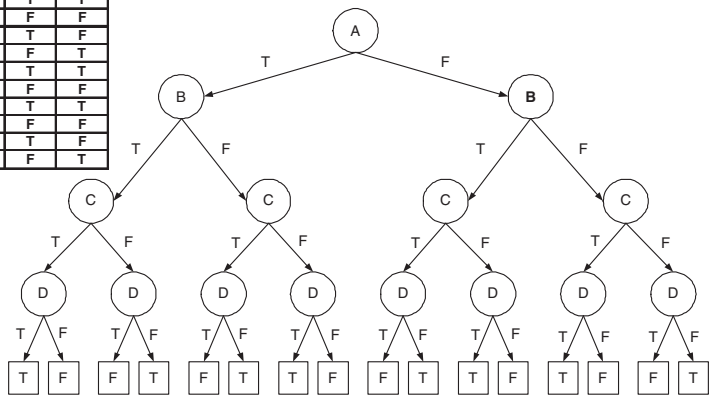



**Figure 4.1.** Decision tree for parity function of four Boolean attributes.

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:

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

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.

**Answer:**

Yes, even though these measures have similar range and monotonous behavior, their respective gains, $\Delta$, 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.

| $X$ | $Y$ | $Z$ | 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 compute 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 involve 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 $(15 + 15)/100 = 0.3$.

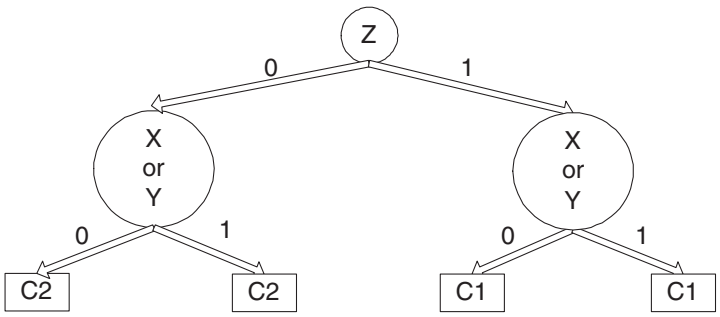
For $Z = 1$, the corresponding counts for attributes $X$ and $Y$ are shown in the tables below.

| $X$ | 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 subsequent 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.

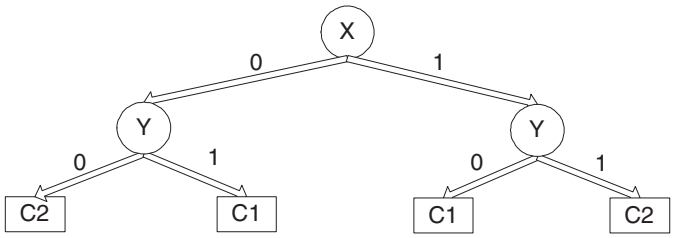
For $X = 1$, the corresponding counts for attributes $Y$ and $Z$ are shown in the tables below.

| $Y$ | 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 examples shows that a greedy heuristic does not always produce an optimal solution.

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 $\frac{20}{100}$.)

(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:
$E_{orig} = \frac{25}{50}$.
After splitting on attribute $A$, the gain in error rate is:

| | $A=T$ | $A=F$ |
|---|---|---|
| + | 25 | 0 |
| − | 0 | 25 |

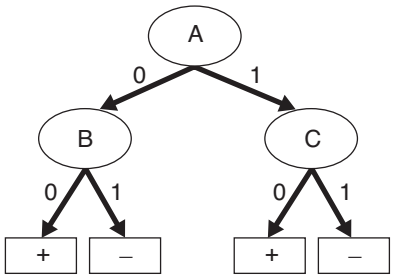
$$E_{A=T} = 0$$
$$E_{A=F} = 0$$
$$\Delta_A = \frac{25}{50}$$

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.




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 | + |

**Figure 4.2.** Decision tree and data sets for Exercise 8.

For the $C = F$ child, the error rate before splitting is: $E_{orig} = \frac{25}{50}$.
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.
The greedy heuristic does not necessarily lead to the best tree.

8. Consider the decision tree shown in Figure 4.2.

(a) Compute the generalization error rate of the tree using the optimistic approach.

**Answer:**

According to the optimistic approach, the generalization error rate is $3/10 = 0.3$.

(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.)

**Answer:**

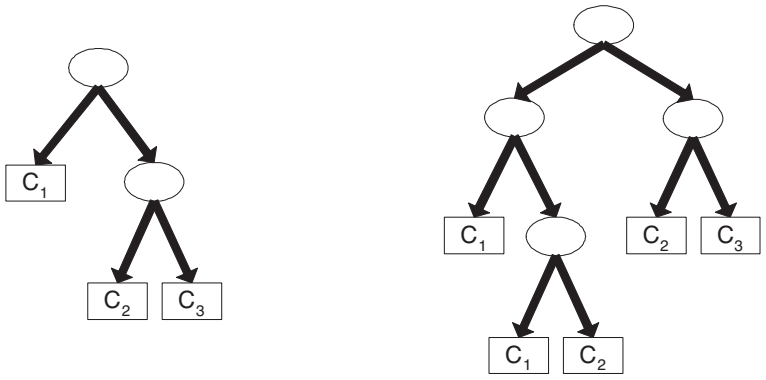
According to the pessimistic approach, the generalization error rate is $(3 + 4 \times 0.5)/10 = 0.5$.

(c) Compute the generalization error rate of the tree using the validation set shown above. This approach is known as **reduced error pruning**.

**Answer:**

According to the reduced error pruning approach, the generalization error rate is $4/5 = 0.8$.

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%.

(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%.

(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 error estimate is:

$$\frac{1}{b}\sum_{i=1}^{b}\Big[0.632\times 0.5+0.368\times 1\Big]=0.684.$$

(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 boostrap method.

11. 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:

$$Z=\frac{p_A-p_B}{\sqrt{\frac{2p(1-p)}{N}}}.$$

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

Table 4.3 compares the accuracies of three different classifiers, decision tree classifiers, naïve Bayes classifiers, and support vector machines, on various data sets. (The latter two classifiers are described in Chapter 5.)

**Table 4.3.** Comparing the accuracy of various classification methods.

| Data Set | Size ($N$) | Decision Tree (%) | naïve 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 |
| Zoo | 101 | 93.07 | 93.07 | 96.04 |

**Answer:**

A summary of the relative performance of the classifiers is given below:

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

12. Let $X$ be a binomial random variable with mean $Np$ and variance $Np(1-p)$. Show that the ratio $X/N$ also has a binomial distribution with mean $p$ and variance $p(1-p)/N$.

**Answer:** 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:

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

$$\begin{aligned}
\text{Variance}, E[(r - E[r])^2] &= E[(X/N - E[X/N])^2] \\
&= E[(X - E[X])^2]/N^2 \\
&= Np(1-p)/N^2 \\
&= p(1-p)/N
\end{aligned}$$