Owns home?	*Married*	*Gender*	*Employed*	*Credit rating*	*Risk class*
Yes	Yes	Male	Yes	*A*	*B*
No	No	Female	Yes	*A*	*A*
Yes	Yes	Female	Yes	*B*	*C*
Yes	No	Male	No	*B*	*B*
No	Yes	Female	Yes	*B*	*C*
No	No	Female	Yes	*B*	*A*
No	No	Male	No	*B*	*B*
Yes	No	Female	Yes	*A*	*A*
No	Yes	Female	Yes	*A*	*C*
Yes	Yes	Female	Yes	*A*	*C*

There are 10 ($s = 10$) samples and three classes. The frequencies of these classes are:

$$A = 3$$

$$B = 3$$

$$C = 4$$

Information in the data due to uncertainty of outcome regarding the risk class each person belongs to is given by

$$I = -(3/10)\ \log(3/10) - (3/10)\ \log(3/10) - (4/10)\ \log(4/10) = 1.57$$

Let us now consider using each attribute in turn as a candidate to split the sample.

1. Attribute "Owns Home"

Value = Yes. There are five applicants who own their home. They are in classes $A = 1$, $B = 2$, $C = 2$.

Value = No. There are five applicants who do not own their home. They are in classes $A = 2$, $B = 1$, $C = 2$.

Given the above values, it does not appear as if this attribute will reduce the uncertainty by much. Let us compute the information gain by using this attribute. We divide persons into those who own their home and those who do not. Computing information for each of these two subtrees,

$$I(\text{Yes}) = I(y) = -(1/5)\ \log(1/5) - (2/5)\ \log(2/5) - (2/5)\ \log(2/5) = 1.52$$

$$I(\text{No}) = I(n) = -(2/5)\ \log(2/5) - (1/5)\ \log(1/5) - (2/5)\ \log(2/5) = 1.52$$

Total information of the two subtrees = $0.5I(y) + 0.5I(n) = 1.52$

2. Attribute "Married"

There are five applicants that are married and five that are not.

$$\text{Value} = \text{Yes has } A = 0,\ B = 1,\ C = 4, \text{ total } 5$$

$$\text{Value} = \text{No has } A = 3,\ B = 2,\ C = 0, \text{ total } 5$$

Looking at the values above, it appears that this attribute will reduce the uncertainty by more than the last attribute. Computing the information gain by using this attribute, we have

$$I(y) = -(1/5)\ \log(1/5) - (4/5)\ \log(4/5) = 0.722$$

$$I(n) = -(3/5)\ \log(3/5) - (2/5)\ \log(2/5) = 0.971$$

Information of the subtrees $= 0.5I(y) + 0.5I(n) = 0.846$

3. Attribute "Gender"

There are three applicants that are male and seven are female.

$$\text{Value} = \text{Male has } A = 0,\ B = 3,\ C = 0, \text{ total } 3$$

$$\text{Value} = \text{Female has } A = 3,\ B = 0,\ C = 4, \text{ total } 7$$

The values above show that the uncertainty is reduced even more by using this attribute since for Value = Male we have only one class. Let us compute the information gain by using this attribute.

$$I(\text{Male}) = 0$$

$$I(\text{Female}) = -(3/7)\ \log(3/7) - (4/7)\ \log(4/7) = 0.985$$

Total information of the subtrees $= 0.3I(\text{Male}) + 0.7I(\text{Female}) = 0.69$

4. Attribute "Employed"

There are eight applicants that are employed and two that are not.

$$\text{Value} = \text{Yes has } A = 3,\ B = 1,\ C = 4, \text{ total } 8$$

$$\text{Value} = \text{No has } A = 0,\ B = 2,\ C = 0, \text{ total } 2$$

The values above show that this attribute will reduce uncertainty but most attribute values are Yes while the No value leads to only one class. Computing the information gain by using this attribute, we have

$$I(y) = -(3/8)\ \log(3/8) - (1/8)\ \log(1/8) - (4/8)\ \log(4/8) = 1.41$$

$$I(n) = 0$$

Total information of the subtrees $= 0.8I(y) + 0.2I(n) = 1.12$

5. Attribute "Credit Rating"

There are five applicants that have credit rating A and five that have B.

$$\text{Value} = A \text{ has } A = 2,\ B = 1,\ C = 2, \text{ total } 5$$

$$\text{Value} = B \text{ has } A = 1,\ B = 2,\ C = 2, \text{ total } 5$$

Looking at the values above, we can see that this is like the first attribute that does not reduce uncertainty by much. The information gain for this attribute is the same as for the first attribute.

$$I(A) = -(2/5)\ \log(2/5) - (1/5)\ \log(1/5) - (2/5)\ \log(2/5) = 1.52$$

$$I(B) = -(1/5)\ \log(1/5) - (2/5)\ \log(2/5) - (2/5)\ \log(2/5) = 1.52$$

Total information of the subtrees = $0.5I(A) + 0.5I(B) = 1.52$

The values for information gain can now be computed. See Table 3.3.

Table 3.3 Information gain for the five attributes

Potential split attribute	*Information before split*	*Information after split*	*Information gain*
Owns Home	1.57	1.52	0.05
Married	1.57	0.85	0.72
Gender	1.57	0.69	0.88
Employed	1.57	1.12	0.45
Credit Rating	1.57	1.52	0.05

Hence the largest information gain is provided by the attribute "Gender" and that is the attribute that is used for the split.

Now we can reduce the data by removing the attribute Gender and removing the class B since all Class B have Gender = Male. See Table 3.4.

Table 3.4 Data after removing attribute "Gender" and Class B

Owns home?	*Married*	*Employed*	*Credit rating*	*Risk class*
No	No	Yes	*A*	*A*
Yes	Yes	Yes	*B*	*C*
No	Yes	Yes	*B*	*C*
No	No	Yes	*B*	*A*
Yes	No	Yes	*A*	*A*
No	Yes	Yes	*A*	*C*
Yes	Yes	Yes	*A*	*C*

The information in this data of two classes due to uncertainty of outcome regarding the class each person belongs to is given by

$$I = -(3/7)\ \log(3/7) - (4/7)\ \log(4/7) = 1.33$$

Let us now consider each attribute in turn as a candidate to split the sample.

1. Attribute "Owns Home"

Value = Yes. There are three applicants who own their home, who are in classes $A = 1$ and $C = 2$.

Value = No. There are four applicants who do not own their home, who are in classes $A = 2$ and $C = 2$.

Given the above values, it does not appear as if this attribute will reduce the uncertainty by much. Computing information for each of these two subtrees:

$$I(\text{Yes}) = I(y) = -(1/3)\ \log(1/3) - (2/3)\ \log(2/3) = 0.92$$

$$I(\text{No}) = I(n) = -(2/4)\ \log(2/4) - (2/4)\ \log(2/4) = 1.00$$

Total information of the two subtrees = $(3/7)I(y) + (4/7)I(n) = 0.96$

2. Attribute "Married"

There are four applicants that are married and three that are not.

Value = Yes has $A = 0$, $C = 4$, total 4

Value = No has $A = 3$, $C = 0$, total 3

Looking at the values above, it appears that this attribute will reduce the uncertainty by a lot more than the last attribute since for each value the persons belong to only one class and therefore information is zero.

$$I(y) = -(4/4)\ \log(4/4) = 0.0$$

$$I(n) = -(3/3)\ \log(3/3) = 0.0$$

Information of the subtrees = 0.0

There is no need to consider other attributes now, since no other attribute can be better. The split attribute therefore is "Married" and we now obtain the following decision tree in Figure 3.2 which concludes this very simple example. It should be noted that a number of attributes that were available in the data were not required in classifying it.

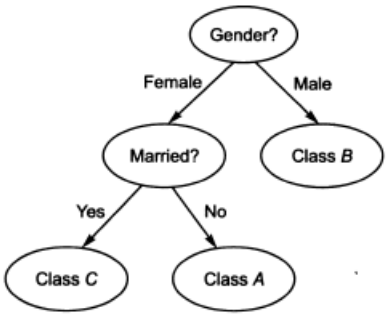

Figure 3.2 Decision tree for Example 3.1.

The decision tree may be written as a set of rules like "if Gender = Male then Class *B*". We will discuss rule generation in Section 3.7.

Note: The attribute comparison based on information gain is not necessarily always fair. The approach tends to favour attributes that have a large number of values. For example, if there was an attribute that had a large number of values such that each value resulted in only one element of the dataset then that attribute will be selected although it might not be the optimum split attribute in that it leads to a very wide tree with depth one. Such a tree may be even quite inaccurate for unseen data.