Consider the data in Table 8.11. The target variable is salary. Start by discretizing salary as follows:

a. Less than $35,000 Level 1
b. $35,000 to less than $45,000 Level 2
c. $45,000 to less than $55,000 Level 3
d. Above $55,000 Level 4

The table below includes the additional field called “Level” for the discretized values of Salary.

Occupation	Gender	Age	Salary	Level
Service	Female	45	$48,000	3
Service	Male	25	$25,000	1
Service	Male	33	$35,000	2
Management	Male	25	$45,000	3
Management	Female	35	$65,000	4
Management	Male	26	$45,000	3
Management	Female	45	$70,000	4
Sales	Female	40	$50,000	3
Sales	Male	30	$40,000	2
Staff	Female	50	$40,000	2
	Male	25	$25,000	1

Table 4.1. Records with variable Salary discretized to Levels 1-4

TABLE 8.11 Decision tree data

Occupation	Gender	Age	Salary
Service	**Female**	**45**	**$48,000**
	Male	**25**	**$25,000**
	Male	**33**	**$35,000**
Management	**Male**	**25**	**$45,000**
	Female	**35**	**$65,000**
	Male	**26**	**$45,000**
	Female	**45**	**$70,000**
Sales	**Female**	**40**	**$50,000**
	Male	**30**	**$40,000**
Staff	**Female**	**50**	**$40,000**
	Male	**25**	**$25,00**

5. **Construct a classification and regression tree to classify salary based on the other variables. Do as much as you can by hand, before turning to the software.**

The table below shows all possible candidate splits that could occur at the root note using predictor variables *Occupation*, *Gender*, and *Age*.

Candidate	L Node	R Node
1	Occupation = Service	Occupation = Management or Sales or Staff
2	Occupation = Management	Occupation = Service or Sales or Staff
3	Occupation = Sales	Occupation = Service or Management or Staff
4	Occupation = Staff	Occupation = Service or Management or Sales
5	Occupation = Service or Management	Occupation = Sales or Staff
6	Occupation = Service or Sales	Occupation = Management or Staff
7	Occupation = Service or Staff	Occupation = Management or Sales
8	Gender = Male	Gender = Female
9	Age <= 25	Age > 25
10	Age <= 35	Age > 35
11	Age <= 45	Age > 45

Table 5.2. Candidate splits for CART algorithm using Occupation, Gender, and Age as predictors

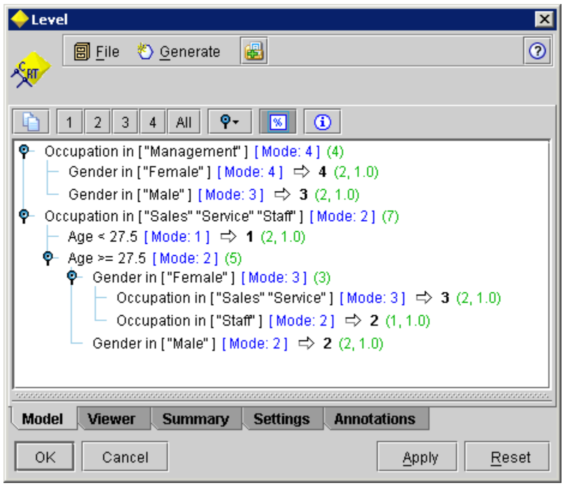
Next, using the candidate splits defined in the table above, the optimality measure for each split is calculated.

Candidate	P_L	P_R	$P(j\|t_L)$	$P(j\|t_R)$	$2P_LP_R$	$Q(s\|t)$	$\Phi(s\|t)$
1	0.273	0.727	1: 0.333 2: 0.333 3: 0.333 4: 0	1: 0.125 2: 0.25 3: 0.375 4: 0.25	0.397	0.583	0.231
2	0.364	0.636	1: 0 2: 0 3: 0.50 4: 0.50	1: 0.286 2: 0.429 3: 0.286 4: 0	0.463	1.429	0.662
3	0.182	0.818	1: 0 2: 0.50 3: 0.50 4: 0	1: 0.222 2: 0.222 3: 0.333 4: 0.222	0.298	0.889	0.265
4	0.182	0.818	1: 0.50 2: 0.50 3: 0 4: 0	1: 0.111 2: 0.222 3: 0.444 4: 0.222	0.298	1.333	0.397
5	0.636	0.364	1: 0.143 2: 0.143 3: 0.429 4: 0.286	1: 0.25 2: 0.50 3: 0.25 4: 0	0.463	0.929	0.430
6	0.455	0.545	1: 0.20 2: 0.40 3: 0.40 4: 0	1: 0.167 2: 0.167 3: 0.333 4: 0.333	0.496	0.666	0.330
7	0.455	0.545	1: 0.40 2: 0.40 3: 0.20 4: 0	1: 0 2: 0.167 3: 0.50 4: 0.333	0.496	1.266	0.628
8	0.545	0.455	1: 0.333 2: 0.333 3: 0.333 4: 0	1: 0 2: 0.20 3: 0.40 4: 0.40	0.496	0.933	0.463
9	0.273	0.727	1: 0.667 2: 0 3: 0.333 4: 0	1: 0 2: 0.375 3: 0.375 4: 0.25	0.397	1.334	0.530
10	0.636	0.364	1: 0.286 2: 0.286 3: 0.286 4: 0.143	1: 0 2: 0.25 3: 0.50 4: 0.25	0.463	0.643	0.298
11	0.909	0.091	1: 0.20 2: 0.20 3: 0.40 4: 0.20	1: 0 2: 1 3: 0 4: 0	0.165	1.6	0.264

Table 5.2. Optimality measures for each candidate split

According to the optimality measure, candidate 2 achieves the best split with a value of **0.662**. This occurs at the split *Occupation* = Management (L) and *Occupation* = Service or Sales or Staff (R). The second best split is achieved using candidate 7 with an optimality measure equal to **0.628**.

The records in Table 4.1 are placed in a comma-separated file that is read by SPSS Modeler. The decision tree produced by CART is shown next:



```
Occupation in ["Management"] [Mode: 4] (4)
    Gender in ["Female"] [Mode: 4] ⇨ 4 (2, 1.0)
    Gender in ["Male"] [Mode: 3] ⇨ 3 (2, 1.0)
Occupation in ["Sales" "Service" "Staff"] [Mode: 2] (7)
    Age < 27.5 [Mode: 1] ⇨ 1 (2, 1.0)
    Age >= 27.5 [Mode: 2] (5)
        Gender in ["Female"] [Mode: 3] (3)
            Occupation in ["Sales" "Service"] [Mode: 3] ⇨ 3 (2, 1.0)
            Occupation in ["Staff"] [Mode: 2] ⇨ 2 (1, 1.0)
        Gender in ["Male"] [Mode: 2] ⇨ 2 (2, 1.0)
```

Figure 5.3. CART model generated by SPSS Modeler

The results generated by CART agree with the calculations performed by hand. That is, the root-level split occurs where *Occupation* = Management, or *Occupation* = {Sales, Service, Staff}.

6. Construct a C4.5 decision tree to classify salary based on the other variables. Do as much as you can by hand, before turning to the software.

For C4.5, the list of candidate splits for the root node are shown below.

Candidate	Attribute = Value	Branches

1	Occupation	Occupation = Service
		Occupation = Management
		Occupation = Sales
		Occupation = Staff

2	Gender	Gender = Male
		Gender = Female

3	Age <= 25	Age <= 25 Age > 25
4	Age <= 35	Age <= 35 Age > 35
5	Age <= 45	Age <= 45 Age > 45

Table 6.4. Candidate splits for C4.5 algorithm using Occupation, Gender, and Age as predictors

Using the records defined in Table 4.1, the Information Gain for the 5 candidate splits is calculated. First, however, the Entropy is calculated for the data set before splitting occurs.

Entropy Calculation:

$$P_{Level1} = \frac{2}{11}, P_{Level2} = \frac{3}{11}, P_{Level3} = \frac{4}{11}, P_{Level4} = \frac{2}{11}$$

$$H(T) = -\frac{2}{11}\log_2\left(\frac{2}{11}\right) - \frac{3}{11}\log_2\left(\frac{3}{11}\right) - \frac{4}{11}\log_2\left(\frac{4}{11}\right) - \frac{2}{11}\log_2\left(\frac{2}{11}\right) = 1.936$$

Candidate 1:

$$P_{Service} = \frac{3}{11}, P_{Management} = \frac{4}{11}, P_{Sales} = \frac{2}{11}, P_{Staff} = \frac{2}{11}$$

$$H_{Occupation}(Service) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right) = 1.585$$

$$H_{Occupation}(Management) = -\frac{0}{4}\log_2\left(\frac{0}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1.0$$

$$H_{Occupation}(Sales) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) = 1.0$$

$$H_{Occupation}(Staff) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) = 1.0$$

$$H_{Occupation}(T) = \frac{3}{11}(1.585) + \frac{4}{11}(1.0) + \frac{2}{11}(1.0) + \frac{2}{11}(1.0) = 1.160$$

$$1.936 - 1.160 = 0.776\,bits$$

Candidate 2:

$$P_{Male} = \frac{6}{11}, P_{Female} = \frac{5}{11}$$

$$H_{Gender}(Male) = -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{0}{6}\log_2\left(\frac{0}{6}\right) = 1.585$$

$$H_{Gender}(Female) = -\frac{0}{5}\log_2\left(\frac{0}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 1.522$$

$$H_{Gender}(T) = \frac{6}{11}(1.585) + \frac{5}{11}(1.522) = 1.556$$

$$1.936 - 1.556 = 0.38 bits$$

Candidate 3:

$$P_{Age \le 25} = \frac{3}{11}, P_{Age > 25} = \frac{8}{11}$$

$$H_{Age}(\le 25) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right) = 0.918$$

$$H_{Age}(> 25) = -\frac{0}{8}\log_2\left(\frac{0}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 1.531$$

$$H_{Gender}(T) = \frac{3}{11}(0.918) + \frac{8}{11}(1.531) = 1.364$$

$$1.936 - 1.364 = 0.572 bits$$

Candidate 4:

$$P_{Age\leq 35}=\frac{7}{11},P_{Age>35}=\frac{4}{11}$$
$$H_{Age}(\leq 35)=-\frac{2}{7}\log_2\left(\frac{2}{7}\right)-\frac{2}{7}\log_2\left(\frac{2}{7}\right)-\frac{2}{7}\log_2\left(\frac{2}{7}\right)-\frac{1}{7}\log_2\left(\frac{1}{7}\right)=1.950$$
$$H_{Age}(>35)=-\frac{0}{4}\log_2\left(\frac{0}{4}\right)-\frac{1}{4}\log_2\left(\frac{1}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)-\frac{1}{4}\log_2\left(\frac{1}{4}\right)=1.5$$

$$H_{Age}(T)=\frac{7}{11}(1.950)+\frac{4}{11}(1.5)=1.786$$
$$1.936-1.786=0.15\,bits$$

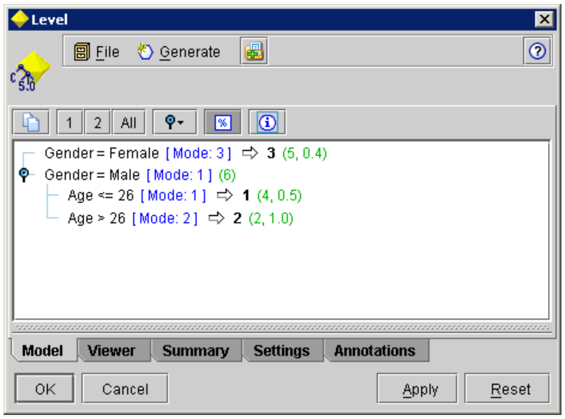
Candidate 5:

$$P_{Age\leq 45}=\frac{10}{11},P_{Age>45}=\frac{1}{11}$$
$$H_{Age}(\leq 35)=-\frac{2}{10}\log_2\left(\frac{2}{10}\right)-\frac{2}{10}\log_2\left(\frac{2}{10}\right)-\frac{4}{10}\log_2\left(\frac{4}{10}\right)-\frac{2}{10}\log_2\left(\frac{2}{10}\right)=1.922$$
$$H_{Age}(>35)=-\frac{0}{1}\log_2\left(\frac{0}{1}\right)-\frac{1}{1}\log_2\left(\frac{1}{1}\right)-\frac{0}{1}\log_2\left(\frac{0}{1}\right)-\frac{0}{1}\log_2\left(\frac{0}{1}\right)=0$$

$$H_{Age}(T)=\frac{10}{11}(1.922)+\frac{1}{11}(0)=1.747$$
$$1.936-1.747=0.189\,bits$$

All five candidate splits are evaluated and Information Gain is maximized to 0.776 bits when splitting on the *Occupation* attribute at the root-level node, according to the calculations performed above.

Now, the records are classified using the C5.0 algorithm in Clementine. Results, using default options, are shown below:



```
Gender = Female [ Mode: 3 ] ⇨ 3 (5, 0.4)
Gender = Male [ Mode: 1 ] (6)
    Age <= 26 [ Mode: 1 ] ⇨ 1 (4, 0.5)
    Age > 26 [ Mode: 2 ] ⇨ 2 (2, 1.0)
```

Figure 6.5. C5.0 model generated by SPSS Modeler

C5.0 splits on the *Gender* attribute at the root level. Interestingly, these results do not agree with the calculations performed manually, above. Perhaps, one or more options such as setting the "Minimum records per child branch" must have a non-default value specified.

7. **Compare the two decision trees and discuss the benefits and drawbacks of each.**

C4.5 has the benefit that it is able to create branches for each categorical attribute value at a decision node split. Depending on the characteristics of the training set, this feature may likely improve classification accuracy. Conversely, this functionality may also suffer from the drawback of creating overly "bushy" trees. CART is limited in its ability to only perform binary splits, which may result in sub-optimal classification accuracy, as compared to results produced by C4.5.

Both decision trees have the benefit of being extremely efficient, while also producing output that is easily interpretable by an analyst. In this way, it is easier to learn what attributes are most important when the algorithm is applied to new domains.

8. Generate the full set of decision rules for the CART decision tree.

The decision rules, as generated by CART, are shown below:

```
Rules for 1 - contains 1 rule(s)
Rule 1 for 1 (2, 1.0)
if Occupation in [ "Sales" "Service" "Staff" ] and Age < 27.5 then 1
Rules for 2 - contains 2 rule(s)
Rule 1 for 2 (1, 1.0)
if Occupation in [ "Sales" "Service" "Staff" ] and Age >= 27.5 and Gender in [ "Female" ]
and Occupation in [ "Staff" ] then 2
Rule 2 for 2 (2, 1.0)
if Occupation in [ "Sales" "Service" "Staff" ] and Age >= 27.5 and Gender in [ "Male" ]
then 2
Rules for 3 - contains 2 rule(s)
Rule 1 for 3 (2, 1.0)
if Occupation in [ "Management" ] and Gender in [ "Male" ] then 3
Rule 2 for 3 (2, 1.0)
if Occupation in [ "Sales" "Service" "Staff" ] and Age >= 27.5 and Gender in [ "Female" ]
and Occupation in [ "Sales" "Service" ] then 3
Rules for 4 - contains 1 rule(s)
Rule 1 for 4 (2, 1.0)
if Occupation in [ "Management" ] and Gender in [ "Female" ] then 4
Default: 3
```

9. **Generate the full set of decision rules for the C4.5 decision tree.**

The decision rules below are generated using C5.0, in Clementine:

```
Rules for 1 - contains 1 rule(s)
Rule 1 for 1 (4, 0.5)
if Gender = Male and Age <= 26 then 1
Rules for 2 - contains 1 rule(s)
Rule 1 for 2 (2, 1.0)
if Gender = Male and Age > 26 then 2
Rules for 3 - contains 1 rule(s)
Rule 1 for 3 (5, 0.4)
if Gender = Female then 3
Default: 3
```

10. **Compare the two sets of decision rules and discuss the benefits and drawbacks of each.**

This time CART grows a slightly more complex tree, as compared to C5.0. All rules in CART result in 100% confidence, while C5.0 has two rules with 40% and 50% confidence. Most likely, if the pruning or records per child node options are further analyzed, a tree of greater depth and higher accuracy will be created.

HANDS-ON ANALYSIS