CS 229 - Machine Learning
Xiangliang Zhang
Computer Science (CS)/Statistics (STAT) Program
Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

HOMEWORK V

Henrique Aparecido Laureano

Spring Semester 2018

Contents

Question 1

Consider the training data shown in Table 1:

Table 1: Data set for decision tree classification.

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

Construct a decision tree by splitting based on the gain in the *Gini index* or *Gain Ratio* (Hint: if meeting an outlier sample when constructing the tree, you can stop splitting if the splitting is not helpful to reach pure class at children nodes. You can make the parent node as a leaf node, whose class label is the majority class of all samples there.)

Solution:

Gini index:

$$\text{GINI}(t) = 1 - \sum_j p(j \mid t)^2, \quad p(j \mid t) \text{ is the relative frequency of class } j \text{ at node } t.$$

For Class we have

$$\text{GINI(Class)} = 1 - (10/20)^2 - (10/20)^2 = 0.5.$$

To select the root node we compute the GAIN for the three nodes

$$\text{GAIN} = \text{GINI(node)} - \sum_{i=1}^{k} \frac{n_i}{n} \text{GINI}(i).$$

Gender:

$$\begin{aligned}\text{GAIN} &= 0.5 - \frac{10}{20}(1 - (4/10)^2 - (6/10)^2) - \frac{10}{20}(1 - (6/10)^2 - (4/10)^2) \\ &= 0.5 - \frac{20}{20}(1 - (4/10)^2 - (6/10)^2) \\ &= 0.02\end{aligned}$$

Car Type:

$$\begin{aligned}\text{GAIN} &= 0.5 - \frac{4}{20}(1 - (1/4)^2 - (3/4)^2) - \frac{8}{20}(1 - (1/8)^2 - (7/8)^2) - \frac{8}{20}(1 - (8/8)^2 - (0/8)^2) \\ &= 0.3375\end{aligned}$$

Shirt Size:

$$\begin{aligned}\text{GAIN} = 0.5 &- \frac{4}{20}(1 - (2/4)^2 - (2/4)^2) - \frac{4}{20}(1 - (2/4)^2 - (2/4)^2) \\ &- \frac{7}{20}(1 - (3/7)^2 - (4/7)^2) - \frac{5}{20}(1 - (3/5)^2 - (2/5)^2) \\ &= 0.008571429\end{aligned}$$

The biggest GAIN is obtained with Car Type, so this is selected as the root node. See Figure 1.

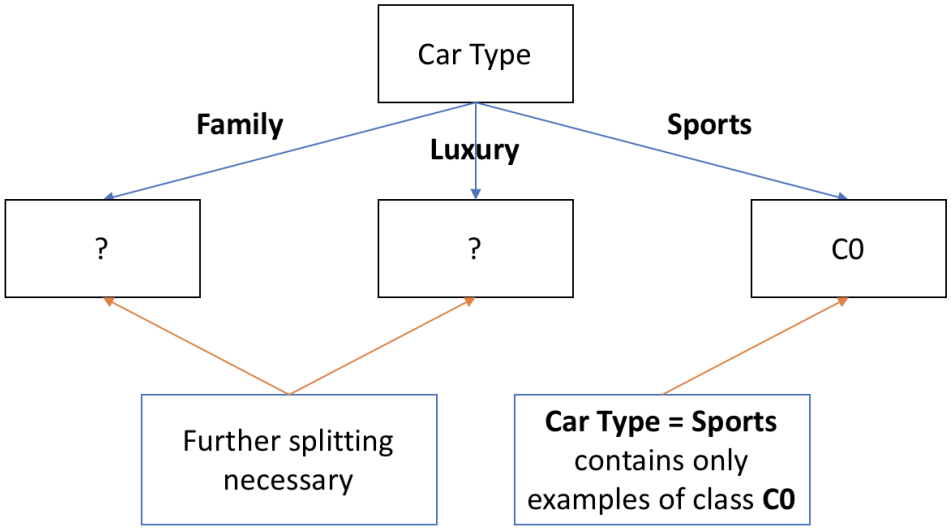

Figure 1: Decision tree root node.

Now, to select the node connected to Car Type = Family we compute new GAIN's.

For Car Type = Family we have

$$\text{GINI}(\text{Car Type} = \text{Family}) = 1 - (1/4)^2 - (3/4)^2 = 0.375.$$

Gender:

$$\begin{aligned}\text{GAIN} &= 0.375 - \frac{4}{4}(1 - (1/4)^2 - (3/4)^2) \\ &= 0.375 - 0.375 \\ &= 0\end{aligned}$$

Shirt Size:

$$\begin{aligned}\text{GAIN} = 0.375 &- \frac{1}{4}(1 - (0/1)^2 - (1/1)^2) - \frac{1}{4}(1 - (0/1)^2 - (1/1)^2) \\ &- \frac{1}{4}(1 - (0/1)^2 - (1/1)^2) - \frac{1}{4}(1 - (1/1)^2 - (0/1)^2) \\ = 0.375\end{aligned}$$

The biggest GAIN is obtained with Shirt Size, so this is selected node. See Figure 2.

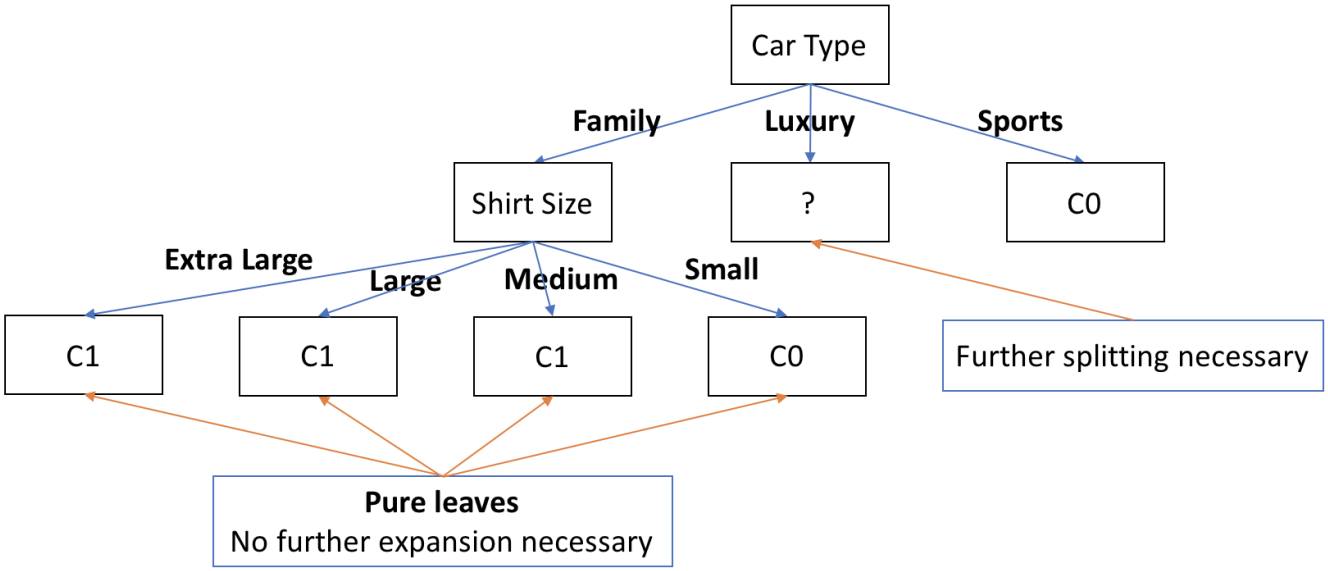

Figure 2: Decision tree updated with Shirt Size selected.

Splitting the Car Type = Luxury node and using the **Hint** in the problem statement we have the final decision tree presented in Figure 3.

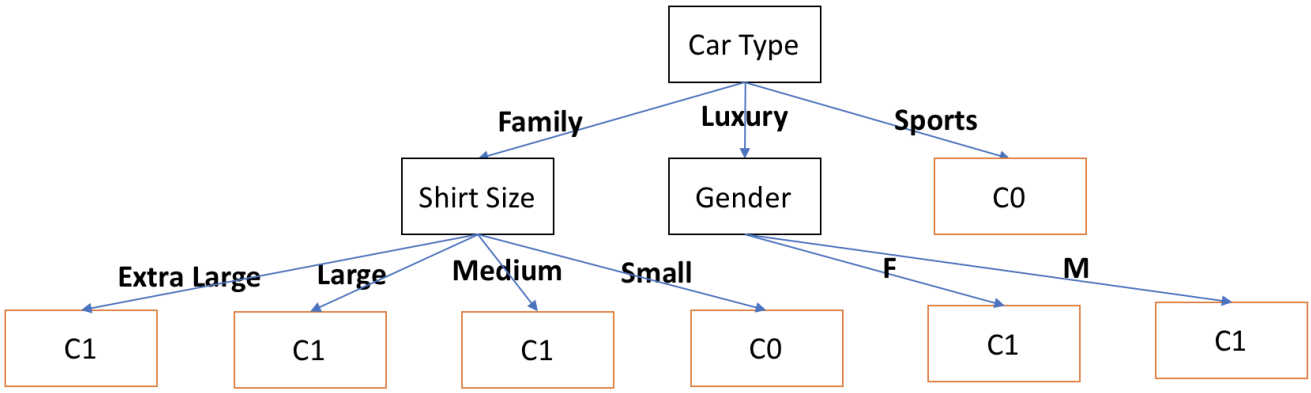

Figure 3: Final decision tree.

For the Gender M in Car Type = Luxury we have only one sample, of class C1. For the Gender F we have 8 samples, only one is of class C0. So we made use of the statement **Hint** and considered the majority class of all samples, therefore the result is C1.

□

Question 2

Table 2 consists of training data from an employee database:

Table 2: Data set of an employee database.

Department	Status	Age	Salary	Count
Sales	Senior	31 . . . 35	46K-50K	30
Sales	Junior	26 . . . 30	26K-30K	40
Sales	Junior	31 . . . 35	31K-35K	40
Systems	Junior	21 . . . 25	46K-50K	20
Systems	Senior	31 . . . 35	66K-70K	5
Systems	Junior	26 . . . 30	46K-50K	3
Systems	Senior	41 . . . 45	66K-70K	3
Marketing	Senior	36 . . . 40	46K-50K	10
Marketing	Junior	31 . . . 35	41K-45K	4
Secretary	Senior	46 . . . 50	36K-40K	4
Secretary	Junior	26 . . . 30	26K-30K	6

The data have been generalized. For a given row entry, *count* represents the number of data examples having the values for *departments*, *status*, *age*, and *salary* given in that row. Let the *status* be the class label attribute.

(1)

How to modify C4.5 algorithm to take into consideration the *count* of each generalized data tuple (i.e. of each row entry)?

Solution:

Integrating the count of each tuple into the calculation of the attribute selection measure (such as GainRATIO). Taking the count into consideration to determine the most common class among the tuples.

□

(2)

Construct a decision tree from the given data by using the modified C4.5 algorithm (Hint: *Age* and *Salary* have been discretized into intervals. You can consider them like ordinal attributes. When trying multi-splitting, you can merge values by their closeness. For example, if you have a three-way split of age, you can have [26-30] at one branch, [31 35] at one, and [36 40] [41 45] [46 50] at one. It is ok as long as you try a number of reasonable splits.)

Solution:

To choose the root node we have to run all the possible nodes and compute the GainRATIO

$$\text{GainRATIO} = \frac{\text{GAIN}}{\text{SplitINFO}}, \quad \text{with} \quad \text{SplitINFO} = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}.$$

n_i is the number of records in partition i.

For Department

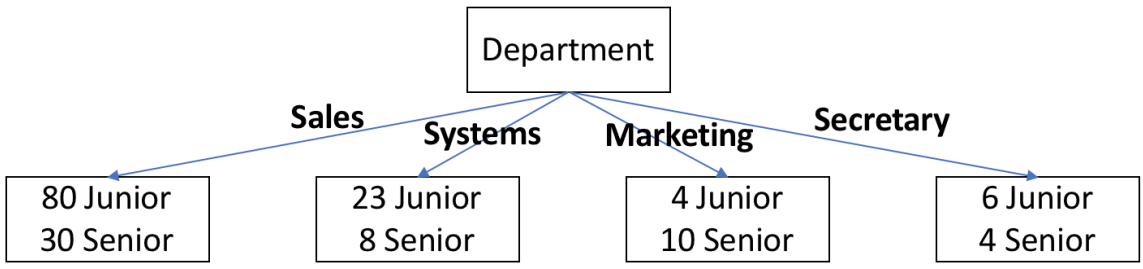

Figure 4: Simplified decision tree for Department.

$$\begin{aligned}
\text{GainRATIO} &= \frac{\text{GAIN}}{\text{SplitINFO}} \\
\text{GAIN} &= \Big(1 - \Big(\frac{113}{165}\Big)^2 - \Big(\frac{52}{165}\Big)^2\Big) \\
&\quad - \frac{110}{165}\Big(1 - \Big(\frac{80}{110}\Big)^2 - \Big(\frac{30}{110}\Big)^2\Big) - \frac{31}{165}\Big(1 - \Big(\frac{23}{31}\Big)^2 - \Big(\frac{8}{31}\Big)^2\Big) \\
&\quad - \frac{14}{165}\Big(1 - \Big(\frac{4}{14}\Big)^2 - \Big(\frac{10}{14}\Big)^2\Big) - \frac{10}{165}\Big(1 - \Big(\frac{6}{10}\Big)^2 - \Big(\frac{4}{10}\Big)^2\Big) = 0.4316 - 0.4001 \\
&= 0.0315 \\
\text{SplitINFO} &= -\Big(\frac{110}{165}\log\frac{110}{165} + \frac{31}{165}\log\frac{31}{165} + \frac{14}{165}\log\frac{14}{165} + \frac{10}{165}\log\frac{10}{165}\Big) = 0.9636 \\
\text{GainRATIO} &= \frac{0.0315}{0.9636} = 0.0327.
\end{aligned}$$

For Age

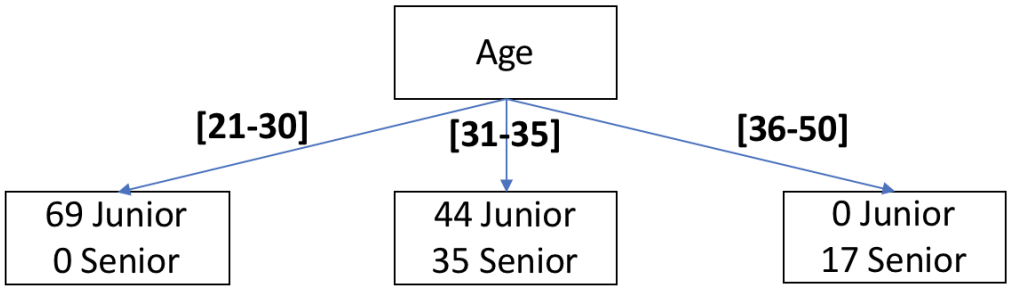

Figure 5: Simplified decision tree for Age.

$$\begin{aligned}
\text{GainRATIO} &= \frac{\text{GAIN}}{\text{SplitINFO}} \\
\text{GAIN} &= \Big(1 - \Big(\frac{113}{165}\Big)^2 - \Big(\frac{52}{165}\Big)^2\Big) \\
&\quad - \frac{69}{165}\Big(1 - \Big(\frac{69}{69}\Big)^2 - \Big(\frac{0}{69}\Big)^2\Big) - \frac{79}{165}\Big(1 - \Big(\frac{44}{79}\Big)^2 - \Big(\frac{35}{79}\Big)^2\Big) \\
&\quad - \frac{17}{165}\Big(1 - \Big(\frac{0}{17}\Big)^2 - \Big(\frac{17}{17}\Big)^2\Big) = 0.4316 - 0.2363 \\
&= 0.1953 \\
\text{SplitINFO} &= -\Big(\frac{69}{165}\log\frac{69}{165} + \frac{79}{165}\log\frac{79}{165} + \frac{17}{165}\log\frac{17}{165}\Big) = 0.9513 \\
\text{GainRATIO} &= \frac{0.1953}{0.9513} = 0.2053.
\end{aligned}$$

For Salary (merging values by their closeness, as mencioned in the **Hint**, and having so three branchs)

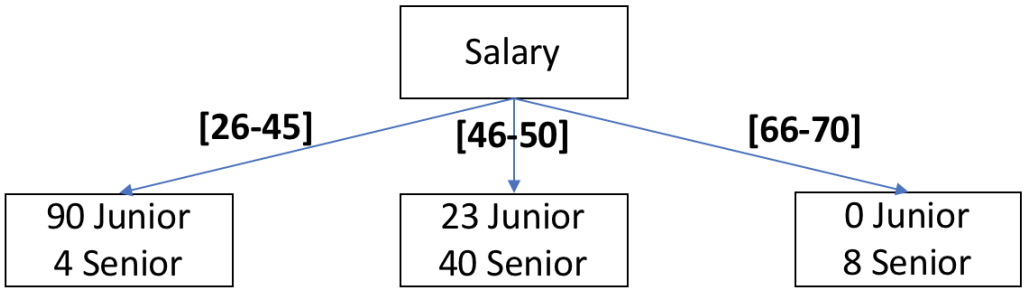

Figure 6: Simplified decision tree for Salary.

$$\begin{aligned}
\text{GAIN} &= \left(1-\left(\frac{113}{165}\right)^2-\left(\frac{52}{165}\right)^2\right) \\
&\quad -\frac{94}{165}\left(1-\left(\frac{90}{94}\right)^2-\left(\frac{4}{94}\right)^2\right)-\frac{63}{165}\left(1-\left(\frac{23}{63}\right)^2-\left(\frac{40}{63}\right)^2\right) \\
&\quad -\frac{8}{165}\left(1-\left(\frac{0}{8}\right)^2-\left(\frac{8}{8}\right)^2\right) = 0.4316-0.2234 \\
&= 0.2082 \\
\text{SplitINFO} &= -\left(\frac{94}{165}\log\frac{94}{165}+\frac{63}{165}\log\frac{63}{165}+\frac{8}{165}\log\frac{8}{165}\right) = 0.8349 \\
\text{GainRATIO} &= \frac{0.2082}{0.8349} = 0.2494.
\end{aligned}$$

The biggest GainRATIO is obtained with Salary, therefore Salary is the root node.

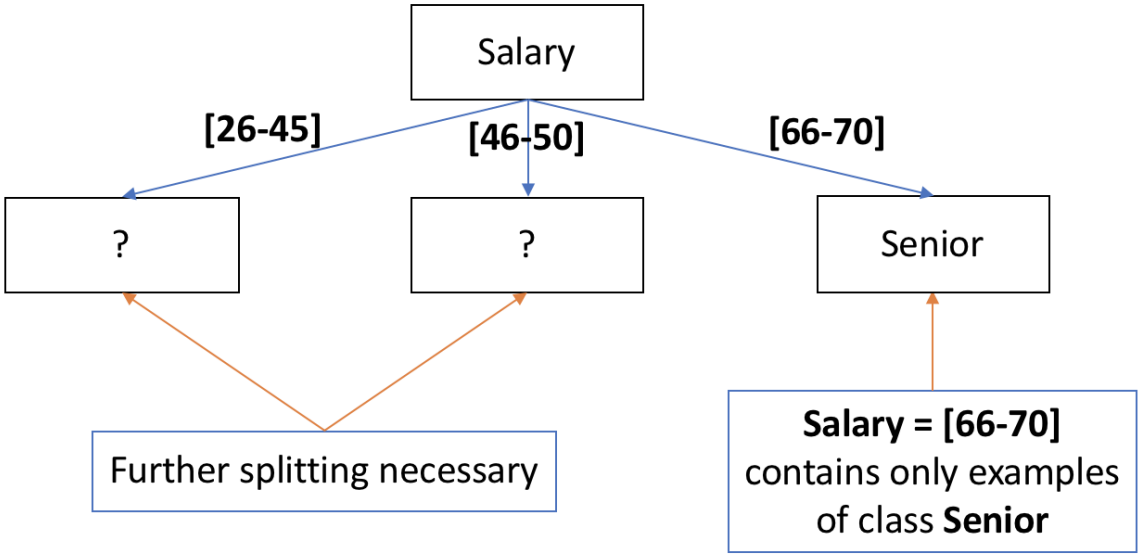

Figure 7: Decision tree root node.

Now, to select the node connected to Salary = [26-45] we compute new GainRATIO's.

For Department

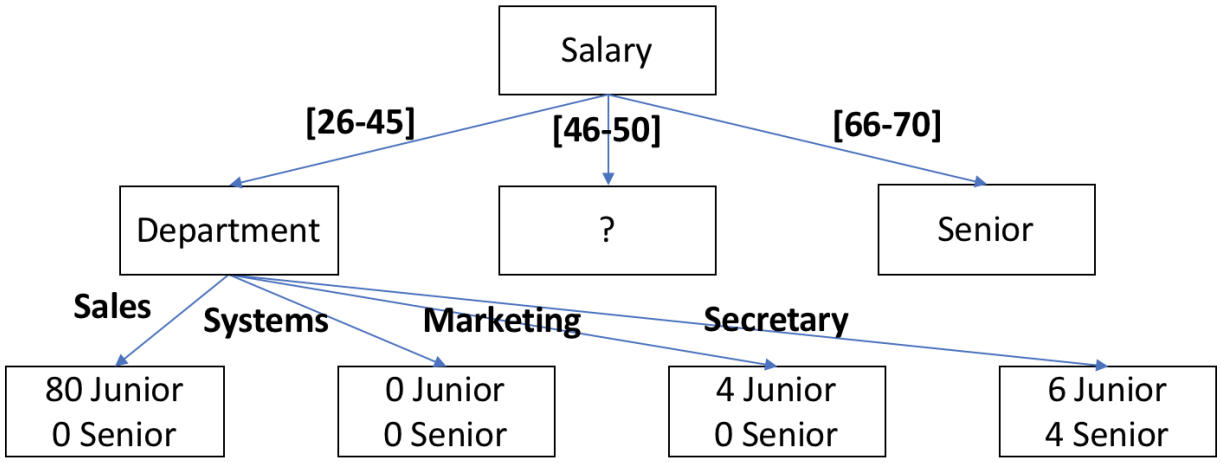

Figure 8: Simplified updated decision tree for Department.

$$
\begin{aligned}
\text{GAIN} &= \left(1-\left(\frac{90}{94}\right)^2-\left(\frac{4}{94}\right)^2\right) \\
&\quad -\frac{80}{94}\left(1-\left(\frac{80}{80}\right)^2-\left(\frac{0}{80}\right)^2\right)-\frac{4}{94}\left(1-\left(\frac{4}{4}\right)^2-\left(\frac{0}{4}\right)^2\right) \\
&\quad -\frac{10}{94}\left(1-\left(\frac{6}{10}\right)^2-\left(\frac{4}{10}\right)^2\right) = 0.0815-0.051 \\
&= 0.0305 \\
\text{SplitINFO} &= -\left(\frac{80}{94}\log\frac{80}{94}+\frac{4}{94}\log\frac{4}{94}+\frac{10}{94}\log\frac{10}{94}\right) = 0.5099 \\
\text{GainRATIO} &= \frac{0.0305}{0.5099} = 0.0598.
\end{aligned}
$$

For Age

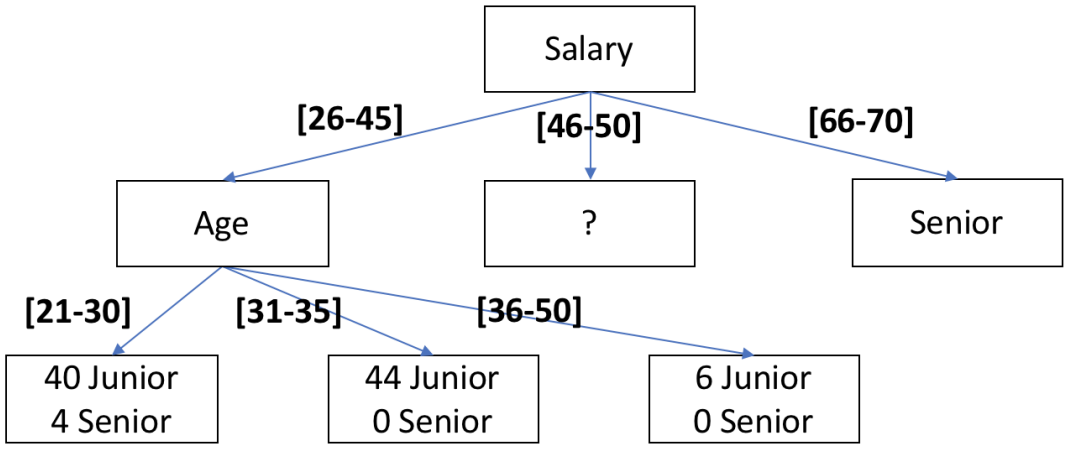

Figure 9: Simplified updated decision tree for Age.

$$
\begin{aligned}
\text{GAIN} &= \left(1 - \left(\frac{90}{94}\right)^2 - \left(\frac{4}{94}\right)^2\right) \\
&\quad - \frac{44}{94}\left(1 - \left(\frac{40}{44}\right)^2 - \left(\frac{4}{44}\right)^2\right) - \frac{44}{94}\left(1 - \left(\frac{44}{44}\right)^2 - \left(\frac{0}{44}\right)^2\right) \\
&\quad - \frac{6}{94}\left(1 - \left(\frac{6}{6}\right)^2 - \left(\frac{0}{6}\right)^2\right) = 0.0815 - 0.0773 \\
&= 0.0042 \\
\text{SplitINFO} &= -\left(\frac{44}{94}\log\frac{44}{94} + \frac{44}{94}\log\frac{44}{94} + \frac{6}{94}\log\frac{6}{94}\right) = 0.8863 \\
\text{GainRATIO} &= \frac{0.0042}{0.8863} = 0.0047.
\end{aligned}
$$

The biggest GainRATIO is obtained with Department, therefore Department is the connection with Salary = [26-45]. **Notes**: For this salary level we don't have samples for Department = Systems. For Department = Secretary the connections to two levels of Age are automatically, as show the Figure 10.

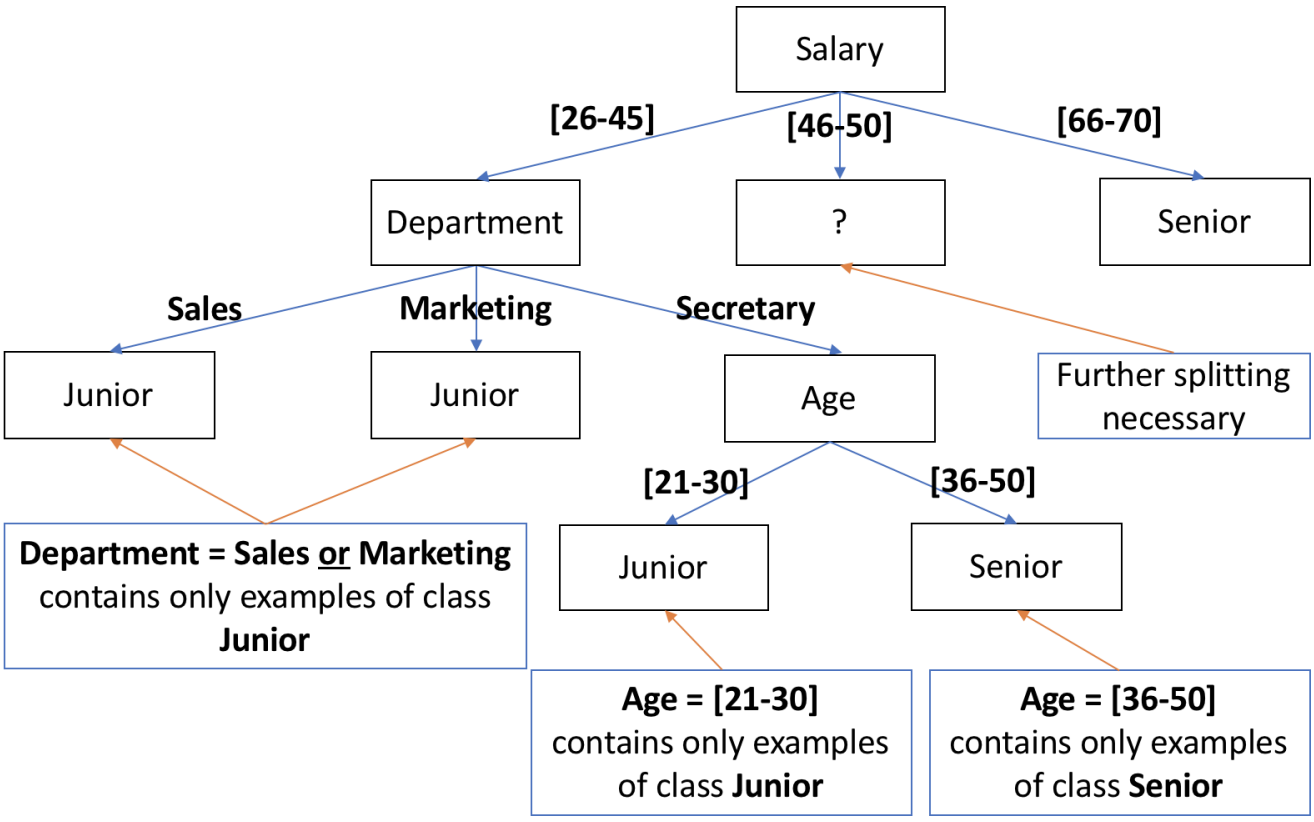

Figure 10: Updated decision tree.

When we connect the node Age in Salary = [46-50] we already arrive in leaf nodes for all the Age classes. Theferore, this is our final tree (Figure 11).

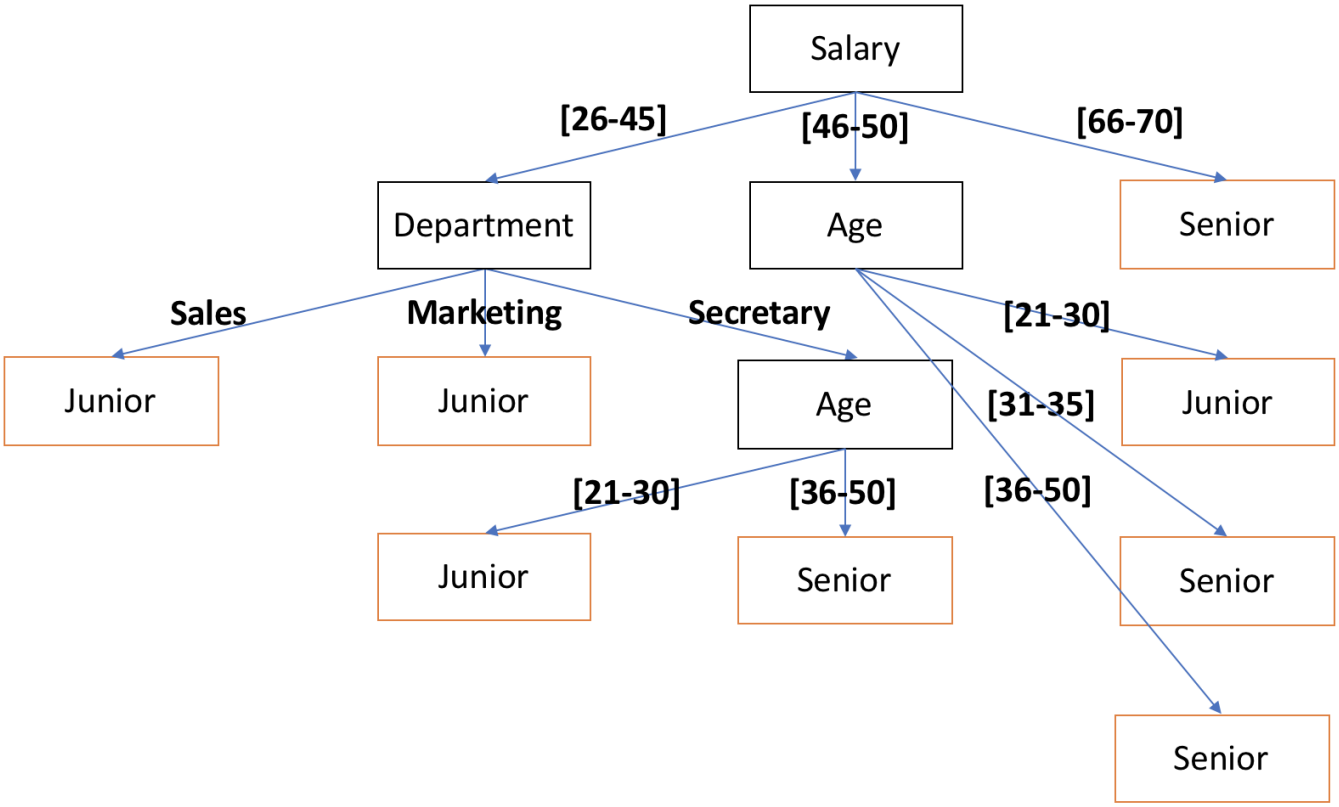

Figure 11: Final decision tree.

□

(3)

Use the tree you learned to classify a given example with the values "system", "26 . . . 30" and "46-50K" for the attributes *departments*, *age*, and *salary*. The *status* of this employee is?

Solution:

$$\text{Salary} \xrightarrow{[46-50]} \text{Age} \xrightarrow{[21-30]} \Rightarrow \text{status : Junior.}$$

Following the tree in Figure 11 we don't need the value "system" of the attribute *departments* to classify an employee with the given characteristics (*salary* and *age*).

□

(4)

Use the training data in Table 2 to learn a Naive Bayes classifier, and classify the same given example with the values "system", "26 . . . 30" and "46-50K" for the attributes *departments*, *age*, and *salary*. The *status* of this employee is?

Solution:

$$X = (\text{Department} = \text{Systems}, \text{Age} = [26-30], \text{Salary} = [46-50])$$

(Using Laplace probability estimation to avoid the 0-probability problem)

$$\begin{aligned}
\mathbb{P}(X \mid \text{Status} = \text{Junior}) &= \mathbb{P}(\text{Department} = \text{Systems} \mid \text{Status} = \text{Junior}) \times \\
&\quad \mathbb{P}(\text{Age} = [26-30] \mid \text{Status} = \text{Junior}) \times \\
&\quad \mathbb{P}(\text{Salary} = [46-50] \mid \text{Status} = \text{Junior}) \\
&= \frac{23+1}{113+4} \times \frac{49+1}{113+6} \times \frac{23+1}{113+6} \\
&= 0.0174 \\
\mathbb{P}(X \mid \text{Status} = \text{Senior}) &= \mathbb{P}(\text{Department} = \text{Systems} \mid \text{Status} = \text{Senior}) \times \\
&\quad \mathbb{P}(\text{Age} = [26-30] \mid \text{Status} = \text{Senior}) \times \\
&\quad \mathbb{P}(\text{Salary} = [46-50] \mid \text{Status} = \text{Senior}) \\
&= \frac{8+1}{52+4} \times \frac{0+1}{52+6} \times \frac{40+1}{52+6} \\
&= 0.0019
\end{aligned}$$

$$\mathbb{P}(X \mid \text{Status} = \text{Junior}) \times \mathbb{P}(\text{Status} = \text{Junior}) > \mathbb{P}(X \mid \text{Status} = \text{Senior}) \times \mathbb{P}(\text{Status} = \text{Senior})$$

$$0.0174 \times \frac{113}{165} > 0.0019 \times \frac{52}{165}$$

$$0.0119 > 0.0006$$

Therefore,

$$\boxed{\mathbb{P}(\text{Status} = \text{Junior} \mid X) > \mathbb{P}(\text{Status} = \text{Senior} \mid X) \quad \Rightarrow \quad \text{Employee } \textit{status} = \text{Junior}.}$$

□

Question 3

Why is *tree pruning* useful in decision tree induction? What are the pros and cons of using a separate set of samples to evaluate pruning?

Solution:

A *tree pruning* is useful in decision tree induction because induced trees may overfit the training data. With too many branchs, e.g., some may reflect anomalies due to noise or outliers, or, also e.g., poor accucary for unseen samples may happen.

Pros of using a separate set of samples to evaluate pruning
: Reduces overfit and error pruning for using different samples that may have different characteristics and patterns.

Cons of using a separate set of samples to evaluate pruning
: May overprune the decision tree, deleting relevant parts from it. Less data is available for training.

■
