Course: Data Mining

Homework 4

Student ID: 20C14001

Student Name: Le Duong Tuan Anh

Question 1

Training set

Shape	Country	Status	Group
Small	Germany	Single	A
Big	French	Single	A
Big	Germany	Single	A
Small	Italy	Single	B
Big	Germany	Married	B
Big	Italy	Single	B
Big	Italy	Married	B
Small	Germany	Married	B

a. Decision Tree by Information Gain measure

Class P: Group = A

Class N: Group = B

$$\text{Info(Group)} = \text{I(3,5)} = -\frac{3}{8}log_2\left(\frac{3}{8}\right) - \frac{5}{8}log_2\left(\frac{5}{8}\right) \approx 0.954$$

Shape:

Shape	Country	Status	Group
Small	Germany	Single	A
Small	Italy	Single	B
Small	Germany	Married	B

Shape	Country	Status	Group
Big	French	Single	A
Big	Germany	Single	A
Big	Germany	Married	B
Big	Italy	Single	B
Big	Italy	Married	B

$$\text{Info}_{Shape}(\text{Group}) = \frac{3}{8}I(1,2) + \frac{5}{8}I(2,3) \approx 0.951$$

$$Gain(shape) = Info(Group) - Info_{Shape}(Group) = 0.003$$

Country:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	Germany	Single	A
Big	Germany	Married	B
Small	Germany	Married	B

Shape	**Country**	**Status**	**Group**
Big	French	Single	A

Shape	**Country**	**Status**	**Group**
Small	Italy	Single	B
Big	Italy	Single	B
Big	Italy	Married	B

$$\text{Info}_{Country}(\text{Group}) = \frac{4}{8}I(2,2) + \frac{1}{8}I(1,0) + \frac{3}{8}I(0,3) = 0.5$$

$$Gain(Country) = Info(Group) - Info_{Country}(Group) = 0.454$$

Status:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	French	Single	A
Big	Germany	Single	A
Small	Italy	Single	B
Big	Italy	Single	B

Shape	**Country**	**Status**	**Group**
Big	Germany	Married	B
Big	Italy	Married	B
Small	Germany	Married	B

$$\text{Info}_{Status}(\text{Group}) = \frac{5}{8}I(3,2) + \frac{3}{8}I(0,3) \approx 0.607$$

$$Gain(Status) = Info(Group) - Info_{Status}(Group) = 0.347$$

$$Gain(shape) = 0.003$$
$$Gain(country) = 0.454$$
$$Gain(status) = 0.347$$

So, choose attribute **Country = Germany *(as** Country = French has only Group = A and Italy only has Group = B)*

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	Germany	Single	A
Big	Germany	Married	B
Small	Germany	Married	B

$\text{Info}(\text{Group}) = \text{I}(2,2) = 1$

Shape:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Small	Germany	Married	B

Shape	**Country**	**Status**	**Group**
Big	Germany	Single	A
Big	Germany	Married	B

$$\text{Info}_{shape}(\text{Group}) = \frac{2}{4}I(1,1) + \frac{2}{4}I(1,1) = 1$$
$$Gain(Status) = Info(Group) - Info_{Status}(Group) = 0$$

Status

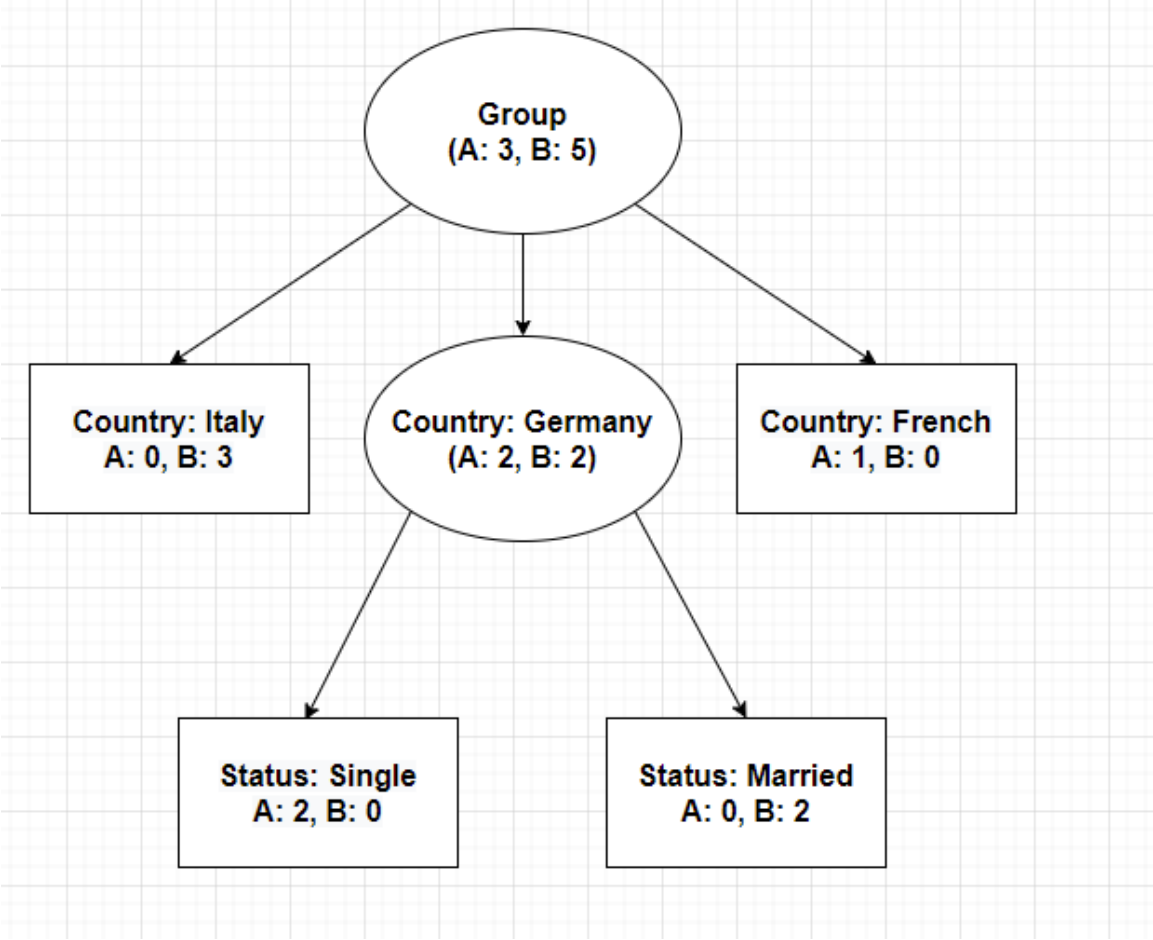
Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	Germany	Single	A

Shape	**Country**	**Status**	**Group**
Big	Germany	Married	B
Small	Germany	Married	B

$$\text{Info}_{shape}(\text{status}) = \frac{2}{4}I(2,0) + \frac{2}{4}I(0,2) = 0$$
$$Gain(Status) = Info(Group) - Info_{Status}(Group) = 1$$

We choose **Status** is the next node. Now, we stop as there is only one result per split.

Decision Tree



b. Classification rules from (a):

If (country == Italy) then **Group = B**

If (country == French) then **Group = A**

If (country == Germany) then:

If (Status == Single) then **Group = A**

If (Status == Married) then **Group = B.**

c. Decision Tree by Gain Ratio measure:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	French	Single	A
Big	Germany	Single	A
Small	Italy	Single	B
Big	Germany	Married	B
Big	Italy	Single	B
Big	Italy	Married	B
Small	Germany	Married	B

Class P: Group = A

Class N: Group = B

$$\text{Info(Group)} = \text{I(3,5)} = -\frac{3}{8}log_2\left(\frac{3}{8}\right) - \frac{5}{8}log_2\left(\frac{5}{8}\right) \approx 0.954$$

Shape:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Small	Italy	Single	B
Small	Germany	Married	B

Shape	**Country**	**Status**	**Group**
Big	French	Single	A
Big	Germany	Single	A
Big	Germany	Married	B
Big	Italy	Single	B
Big	Italy	Married	B

$$\text{Info}_{Shape}\text{(Group)} = \frac{3}{8}I(1,2) + \frac{5}{8}I(2,3) \approx 0.951$$

$$Gain(shape) = Info(Group) - Info_{Shape}(Group) = 0.003$$

$$\text{SplitInfo(Shape)} = -\frac{3}{8}log_2\left(\frac{3}{8}\right) - \frac{5}{8}log_2\left(\frac{5}{8}\right) \approx 0.954$$

$$\text{GainRatio(Shape)} = \frac{0.003}{0.954} \approx 0.003$$

Country:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	Germany	Single	A
Big	Germany	Married	B
Small	Germany	Married	B

Shape	**Country**	**Status**	**Group**
Big	French	Single	A

Shape	**Country**	**Status**	**Group**
Small	Italy	Single	B
Big	Italy	Single	B
Big	Italy	Married	B

$$\text{Info}_{Country}(\text{Group}) = \frac{4}{8}I(2,2) + \frac{1}{8}I(1,0) + \frac{3}{8}I(0,3) = 0.5$$

$$Gain(Country) = Info(Group) - Info_{Country}(Group) = 0.454$$

$$\text{SplitInfo(Country)} = -\frac{4}{8}log_2\left(\frac{4}{8}\right) - \frac{1}{8}log_2\left(\frac{1}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) \approx 1.406$$

$$\text{GainRatio(Country)} = \frac{0.454}{1.406} \approx 0.323$$

Status:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	French	Single	A
Big	Germany	Single	A
Small	Italy	Single	B
Big	Italy	Single	B

Shape	**Country**	**Status**	**Group**
Big	Germany	Married	B
Big	Italy	Married	B
Small	Germany	Married	B

$$\text{Info}_{Status}(\text{Group}) = \frac{5}{8}I(3,2) + \frac{3}{8}I(0,3) \approx 0.607$$

$$Gain(Status) = Info(Group) - Info_{Status}(Group) = 0.347$$

$$\text{SplitInfo(Status)} = -\frac{5}{8}log_2\left(\frac{5}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) \approx 0.954$$

$$\text{GainRatio(Status)} = \frac{0.347}{0.954} \approx 0.364$$

We have:

$$GainRatio(shape) = 0.003$$
$$GainRatio(country) = 0.323$$
$$Gain(status) = 0.364$$

Splitting Attribute: Status, *Status = Single (as Status = Married only has result Group = B)*

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	French	Single	A
Big	Germany	Single	A
Small	Italy	Single	B
Big	Italy	Single	B

$$\text{Info(Group)} = \text{I(3,2)} = -\frac{3}{5}log_2\left(\frac{3}{5}\right) - \frac{2}{5}log_2\left(\frac{2}{5}\right) \approx 0.971$$

Shape:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Small	Italy	Single	B

Shape	**Country**	**Status**	**Group**
Big	French	Single	A
Big	Germany	Single	A
Big	Italy	Single	B

$$\text{Info}_{Shape}(\text{Group}) = \frac{2}{5} I(1,1) + \frac{3}{5} I(2,1) \approx 0.117$$

$$Gain(shape) = Info(Group) - Info_{Shape}(Group) = 0.854$$

$$\text{SplitInfo}(\text{Shape}) = -\frac{2}{5} log_2\left(\frac{2}{5}\right) - \frac{3}{5} log_2\left(\frac{3}{5}\right) \approx 0.971$$

$$\text{GainRatio}(\text{Shape}) = \frac{0.883}{0.971} \approx 0.88$$

Country:

Shape	**Country**	**Status**	**Group**
Small	Germany	Single	A
Big	Germany	Single	A

Shape	**Country**	**Status**	**Group**
Big	French	Single	A

Shape	**Country**	**Status**	**Group**
Small	Italy	Single	B
Big	Italy	Single	B

$$\text{Info}_{Shape}(\text{Country}) = \frac{2}{5} I(2,0) + \frac{1}{5} I(1,0) + \frac{2}{5} I(0,2) = 0$$

$$Gain(\text{Country}) = Info(Group) - Info_{Shape}(Group) = 0.971$$

$$\text{SplitInfo}(\text{Country}) = -\frac{2}{5} log_2\left(\frac{2}{5}\right) - \frac{1}{5} log_2\left(\frac{1}{5}\right) - \frac{2}{5} log_2\left(\frac{2}{5}\right) \approx 1.522$$

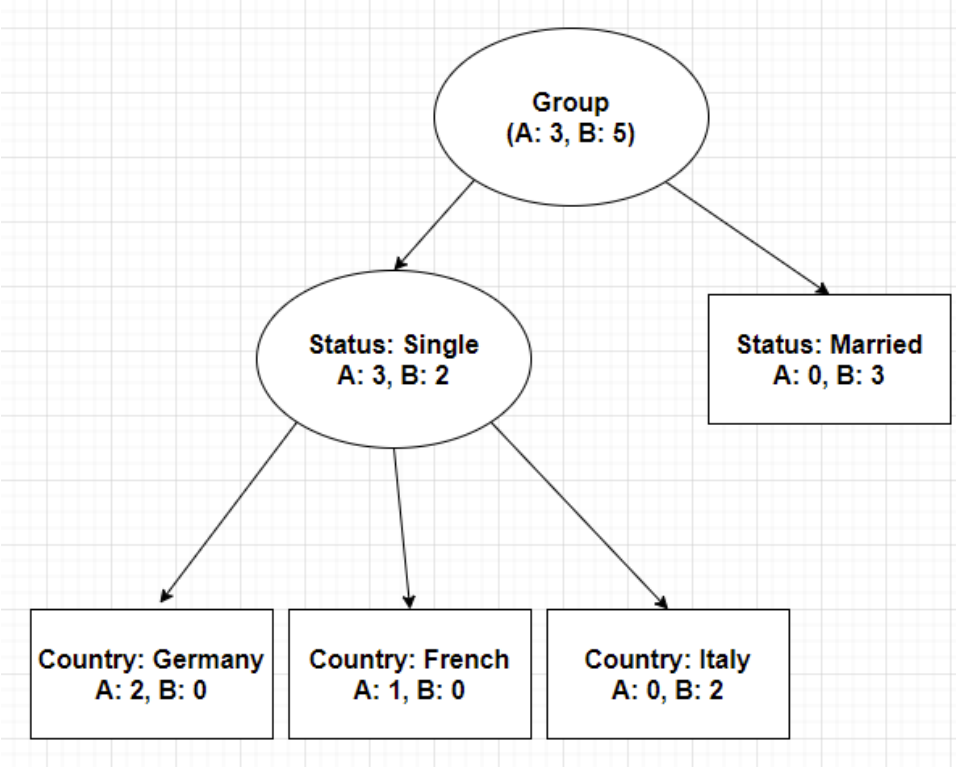
$$\text{GainRatio}(\text{Country}) = \frac{0.971}{1.522} \approx 0.909$$

We have:

$$GainRatio(shape) = 0.88$$
$$GainRatio(country) = 0.909$$

Splitting Attribute: Country. Stopped, as for every split we have similar Group value.

Decision Tree:



Classification Rule:
If (status == Married) then **Group = B**

If (status == Single) then:

If (Country == Germany) or (Country == French) then **Group = A**

If (Country == Italy) then **Group = B.**

d. Evaluate the tree of (a) and (c)

Shape	**Country**	**Status**	**(a) result**	**(c) result**	**Expected Result**
Small	Germany	Single	A	A	A
Big	French	Single	A	A	A
Big	Germany	Single	A	A	A
Small	Italy	Single	B	B	B
Big	Germany	Married	B	B	B
Big	Italy	Single	B	B	B
Big	Italy	Married	B	B	B
Small	Germany	Married	B	B	B

Confusion Matrix for (a)

Actual\Predicted	**A**	**B**	
A	**3**	**0**	
B	**0**	**5**	
			Accuracy: 100%

Confusion Matrix for (b)

Actual\Predicted	**A**	**B**	
A	**3**	**0**	
B	**0**	**5**	
			Accuracy: 100%

Question 2

Customer	Article1	Article2	Article3	Article4	Article5	Artical6
1	0	1	0	1	1	0
2	0	1	1	1	0	1
3	1	0	1	0	1	0
4	1	0	0	1	0	1
5	1	0	0	0	1	0

a. K-means with Cosine distance

K=2

Initial centroids:

Center 1: (0,0,1,1,1)

Center 2: (0,1,1,0,0)

Iteration 1:

Center 1: (0,0,1,1,1)

Center 2: (0,1,1,0,0)

Article	Cosine Distance to Center 1	Cosine distance to Center 2
Article 1 (0,0,1,1,1)	**0.0**	0.5917517095361369
Article 2 (1,1,0,0,0)	1.0	**0.5**
Article 3 (0,1,1,0,0)	0.5917517095361369	**0.0**
Article 4 (1,1,0,1,0)	0.6666666666666667	**0.5917517095361369**
Article 5 (1,0,1,0,1)	**0.3333333333333333**	0.5917517095361369
Article 6 (0,1,0,1,0)	0.5917517095361369	**0.5**

Update centroids:

Center 1: (0,0,1,1,1) ; (1,0,1,0,1) ➔ **(0.5, 0, 1, 0.5, 1)**

Center 2: (1,1,0,0,0) ; (0,1,1,0,0) ; (1,1,0,1,0) ; (0,1,0,1,0) ➔ **(0.5, 1, 0.25, 0.5, 0)**

Iteration 2:

Center 1: **(0.5, 0, 1, 0.5, 1)**

Center 2: **(0.5, 1, 0.25, 0.5, 0)**

Article	Cosine distance to Center 1	Cosine distance to Center 2
Article 1 (0,0,1,1,1)	**0.08712907**	0.65358984
Article 2 (1,1,0,0,0)	0.7763932	**0.15147186**
Article 3 (0,1,1,0,0)	0.5527864	**0.29289322**
Article 4 (1,1,0,1,0)	0.63485163	**0.07623957**
Article 5 (1,0,1,0,1)	**0.08712907**	0.65358984
Article 6 (0,1,0,1,0)	0.7763932	**0.15147186**

Update centroids:

There is no chance on cluster, therefore, stop here.

Final centroids:

Center 1: **(0.5, 0, 1, 0.5, 1) ; Article (1,5)**

Center 2: **(0.5, 1, 0.25, 0.5, 0) ; Article (2,3,4,6)**

b. BetaCV measure

$$Cosine(i,j) = 1 - \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} * \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

$W(C_1, C_1) = Cosine(1,5) + Cosine(5,1) = 0.6666666666666667$

$W(C_2, C_2) = Cosine(2,3) + Cosine(2,4) + Cosine(2,6) + Cosine(3,2) + Cosine(3,4)$
$+ Cosine(3,6) + Cosine(4,2) + Cosine(4,3) + Cosine(4,6)$
$+ Cosine(6,2) + Cosine(6,3) + Cosine(6,4) = 4.917517095361369$

$$\boldsymbol{W_{in} = \frac{1}{2}\sum_{i=1}^{k=2} W(C_i, C_i) = 2.792091881014018}$$

$$\boldsymbol{W_{out} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, \overline{C_i})}$$

$$\boldsymbol{= \sum W(C_1, C_2) = Cosine(1,2) + Cosine(1,3) + Cosine(1,4) + Cosine(1,6)}$$
$$\boldsymbol{+ Cosine(5,2) + Cosine(5,3) + Cosine(5,4) + Cosine(5,6)}$$
$$\boldsymbol{= 5.700340171477881}$$

$$\boldsymbol{N_{in} = \sum_{i=1}^{k=2} \binom{n_i}{2} = \binom{2}{2} + \binom{4}{2} = 7}$$

$$\boldsymbol{N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i n_j = 2 * 4 = 8}$$

$$\boldsymbol{BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{2.792091881014018\ /\ 7}{5.700340171477881\ /\ 8} = 0.5597845135 \approx 0.56}$$