**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**

4

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 |

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 |

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 |

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 |

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 |

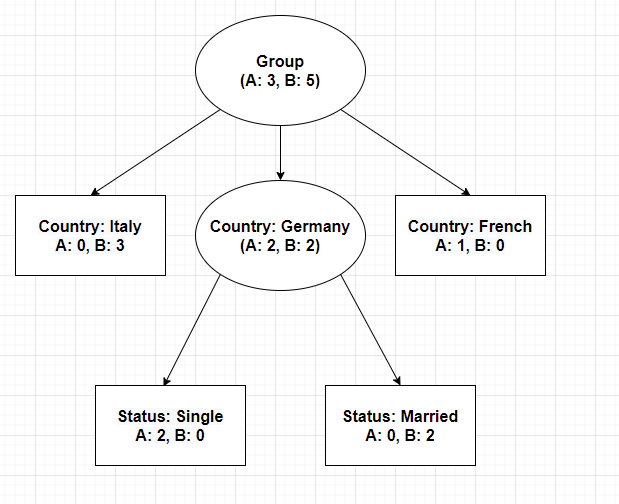
Status

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

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

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**

4

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 |

4

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 |

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 |

**We have:**

**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 |

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 |

Country:

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

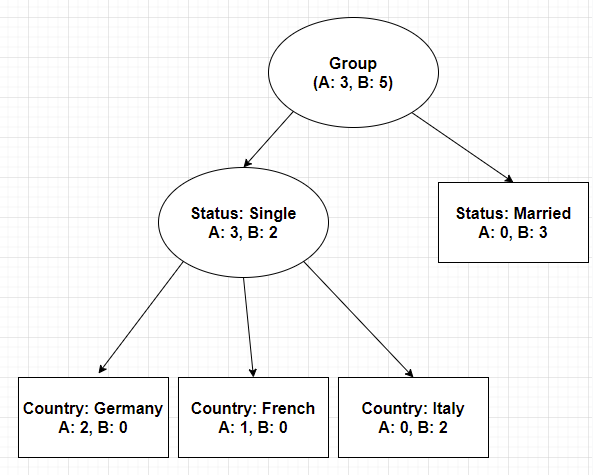
|  |  |  |  |
| --- | --- | --- | --- |
| **Shape** | **Country** | **Status** | **Group** |
| Big | French | Single | A |

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

**We have:**

**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*