# Assignment #3

By: Daria Yip

# Q1: Association Rules

**Candidate Set (C1)**

|  |  |
| --- | --- |
| **Item** | **Support Count** |
| Meat | 5/10 |
| Potato | 5/10 |
| Onion | 4/10 |
| Noodle | 4/10 |
| Spinach | 3/10 |
| Eggs | 2/10 |
| Salt | 2/10 |

**Minimum Support Value = 0.3 \* 10 transactions = 3**

Therefore, Eggs and Salt do not meet the minimum support value of 3, or 3/10 Support level, so they are not included in the Item Set.

**Item Set (I1)**

|  |  |
| --- | --- |
| **Item** | **Frequency** |
| Meat, Potato | 4/10 |
| Meat, Onion | 3/10 |
| Meat, Noodle | 1/10 |
| Meat, Spinach | 0/10 |
| Potato, Onion | 4/10 |
| Potato, Noodle | 1/10 |
| Potato, Spinach | 0/10 |
| Onion, Noodle | 1/10 |
| Onion, Spinach | 0/10 |
| Noodle, Spinach | 1/10 |

**Item Set (I2)**

|  |  |
| --- | --- |
| **Item** | **Frequency** |
| Meat, Potato, Onion | 3/10 |

Therefore, the most frequent item-set determining the sub-item set is [meat, potato, onion].

# 1.2

Using the following Confidence formula…

**Confidence = Support (X, Y, Z)/ Support (X, Y)**

**1:** {Meat, Potato} => {Onion}

Confidence = Support(Meat, Potato, Onion)/Support(Meat, Potato) = ¾ = 0.75, 0.75 < 0.8

**NOT ACCEPTED**

**2:** {Meat, Onion} => {Potato}

Confidence = {Meat, Potato, Onion}/{Meat, Onion} = 3/3 = 1 > 0.8

**ACCEPTED**

**3:**{Onion, Potato} => {Meat}

Confidence = {Meat, Potato, Onion}/{Onion, Potato} = ¾ = 0.75 < 0.8

**NOT ACCEPTED**

**4:** {Onion} => {Meat, Potato}

Confidence = 0.3/0.4 = 0.75 < 0.8

**NOT ACCEPTED**

**5.** {Potato} => {Meat, Onion}

Confidence = 0.3/0.5 = 0.6 < 0.8

**NOT ACCEPTED**

**6.** {Meat} => {Potato, Onion}

Confidence = 0.3/0.5 = 0.6 < 0.8

**NOT ACCEPTED**

Therefore, only:

{Meat, Onion} => {Potato}

Confidence = {Meat, Potato, Onion}/{Meat, Onion} = 3/3 = 1 > 0.8

Meets the criteria of Confidence > 0.8.

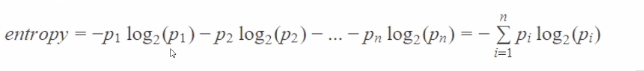
# Q2: Classification

The Entropy Formula for our Class Attribute of Profile is:

= -0.6log2(0.6) – 0.4log2(0.4) = **0.97095059445**

Calculating the Root Node by using Information Gain:

Using the formula from the lecture:



City:

0.97095059445 -

= 0.97095059445 – 0.604184397996641701 – 0.2 – undefined

= **0.166766196453358299**

Gender:

0.97095059445

= 0. 97095059445 - 0.275488750216346 - 0.604184397996641701 = **0.091277446237012299**

Age:

0.97095059445

= 0.97095059445 - 0.4854752972273343194 – undefined – undefined – undefined = **0.4854752972226656806**

Education:

= 0.97095059445 - 0.604184397996641701 = **0.366766196453358299**

Therefore, as the largest information gain, Age will be the root node.

We can see that for the age groups 51-60, 41-50, and 31-40, they all have homogenous results (i.e. all employed, or all unemployed). Therefore, we will only need to split 20-30 age group.

Calculate the Information Gain for the Next Split

Using the same formula, we will calculate the Information Gain for the next split node.

First, we will calculate the formula for the Class Attribute of Profile (new), with the new sample size of 20-30 age group with 5 unemployed and employed members:

= -0.6log2(0.6) – 0.4log2(0.4) = **0.97095059445**

City :

= 0.97095059445 -

= -0.97095059445 - 0.5509775004326937088722433 – 0.4

= **0.019973094017306291127756700**

Gender

= 0.97095059445 -

= **0.019973094**

Education

= 0.97095059445 -

= **0.970951**

Therefore, the next split will be Education, as it has the highest information gain.

Age

20-30 31-40 41-50 51-60

Education

Employed

Unemployed

High School

College

# Q3: Clustering

Using the Euclidean Distance formula, we will calculate the distance from records RIDs 103 and 104.

An example of this is RID 101, where we calculate:

22.36068

We will continue this with the rest of the RIDs.

**Iteration 1:**

|  |  |  |
| --- | --- | --- |
| **RID** | **Distance from 103 (50, 15)** | **Distance from 104 (25, 5)** |
| 101 (30, 5) | Sqrt(500) = 22.36068 | 5 |
| 102 (50, 25) | 10 | Sqrt(1025) = 32.015621 |
| 103 (50, 15) | N/A | Sqrt(725) = 26.925824 |
| 104 (25, 5) | Sqrt(725) = 26.925824 | N/A |
| 105 (30, 10) | Sqrt(425) = 20.615528 | Sqrt(50) = 7.071068 |
| 106 (55, 25) | Sqrt(125) = 11.18034 | Sqrt(1300) = 36.055513 |

The datasets will be clustered based on the yellow highlighter (lowest/minimum distance). Therefore, RID 102, and 106 will be clustered with RID 103 (Cluster 1). RID 101, and RID 105 will be clustered with RID 104 (Cluster 2).

We can update the centroid by averaging all x-axis points and y-axis points in each cluster. That is to say, for RID 103’s cluster, the average of all x-axis points is 55 + 50 + 50 = 155/3 = 51.67, and the average of all y-axis points is 15 + 25 + 25 = 65/3 = 21.67. Therefore, the new centroid is **{51.67, 21.67}.**

For the RID 104’s cluster, the average of all x-axis points is 30 + 30 + 25 = 85/3 = 28.33, and the average of all y-axis points is 5 + 10 + 5 = 20/3 = 6.67. Therefore, the new centroid is **{28.33, 6.67}.**

**Iteration 2:**

|  |  |  |
| --- | --- | --- |
| **RID** | **Distance from 103 (51.67, 21.67)** | **Distance from 104 (28.33, 6.67)** |
| 101 (30, 5) | Sqrt(747.4778) = 27.34004 | Sqrt(5. 5778) = 2.361737 |
| 102 (50, 25) | Sqrt(13.8778) = 3.725292 | Sqrt(805.5778) = 28.382702 |
| 103 (51.67, 21.67) | N/A | Sqrt(769.7556) = 27.74447 |
| 104 (25, 5) | Sqrt(989.1778) = 31.451197 | N/A |
| 105 (30, 10) | Sqrt(605.7778) = 24.612554 | Sqrt(13.8778) = 3.725292 |
| 106 (55, 25) | Sqrt(22.1778) = 4.709331 | Sqrt(1047.2778) = 32.361672 |

Because the clusters did not change their contents (102 and 106 still with 103, and 101 and 105 still with 104), and their new centroid would be minimal change to {52.23, 23.89} and {29.44, 7.23}, we can stop the iteration here.

# 3.2

Classification uses predefined classes based upon the attributes of the object, and Clustering identifies similarities/patterns within objects to group objects. Clustering can also group objects based upon distance. Classification is supervised learning, and clustering is unsupervised learning. Therefore, classification already knows what the possibilities/categories are, and in clustering, we want the algorithm to determine these categories for us.