Paul Ycay

500709618

CIND110- Data Organization for Data Analysts

Dr. Tamer Abdou

Due Date: July 25, 2019

Assignment 3

1. On describing discovered knowledge using association rules

One of the major techniques in data mining involves the discovery of association rules. These rules correlate the presence of a set of items with another range of values for another set of variables. The database in this context is regarded as a collection of transactions, each involving a set of items, as shown below.

Trans ID	Items Purchased
101	milk, bread, eggs
102	milk, juice
103	juice, butter
104	milk, bread, eggs
105	eggs, bread
106	coffee
107	coffee, juice
108	milk, bread, cookies, eggs
109	cookies, butter
110	milk, bread

1.1

Apply the Apriori algorithm on this dataset.
Note that, the set of items is {milk, bread, cookies, eggs, butter, coffee, juice}. You may use 0.2 for the minimum support value.

1.2

Show two rules that have a confidence of 0.7 or greater for an itemset containing three items.

1.1

Item-Set	**Count**
Milk	5
Bread	5
Cookies	2
Eggs	4
Butter	2
Coffee	2
Juice	3

Item-Set	**Count**
Milk, Bread	4
Milk, Eggs	3
Milk, Juice	1
Milk, Cookies	1
Bread, Eggs	3
Bread, Cookies	1
Eggs, Coffee	0
Eggs, Cookies	1
Juice, Butter	1
Juice, Coffee	1
Butter, Cookies	1

List with minimum support value = 0.2; support (frequency of item-set/total transactions)

Since 10 transactions…

Item-set	**Count**
Milk, Bread	4
Milk, Eggs	3
Bread, Eggs	3
Milk, Bread, Eggs	3

Thus, the most frequent and highest item-set determining sub-item-set is {milk, bread, eggs}.

1.2

Confidence = {itemset (X and Y)}/ {itemset (X)}

Support= {itemset (X and Y)}/ transactions

Let {itemset 1, itemset 2} = X

Let {itemset 3} = Y

Then {itemset 1, itemset 2} -> itemset 3

X -> Y

Thus, two rules with confidence greater than or equal to 0.7 for an itemset containing 3 items are the following:

{Milk, Eggs} -> {Bread} implies

confidence = 3 (rows containing milk, eggs and bread)/3 (rows containing milk, eggs) =1 > 0.7

{Eggs, Bread} -> {Milk} implies

confidence = 3 (rows containing eggs, bread and milk)/3 (rows containing eggs, bread)

2. On describing discovered knowledge using classification

Classification is the process of learning a model that describes different classes of data and the classes should be pre-determined. Consider the following set of data records:

RID	Age	City	Gender	Education	Repeat Customer
101	20..30	NY	F	College	YES
102	31..40	NY	F	College	YES
103	51..60	NY	F	College	NO
104	20..30	LA	M	High school	NO
105	41..50	NY	F	College	YES
106	41..50	NY	F	Graduate	YES
107	20..30	LA	M	College	YES
108	20..30	NY	F	High school	NO
109	20..30	NY	F	College	YES
110	51..60	SF	M	College	NO

2.1

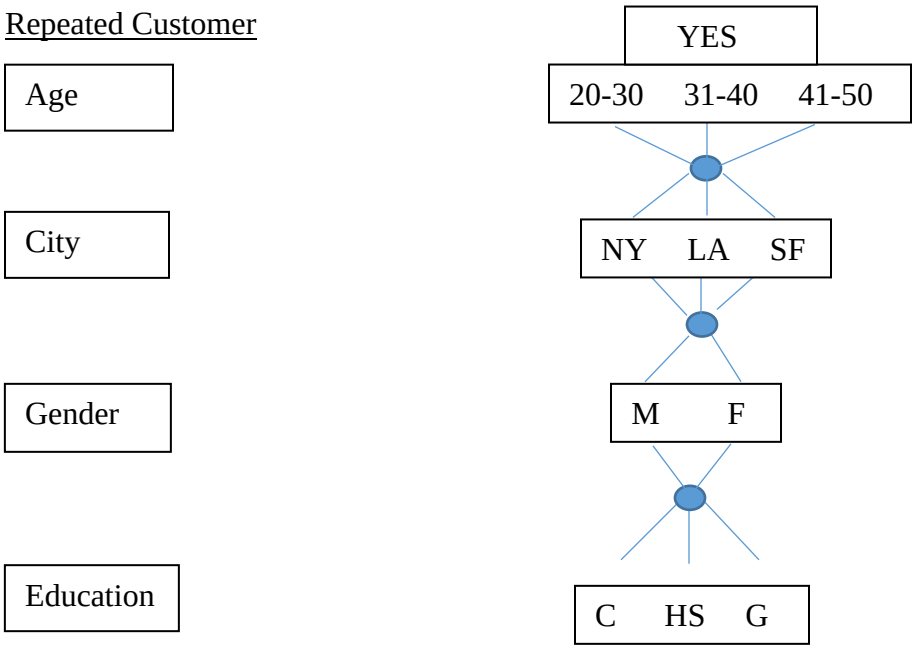
Assuming that the class attribute is Repeat Customer, apply a classification algorithm to this dataset.

2.1

The following video was used to supplement this question

https://www.youtube.com/watch?v=wt-X61BnUCA

The decision tree is as follows:



3. On describing discovered knowledge using clustering

Consider the following set of two-dimensional records:

RID	Dimension 1	Dimension 2
1	8	4
2	4	7
3	2	4
4	2	3
5	2	8
6	4	4

3.1

Use the K-means algorithm to cluster this dataset. You can use a value of 3 for K and can assume that the records with RIDs 1, 3, and 5 are used for the initial cluster centroids (means).

3.2

What is the difference between describing discovered knowledge using clustering and describing it using classification.

3.1

Numerical Example (manual calculation)

The basic step of k-means clustering is simple:

Iterate until *stable* (= no object move group):
1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance

We have k=3 clusters.

Distance [(x,y), (a,b)] = $\sqrt{(x-a)^2+(x-b)^2}$

RID	Distance to 1 (8,4)	Distance to 3 (2,3)	Distance to 5 (2,8)
1 (8,4)	0	-	
2 (4,7)	5	Sqrt(13)	Sqrt(5)
3 (2,4)	-	0	-
4 (2,3)	Sqrt(37)	1	5
5 (2,8)	-	-	0
6 (4,4)	4	2	Sqrt(20)

For dataset 2 (4,7), it will be assigned to C_3, distance 5, since it is the minimum

Now, we update the cluster centroid 5, so we calculate the mean: ((4+2)/2, (8+7)/2))= (3, 15/2