NAME – MOHIT AKHOURI

ROLL NO – 19UCC023

DATA MINING ASSIGNMENT – 2

**Q 1 )** -------------------------------------------------------------------------------------------------------

**Ans 1 )**

Given : Support Threshold = 20 %

**( a )** First, In order to calculate the frequent itemsets, the theory for the same is : Frequent itemsets are best represented by the dense columns in the graph ( that is frequently occurring items).

Frequent itemsets for each dataset are :

Dataset 1 = 5 ( Transactions overlap )

Dataset 2 = 4 ( Transactions 1,2,3,4 and 5 overlap with each other )

Dataset 3 = 10 ( All 10 transactions overlap )

Now, Data set **c** will produce the most number of frequent itemsets as it has the highest overlap.

**( b )** For longest frequent itemset, we have to see where there are more number of 1 item datasets and whether they overlap or not. Dataset **c** will produce the longest frequent itemset, as it has 2-length transactions in which each 1 length subpart overlaps with many other transactions.

**( c )** In dataset **c** , Item number of A occurs around 10000 times, which can be the maximum support for any dataset in this example.

**( d )** Looking at the datasets, we can see that dataset **b** has many frequent item columns of varying lengths, that is there are transactions of length 3 , 2 and 7. Therefore, dataset **b** will produce frequent itemsets containing items with widely varying support levels.

**( e )** Maximal frequent itemsets for each dataset can be as follows :

Dataset 1 = Each item occurs atmost 5 times in this dataset. Therefore maximal frequent itemsets for this dataset = 5.

Dataset 2 = Here Item A occurs 2 times, B and D occurs 3 times, C occurs 4 times and rest of items occurs 1 time. Therefore, maximal frequent itemsets for this dataset = 4.

Dataset 3 = Here item A and B occur 10 times, items D and E occur 1 time and item G occurs 2 times, therefore maximal frequent itemsets for this dataset = 10.

The dataset which will produce the most number of frequent itemsets = **b**

**( f )** Number of closed frequent itemsets for each dataset are as follows :

Dataset 1 = Items A to E occur around 5 times. Items F to J occur around 5 times, therefore number of closed frequent itemsets = 5.

Dataset 2 = Item A occur 2 times, B and D occur 3 times, C occur 4 times and rest of items occur only 1 time. But the number of closed frequent itemsets = 1.

Dataset 3 = Items A and B occur 10 times. Items D and E occur only 1 time. Item G occur 2 time, therefore number of closed frequent itemsets = 10.

Dataset which will produce the most number of closed frequent itemsets = **c**

**Q 2 )** -------------------------------------------------------------------------------------------------------

**Ans 2 ) Introduction to Data Mining : Exercise 6.10 – Q2**

**( a )** By treating each transaction ID as a market basket, the support for different itemsets can be calculated by counting their occurrences in the transactions. The support for different itemsets are as follows :

Support for itemset **{e}** = s({e}) = 8/10

Reason : item e occurs 8 times in the transactions out of 10.

Support for itemset **{b,d}** = s({b,d}) = 2/10

Reason : item {b,d} occur 2 times ( In transaction ID 0012 and 0022 ).

Support for itemset **{b,d,e}** = s({b,d,e}) = 2/10

Reason : item {b,d,e} occur 2 times ( In transaction ID 0012 and 0022 ).

**( b )** The confidence for the different association rules are as follows :

Confidence for the association rule ( bd -> e ) is : C(bd -> e) = 0.2/0.2 = 1 = **100%**

Confidence for the association rule ( e -> bd ) is : C(e -> bd) = 0.2/0.8 = 0.25 = **25%**

Since the association rules bd->e and e->bd are symmetric, but their confidence are not same. Hence, we can say that here confidence is not a symmetric measure.

**( c )** If we treat each customer ID as a market basket and each item as a binary variable ( 1 if item appears in at least one transaction and 0 otherwise ), Support count for different itemsets are as follows :

Support for itemset **{e}** = s({e}) = 4/5 = 0.8 = 80%

Support for itemset **{b,d}** = s({b,d}) = 5/5 = 1 = 100%

Support for itemset **{b,d,e}** = s({b,d,e}) = 4/5 = 0.8 = 80%

**( d )** Using the results of previous part c, Confidence for the different association rules can be as follows :

Confidence for the association rule ( bd -> e ) is : C(bd -> e) = 0.8/1 = 0.8 = **80%**

Confidence for the association rule ( e -> bd ) is : C(e -> bd) = 0.8/0.8 = 1 = **100%**

**( e )** Here, s1 and c1 are the support and confidence values of association rule **r**, if we treat each transaction ID as a market basket. Now, s2 and c2 are support and confidence values of **r** if we treat customer ID as a market basket. Looking from the above table and different values we got in the previous parts, we can see that there is no visible relationships between s1,s2,c1 and c2.

**Q 3 )** -------------------------------------------------------------------------------------------------------

**Ans 3 ) Introduction to Data Mining : Exercise 6.10 – Q3**

**( a )** Confidence for the rules φ -> A and A -> φ are as follows :

Confidence of φ -> A = Support of φ -> A : c(φ -> A) = s(φ -> A)

Confidence of A -> φ = **100%**

**( b )** The formulas for c1,c2 and c3 are as follows :

c1 = s(p U q) / s(p)

c2 = s(p U q U r) / s(p)

c3 = s(p U q U r) / s(p U r)

In the above formulas, **s** indicates the support count.

We can consider s(p) >= s(p U q) >= s(p U q U r), therefore the conclusion for the relationship between c1,c2 and c3 are as follows :

Relation : c1 >= c2 and c3 >= c2

Therefore we can say that c2 has the lowest confidence.

**( c )** Given the Assumption : Rules have identical support.

We can consider s(p U q) = s(p U q U r)

But we have deduced that s(p) >= s(p U r), therefore we say that : c3 >= (c1 = c2).

So the conclusion that we can make is that either for all the rules, confidence is the same or c3 has the highest confidence among all.

**( d )** Given some pre-conditions : c(A ->B) and c(B -> C) > minconf.

YES , it is possible for A -> C has a confidence less than minconf. It depends on the support of items A,B and C. The example for the same can be discussed as below :

Let s(A,B) = 60% , s(A) = 90% , s(A,C) = 20% , s(B) = 70% , s(B,C) = 50% and s(C) = 60%. Let minconf = 50%. Therefore the conclusions can be made as follows :

c(A -> B) = 66% which is greater than minconf.

c(B -> C) = 71% which is greater than minconf.

But confidence for A -> C, c(A -> C) = 22% which is less than minconf.

**Q 4 )** -------------------------------------------------------------------------------------------------------

**Ans 4 ) Introduction to Data Mining : Exercise 6.10 – Q6**

**( a )** Observing the table, we can conclude that there are six items in the dataset. The items are : { Milk, Beer, Diapers, Bread, Butter and Cookies }. Therefore, total number of rules that can be extracted from this dataset are : **602**.

**( b )** Assuming minsup > 0, we can see from the table that longest transaction which are transactions 6 and 9 contains **4** items each. Therefore, maximum size of the frequent itemset = **4**.

**( c )** To derive the maximum number of size-3 itemsets , we can use the probability concepts : 6C3 = **20**.

**( d )** The itemset of size 2 or larger and which has the largest support is as follows :

{Bread, Butter}

**( e )** The pair of items **a** and **b,** such that the rules : {a} -> {b} and {b} -> {a} have the same confidence are :

{Beer, Cookies}

{Bread, Butter}

**Q 5 )** -------------------------------------------------------------------------------------------------------

**Ans 5 ) Introduction to Data Mining : Exercise 6.10 – Q7**

**( a )** Using the Fk-1 x F1 merging strategy, all candidate 4-itemsets are as follows :

{1,2,3,4} , {1,2,3,5} , {1,2,3,6} , {1,2,4,5} , {1,2,4,6} , {1,2,5,6} , {1,3,4,5} , {1,3,4,6} , {2,3,4,5} , {2,3,4,6} , {2,3,5,6}

**( b )** Using the candidate procedure Apriori , candidate-4 itemsets are as follows :

{1,2,3,4} , {1,2,3,5} , {1,2,4,5} , {2,3,4,5} , {2,3,4,6}

**( c )** All candidate-4 itemsets that survive the candidate pruning step of Apriori algorithm are : {1,2,3,4} .

**Q 6 )** -------------------------------------------------------------------------------------------------------

**Ans 6 ) Introduction to Data Mining : Exercise 6.10 – Q12**

**( a )** The contingency tables for the rules after observing the given table are as follows :

Rule : {b} - > {c}

|  |  |  |
| --- | --- | --- |
|  | c | c’ |
| b | 3 | 4 |
| b’ | 2 | 1 |

Rule : {a} - > {d}

|  |  |  |
| --- | --- | --- |
|  | d | d’ |
| a | 4 | 1 |
| a’ | 5 | 0 |

Rule : {b} - > {d}

|  |  |  |
| --- | --- | --- |
|  | d | d’ |
| b | 6 | 1 |
| b’ | 3 | 0 |

Rule : {e} - > {c}

|  |  |  |
| --- | --- | --- |
|  | c | c’ |
| e | 2 | 4 |
| e’ | 3 | 1 |

Rule : {c} - > {a}

|  |  |  |
| --- | --- | --- |
|  | a | a’ |
| c | 2 | 3 |
| c’ | 3 | 2 |

**( b )** Using the contingency tables in the previous part, Ranking the rules :

( i ) Support : The table and rankings are as follows –

|  |  |  |
| --- | --- | --- |
| **Rules** | **Support** | **Rank** |
| b -> c | 0.3 | 3 |
| a -> d | 0.4 | 2 |
| b -> d | 0.6 | 1 |
| e -> c | 0.2 | 4 |
| c - > a | 0.2 | 4 |

( ii ) Confidence : The table and rankings are as follows –

|  |  |  |
| --- | --- | --- |
| **Rules** | **Confidence** | **Rank** |
| b -> c | 3/7 = 0.42 | 3 |
| a -> d | 4/5 = 0.80 | 2 |
| b -> d | 6/7 = 0.85 | 1 |
| e -> c | 2/6 = 0.33 | 5 |
| c - > a | 2/5 = 0.40 | 4 |

( iii ) Interest ( X -> Y ) = ( P(X,Y) / P(X) ) \* P(Y) .

The table and rankings are as follows –

|  |  |  |
| --- | --- | --- |
| **Rules** | **Interest** | **Rank** |
| b -> c | 0.214 | 3 |
| a -> d | 0.72 | 2 |
| b -> d | 0.771 | 1 |
| e -> c | 0.167 | 5 |
| c - > a | 0.2 | 4 |

( iv ) IS ( X -> Y ) = P(X,Y) / ( sqrt ( P(X) \* P(Y) ).

The table and rankings are as follows –

|  |  |  |
| --- | --- | --- |
| **Rules** | **IS** | **Rank** |
| b -> c | 0.507 | 3 |
| a -> d | 0.596 | 2 |
| b -> d | 0.756 | 1 |
| e -> c | 0.365 | 5 |
| c - > a | 0.4 | 4 |

( v ) Klosgen ( X -> Y ) = sqrt(P(X,Y)) \* (P(Y|X) – P(Y)).

The table and rankings are as follows –

|  |  |  |
| --- | --- | --- |
| **Rules** | **Klosgen** | **Rank** |
| b -> c | -0.039 | 2 |
| a -> d | -0.063 | 4 |
| b -> d | -0.033 | 1 |
| e -> c | -0.075 | 5 |
| c - > a | -0.045 | 3 |

( vi ) Odds ratio ( X -> Y ) = ( P(X,Y) \* P(X’Y’) ) / ( P(X,Y’) \* P(X’,Y) ).

The table and rankings are as follows –

|  |  |  |
| --- | --- | --- |
| **Rules** | **Odds Ratio** | **Rank** |
| b -> c | 0.375 | 2 |
| a -> d | 0 | 4 |
| b -> d | 0 | 4 |
| e -> c | 0.167 | 3 |
| c - > a | 0.444 | 1 |

**Q 7 )** -------------------------------------------------------------------------------------------------------

**Ans 7 ) Introduction to Data Mining : Exercise 7.8 – Q1**

**( a )** The binarized version of the given dataset is as follows :

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Good** | **Bad** | **Alcohol** | **Sober** | **Exceed Speed** | **None** | **Disobey Stop** | **Disobey Traffic** | **Belt =**  **No** | **Belt = Yes** | **Major** | **Minor** |
| 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |

**( b)** Observing from the binarized table, we can conclude that maximum width of each transaction in the binarized data – **5** .

**( c )** Assuming that support threshold is given as 30%. The number of candidate and frequent itemsets that will be generated are as follows :

The number of candidate itemsets from size one to size three can be : 10+28+3=41

The number of frequent itemsets from size one to size three can be : 8+10+0=18

**( d )** Assuming that support threshold is 30%, after generating the required binarized data for Traffic accident dataset, we come to the following two conclusions :

The number of candidate itemsets from size one to size three can be : 5+10+0=15

The number of frequent itemsets from size one to size three can be : 5+3+0=8

**( e )** We can see from part **c** and **d ,** number of candidate and frequent itemsets in part **d** is less than part **c .**

**Q 8 )** -------------------------------------------------------------------------------------------------------

**Ans 8 ) Introduction to Data Mining : Exercise 7.8 – Q10**

Assuming that there are no timing constraints imposed on the sequences and assuming support >= 50% . After observing the given sequence database, frequent subsequences that can be found are as follows :

< {A} > , < {B} > , < {C} > , < {D} > , < {E} > , < {A},{C} > , < {A},{D} > , < {A},{E} > , < {B},{C} > , < {B},{D} > , < {B},{E} > , < {C},{D} > , < {C},{E} > , < {D,E} >

**Q 9 )** -------------------------------------------------------------------------------------------------------

**Ans 9 ) Introduction to Data Mining : Exercise 7.8 – Q11**

**( a )** The table of event subsequences generated by various sensors are given. According to the given timing constraints and given sequence : < {1,2,3}, {2,4}, {2,4,5}, {3,5}, {6} >.

|  |  |
| --- | --- |
| **Sequence (w)** | **Whether this is subsequence ?** |
| < {1},{2},{3} > | YES |
| < {1,2,3,4} , {5,6} > | NO |
| < {2,4} , {2,4} , {6} > | YES |
| < {1} , {2,4} , {6} > | YES |
| < {1,2} , {3,4} , {5,6} > | NO |

**( b )** Given the timing constraints and table of event subsequences, checking for contiguous subsequences of the following sequence s .

* For sequence : < {1,2,3,4,5,6} , {1,2,3,4,5,6} , {1,2,3,4,5,6} >

|  |  |
| --- | --- |
| **Sequence (w)** | **Whether this is contiguous ?** |
| < {1},{2},{3} > | YES |
| < {1,2,3,4} , {5,6} > | YES |
| < {2,4} , {2,4} , {6} > | YES |
| < {1} , {2,4} , {6} > | YES |
| < {1,2} , {3,4} , {5,6} > | YES |

* For sequence : < {1,2,3,4} , {1,2,3,4,5,6} , {3,4,5,6} >

|  |  |
| --- | --- |
| **Sequence (w)** | **Whether this is contiguous ?** |
| < {1},{2},{3} > | YES |
| < {1,2,3,4} , {5,6} > | YES |
| < {2,4} , {2,4} , {6} > | YES |
| < {1} , {2,4} , {6} > | YES |
| < {1,2} , {3,4} , {5,6} > | YES |

* For sequence : < {1,2} , {1,2,3,4} , {3,4,5,6} , {5,6} >

|  |  |
| --- | --- |
| **Sequence (w)** | **Whether this is contiguous ?** |
| < {1},{2},{3} > | YES |
| < {1,2,3,4} , {5,6} > | YES |
| < {2,4} , {2,4} , {6} > | NO |
| < {1} , {2,4} , {6} > | YES |
| < {1,2} , {3,4} , {5,6} > | YES |

* For sequence : < {1,2,3} , {2,3,4,5} , {4,5,6} >

|  |  |
| --- | --- |
| **Sequence (w)** | **Whether this is contiguous ?** |
| < {1},{2},{3} > | NO |
| < {1,2,3,4} , {5,6} > | NO |
| < {2,4} , {2,4} , {6} > | NO |
| < {1} , {2,4} , {6} > | YES |
| < {1,2} , {3,4} , {5,6} > | YES |

**Q 10 )** -----------------------------------------------------------------------------------------------------

**Ans 10 ) Introduction to Data Mining : Exercise 8.7 – Q7**

Given : There is a dataset with the following properties –

* There are **m** points and **K** clusters
* Half of points and clusters are in “more dense region”
* Half of points and clusters are in “less dense region”
* The two regions are well-separated from each other

The conclusion we obtain after observing the above properties and the available given options is : More centroids should be allocated to the denser region. The reason for the same is – Less dense region require more centroids if we want the squared error is to be minimized.

**Q 11 )** -----------------------------------------------------------------------------------------------------

**Ans 11 ) Introduction to Data Mining : Exercise 8.7 – 10**

Cosine measure is NOT the appropriate similarity measure to use with K-means clustering for time-series date due to these reasons :

* Time series data is a high-dimensional and denser data.
* Cosine measure is appropriate for sparse data.

Alternate Solution – **Euclidean distance** could be used with K-means clustering when the magnitude of time series is important. **Correlation** would be appropriate if shapes of the time series data is important.