CSC: 522 Automated Learning and Data Analysis (Instructor: Dr. Min Chi)

Submitted by:

1. Himangshu Ranjan Borah (Unity ID: hborah, Student ID: 200105222)

2. Sukriti Sharma (Unity ID: ssharm18, Student ID: 200111439)

3. Sushma Ravichandran (Unity ID: sravich, Student ID: 200106275)

HOMEWORK 4

1 a) Items are –

{M},{O},{N},{K},{E},{Y},{D},{A},{U},{C},{I}

Number of items= 11

The number of candidate itemsets are 2^{11} (including null) = 2048

b) USING APRIORI

1-itemsets C1

Item	Sup
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

L1

Item	sup
M	3
O	3
K	5
E	4
Y	3

C2

Itemset	sup
{M,O}	1
{M,K}	3
{M,E}	2
{M,Y}	2
{O,K}	3
{O,E}	3
{O,Y}	2
{K,E}	4
{K,Y}	3
{E,Y}	2

L2

Itemset	sup
{M,K}	3
{O,K}	3
{O,E}	3
{K,E}	4
{K,Y}	3

C3

Itemset	sup
{O,K,E}	3

L3

Itemset	sup
{O,K,E}	3

L1 +L2+L3 contain all frequent itemsets with min. support 3

c) Generating frequent itemset for remaining parts of the question

1-itemsets C1

Item	sup
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

L1

Item	sup
M	3
O	3
K	5
E	4
Y	3
N	2
C	2

C2

Itemset	sup
{M,O}	1
{M,K}	3
{M,E}	2
{M,Y}	2
{M,N}	1
{M,C}	1
{O,K}	3
{O,E}	3
{O,Y}	2
{O,N}	2
{O,C}	1
{K,E}	4
{K,Y}	3
{K,N}	2
{K,C}	2
{E,Y}	2
{E,N}	2
{E,C}	1
{Y,N}	2
{Y,C}	1
{N,C}	0

L2

Itemset	sup
{M,K}	3
{M,E}	2
{M,Y}	2
{O,K}	3
{O,E}	3
{O,Y}	2
{O,N}	2
{K,E}	4
{K,Y}	3
{K,N}	2
{K,C}	2
{E,Y}	2
{E,N}	2
{Y,N}	2

C3

Itemset	sup
{M,K,E}	2
{M,K,Y}	2
{M,E,Y}	1
{O,K,E}	3
{O,K,Y}	2
{O,K,N}	2
{O,E,Y}	2
{O,E,N}	2
{O,Y,N}	2
{K,E,Y}	2
{K,E,N}	2
{K,Y,N}	2
{E,Y,N}	2

L3

Itemset	sup
{M,K,E}	2
{M,K,Y}	2
{O,K,E}	3
{O,K,Y}	2
{O,K,N}	2
{O,E,Y}	2
{O,E,N}	2
{O,Y,N}	2
{K,E,Y}	2
{K,E,N}	2
{K,Y,N}	2
{E,Y,N}	2

C4

Itemset	Sup
{O,K,E,Y}	2
{O,K,E,N}	2
{O,K,Y,N}	2
{O,Y,E,N}	2
{K,E,Y,N}	2

L4

Itemset	Sup
{O,K,E,Y}	2
{O,K,E,N}	2
{O,K,Y,N}	2
{O,Y,E,N}	2
{K,E,Y,N}	2

C5

Itemset	Sup
{O,K,E,Y,N}	2

L5

Itemset	Sup
{O,K,E,Y,N}	2

Frequent itemsets are

Item	sup
M	3
O	3
K	5
E	4
Y	3
N	2
C	2
{M,K}	3
{M,E}	2
{M,Y}	2
{O,K}	3
{O,E}	3
{O,Y}	2
{O,N}	2
{K,E}	4
{K,Y}	3
{K,N}	2
{K,C}	2
{E,Y}	2
{E,N}	2
{Y,N}	2
{M,K,E}	2
{M,K,Y}	2
{O,K,E}	3
{O,K,Y}	2
{O,K,N}	2
{O,E,Y}	2
{O,E,N}	2
{O,Y,N}	2
{K,E,Y}	2
{K,E,N}	2
{K,Y,N}	2
{E,Y,N}	2
{O,K,E,Y}	2
{O,K,E,N}	2
{O,K,Y,N}	2
{O,Y,E,N}	2
{K,E,Y,N}	2
{O,K,E,Y,N}	2

In the table above, Closed frequent itemsets have been marked in yellow and red. The maximal frequent itsemsets are marked in red.

Closed frequent itemsets are those which which are frequent and do not have any immediate superset with exactly same support count as that set. So we look at the frequent itemsets table above. Consider frequent itemset {K} . None of its immediate supersets in the frequent itemset table have same support as {K}. The immediate supersets of {K} which are not frequent, anyway cannot have the same support as {K}, as {K} is frequent. Hence {K} is a closed frequent itemset. Similarly we find all.

The closed frequent itemsets are – {K},{M,K},{K,E},{K,Y},{K,C},{M,K,E},{M,K,Y},{O,K,E},{O,K,E,Y,N}

d) A maximal frequent itemset is one which is frequent and for which none of its immediate supersets are frequent. From the above frequent itemsets table we can see {K,C} does not have any immediate superset which is in the frequent itemset table. Hence it is a maximal itemset. Similarly we find all maximal frequent itemsets

maximal frequent itsemsets are {K,C},{M,K,E},{M,K,Y},{O,K,E,Y,N}

e) Since we need to get association rules of the type item1,item2 -> item 3, we need to consider 3 itemsets . Since we need to get support of 60% , and there are 5 transactions, atleast 3 transactions must contain the itemset. As seen in the table above 3-itemset {0,K,E} has a minimum support of 60%.

The association rules satisfying the metarule are

1. O,K -> E
2. O,E -> K
3. K,E -> O

Out of these we need to find rules that have min confidence of 80%.

1. For rule {O,K } -> {E}
   c({O,K } -> {E}) = S(O,K,E) / S(O,K)
   Using the tables computed above, c({O,K } -> {E}) =3/3 =1
2. For rule {O,E} -> {K}
   c({O,E } -> {K}) = S(O,K,E) / S(O,E)
   Using the tables computed above, c({O,E } -> {K}) = 3/3=1
3. For rule { K,E}->{O}
   c({ K,E}->{O}) = S(O,K,E) / S(K,E)
   Using the tables computed above, c({ K,E}->{O}) = ¾ =0.75

Hence the strong association rules matching the metarule item1, item2 =⇒ item3 with minimum support of 60% And confidence of 80% are:

1. {O,K} ->{ E}
2. {O,E} -> {K}

Ans. 2

Weather Condition	Driver’s Condition	Traffic Violation	Seat Belt	Crash Severity
Good	Alcohol-impaired	Exceed speed limit	No	Major
Bad	Sober	None	Yes	Minor
Good	Sober	Disobey stop sign	Yes	Minor
Good	Sober	Exceed speed limit	Yes	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Disobey stop sign	Yes	Minor
Bad	Alcohol-impaired	None	Yes	Major
Good	Sober	Disobey traffic signal	Yes	Major
Good	Alcohol-impaired	None	No	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Exceed speed limit	Yes	Major
Bad	Sober	Disobey stop sign	Yes	Minor

(a) Show a binarized version of the data set.

Weather Condition = Good	Weather Condition = Bad	Driver’s Condition =Alcohol	Driver’s Condition =Sober	Traffic Violation = Exceed Speed Limit	Traffic Violation = Disobey traffic signal	Traffic Violation = Disobey stop sign	Traffic Violation = None	Seat Belt = Yes	Seat Belt = No	Crash Severity = major	Crash Severity = Minor
1	0	1	0	1	0	0	0	0	1	1	0
0	1	0	1	0	0	0	1	1	0	0	1
1	0	0	1	0	0	1	0	1	0	0	1
1	0	0	1	1	0	0	0	1	0	1	0
0	1	0	1	0	1	0	0	0	1	1	0
1	0	1	0	0	0	1	0	1	0	0	1
0	1	1	0	0	0	0	1	1	0	1	0
1	0	0	1	0	1	0	0	1	0	1	0
1	0	1	0	0	0	0	1	0	1	1	0
0	1	0	1	0	1	0	0	0	1	1	0
1	0	1	0	1	0	0	0	1	0	1	0
0	1	0	1	0	0	1	0	1	0	0	1

(b) The maximum width of each transaction in the binarized data sheet is 5 because in the original data, the total number of attributes is 5.

For each attribute is as follows:

(i) Weather Condition
(ii) Driver's Condition
(iii) Traffic Violation
(iv) Seat belt
(v) Crash Severity

(c) Assuming that support threshold is 30%, how many candidate and frequent itemsets will be generated?

Candidate 1 ItemSets

Item	Count	Support	Greater than threshold?
Weather Condition = Good	7	58.3333%	Y
Weather Condition = Bad	5	41.6667%	Y
Driver's Condition = Alcohol- Impaired	5	41.6667%	Y
Driver's Condition = Sober	7	58.3333%	Y
Traffic Violation = Disobey Stop Sign	3	25.0000%	N
Traffic Violation = Disobey Traffic Signal	3	25.0000%	N
Traffic Violation = Exceed Speed Limit	3	25.0000%	N
Traffic Violation = None	3	25.0000%	N
Seat Belt = Yes	8	66.6667%	Y
Seat Belt = No	4	33.3333%	Y
Crash Severity = Major	8	66.6667%	Y
Crash Severity = Minor	4	33.3333%	Y

Candidate 2 Item Sets

Item			Count	Support	Greater than Threshold?
Weather Condition	Good	Driver's Condition =Alcohol-Impaired	4	33.33333	Y

		Driver's Condition = Sober	3	25	N
		Seat Belt = Yes	5	41.66667	Y
		Seat Belt = No	2	16.66667	N
		Crash Severity = Major	5	41.66667	Y
		Crash Severity = Minor	2	16.66667	N
Weather Condition	Bad	Driver's Condition =Alcohol-Impaired	1	8.333333	N
		Driver's Condition = Sober	4	33.33333	Y
		Seat Belt = Yes	3	25	N
		Seat Belt = No	2	16.66667	N
		Crash Severity = Major	3	25	N
		Crash Severity = Minor	2	16.66667	N
Driver's Condition	Alcohol - Impaired	Seat Belt = Yes	3	25	N
		Seat Belt = No	2	16.66667	N
		Crash Severity = Major	4	33.33333	Y
		Crash Severity = Minor	1	8.333333	N
Driver's Condition	Sober	Seat Belt = Yes	5	41.66667	Y
		Seat Belt = No	2	16.66667	N
		Crash Severity = Major	4	33.33333	Y
		Crash Severity = Minor	3	25	N
Seat Belt	Yes	Crash Severity = Major	4	33.33333	Y
		Crash Severity = Minor	4	33.33333	Y
Seat Belt	No	Crash Severity = Major	4	33.33333	Y

		Crash Severity = Minor	0	0	N

Candidate 3 Item Sets:

Item	Count	Support	Greater than Threshold?
Weather Condition: Good Driver’s Condition: Alcohol-Impaired Crash Severity: Major	3	25%	N
Weather Condition: Good Seat Belt: Yes Crash Severity: Major	3	25%	N
Driver’s Condition: Sober Seat Belt: Yes Crash Severity: Major	2	16.6666%	N

The total number of Frequent Item Sets are: 8 + 10 + 0 = 18

And the total number of Candidate Item sets are: 12 + 24 + 3 = 39

(d) Create a data set that contains only the following asymmetric binary attributes: (Weather = Bad, Driver’s Condition = Alcohol−Impaired, Traffic Violation = Yes, Seat Belt = No, Crash Severity = Major) For Traffic violation, only None has a value of 0. The rest of the attribute values are assigned to 1. Assuming that support threshold is 30%, how many candidate and frequent itemsets will be generated?

Ans.

Weather Condition: Bad	Driver’s Condition: Alcohol Impaired	Traffic Violation: Yes	Seat belt: No	Crash Severity: Major
0	1	1	1	1
1	0	0	0	0
0	0	1	0	0
0	0	1	0	1
1	0	1	1	1
0	1	1	0	0
1	1	0	0	1
0	0	1	0	1
0	1	0	1	1
1	0	1	1	1

0	1	1	0	1
1	0	1	0	0

For 1-Item Sets

Item	Count	Support	Greater than Threshold?
Weather Condition: Bad	5	41.6667%	Y
Driver’s Condition: Alcohol Impaired	5	41.6667%	Y
Traffic Violation: Yes	9	75%	Y
Seat belt: No	4	33.3333%	Y
Crash Severity: Major	8	66.6667%	Y

For 2-Item Sets

Weather Condition: Bad Drivers Condition: Alc. Imp	1	8.3333%	N
Weather Condition: Bad Traffic Violation: Yes	3	25.0000%	N
Weather Cond: Bad Seat belt: No	2	16.6667%	N
Weather Cond: Bad Crash Severity: Major	3	25.0000%	N
Driver’s Condition: Alc. Imp Traffic Violation: Yes	3	25%	N
Driver’s Condition: Alc. Imp Seat belt: No	2	16.6667%	N
Driver’s Condition: Alc. Imp Crash Severity: Major	4	33.3333%	Y
Traffic Violation: Yes Seat Belt: No	3	25%	N
Traffic Violation: Yes Crash Severity: Major	6	50%	Y
Seat Belt: No Crash Severity: Major	4	33.3333%	Y

There are no more attributes left to be added onto the variables that haven’t already been added into the previous subsets. So we conclude that:
The total number of frequent item sets are: 5 + 3 = 8

The total number of candidate sets are: 5 + 10 = 15

(e) Compare the number of candidate and frequent itemsets generated in parts (c) and (d).
Ans.
Part (c)
Candidate Item Sets: 39
Frequent Item Sets: 18

Part (d)
Candidate Item Sets: 8
Frequent Item Sets: 15
There is a drastic reduction in the number of candidate item sets in Part (d) because it is a condensed version where all the attributes have been reduced to a specific binary category. This also helped Part d to increase the number of frequent item sets despite it being a more specific subset.
When it comes to specific attributes for example, traffic violation part (d) has thoroughly reduced the original dataset to a binary form. In the original version traffic violation could take any of four different values. While this may help in the short term, during data analysis at a later stage this important information is lost and cannot be retrieved.

Ans 3:

Code attached with the assignment submission folder.