Recitation Problems

Q2. Consider the data set shown in Table 6.22.

Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

(a) Compute the support for itemsets {e}, {b,d}, and {b,d,e} by treating each transaction ID as a market basket.
Answer:
S({e}) = σ(e)/N=8/10 = 0.8
S({b,d}) = σ(bUd)/N =2/10 = 0.2
S({b,d,e}) = σ(bUdUe)/N =2/10 = 0.2

(b) Use the results in part (a) to compute the confidence for the association rules {b,d} → {e} and {e} →{b ,d}. Is confidence a symmetric measure?
Answer:
C(b,d→e) = σ(b,d U e)/σ(b,d)= 0.2/0.2 = 100%
C(e→b,d) = σ(b,d U e)/σ(e)= 0.2/0.8 = 25%
So, by above example we know that confidence is not symmetric.

(c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in atleast one transaction bought by the customer, and 0 otherwise.)
Answer:
S({e}) = σ(e)/N=4/5 = 0.8
S({b,d}) = σ(bUd)/N =5/5 = 1
S({b,d,e}) = σ(bUdUe)/N =4/5 = 0.8

(d) Use the results in part (c) to compute the confidence for the association rules {b, d} → {e} and {e}→ {b,d,}.
Answer:
C(b,d→e) = σ(b,d U e)/σ(b,d)= 0.8/1 = 80%
C(e→b,d) = σ(b,d U e)/σ(e)= 0.8/0.8 = 100%

(e) Suppose s1 and c1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s2 and c2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s1 and s2 or c1 and c2.
Answer:
No, there is no relationship between s1,s2,c1 and c2.

Q6. Consider the market basket transactions shown in Table 6.23.

Table 6.23. Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

(a) What is the maximum number of association rules that can be extracted
from this data (including rules that have zero support)?
Answer: The maximum number of association rules that can be extracted is calculated by given formula:

$$R=3^n - 2^{n+1} +1$$
$$=3^6 - 2^{6+1} +1$$
$$=602$$

(b) What is the maximum size of frequent itemsets that can be extracted
(assuming minsup > 0)?
Answer: The maximum size of frequent itemsets that can be extracted is 4 because the longest transaction is of size 4.

(c) Write an expression for the maximum number of size-3 itemsets that can
be derived from this data set.
Answer: $({}_{3}{}^{6})$ = 20.

(d) Find an itemset (of size 2 or larger) that has the largest support.
Answer: The itemset which has maximum support is {Bread, Butter} with support =5.

(e) Find a pair of items ,a and b, such that the rules {a} → {b} and {b} →{a} have the same confidence.
Answer: {Beer, Cookies}=2/4, {Butter, Cookies}=1/4, {Milk, Diaper}=4/5,{Milk, Bread}=3/5,{Diaper, Bread}=3/5.

Q7. Consider the following set of frequent 3-itemsets:
{1, 2, 3}, {1, 2, 4), {r, 2, 5}, {r, 3, 4},{ 1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5 } .
Assume that there are only five items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the F_{k-1} x F_k merging strategy.
Answer:
{1,2,3,5},{1,2,3,4},{1,2,4,5},{1,3,4,5},{2,3,4,5}.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.
Answer: {1,2,3,5},{1,2,3,4},{1,2,4,5},{1,3,4,5},{2,3,4,5}.

(c) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori, algorithm.
Answer: {1,2,3,4}

Q9. The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in Figure 6.32.

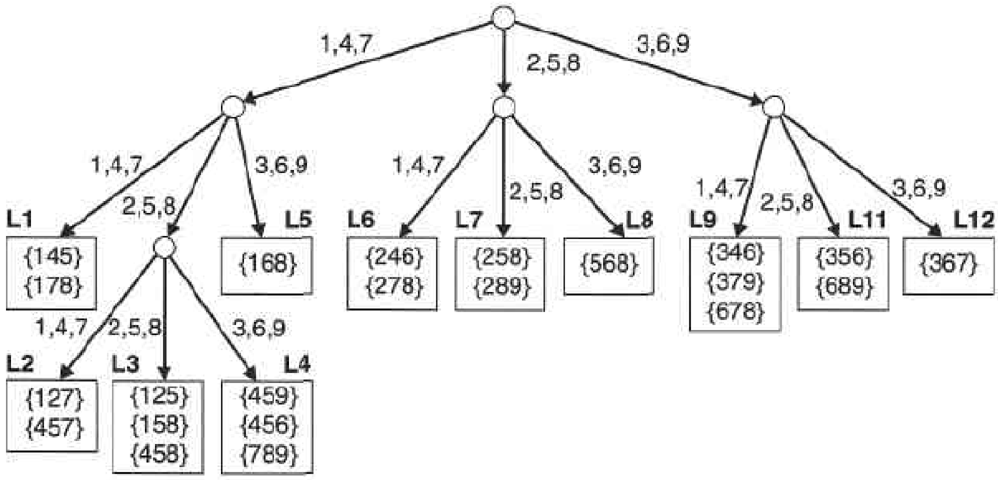
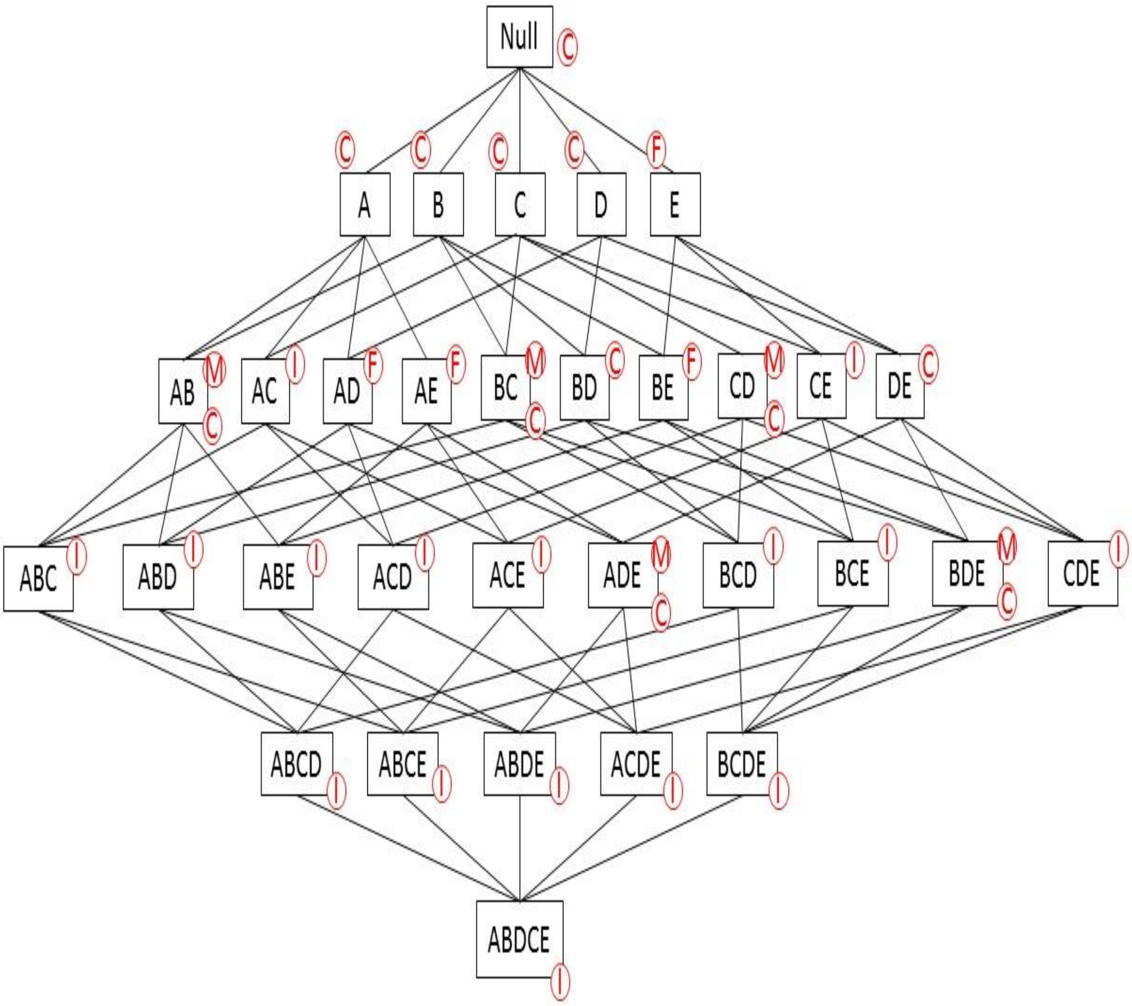


Figure 6.32. An example of a hash tree structure.

(a) Given a transaction that contains items {1,3,4,5,8}, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?
Answer: The leaf nodes visited that will be visited are L1, L3, L5, L9, and L11.

(b) Use the visited leaf nodes in part (b) to determine the candidate itemsets that are contained in the transaction {1,3,4,5,8}.
Answer: The candidates contained in the transaction are {1, 4, 5}, {1, 5, 8}, and {4, 5, 8}.

Q11. Given the lattice structure shown in Figure 6.33 and the transactions given in Table 6.24,label each node with the following letter(s):
M if the node is a maximal frequent itemset,
C if it is a closed frequent itemset,
A if it is frequent but neither maximal nor closed,
I if it is infrequent.
Assume that the support threshold is equal to 30
Answer:



Q12. The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.

Table 6.25. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

(a) Draw a contingency table for each of the following rules using the transactions shown in Table 6.25.

Rules: {b}→{c}, {a} → {d}, {b} → {d}, {e} → {c}, {c}→{a}.

Answer:

	b	$\bar{c}$
b	3	4
$\bar{c}$	2	1

	a	$\bar{d}$
a	4	1
$\bar{d}$	5	0

	b	$\bar{d}$
b	6	1
$\bar{d}$	3	0

	e	$\bar{c}$
e	2	4
$\bar{c}$	3	1

	c	$\bar{a}$
c	2	3
$\bar{a}$	3	2

(b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

1.Support.

Answer:

Rules	**Support**	**Rank**
{b}→{c}	3/10	3
{a} → {d}	4/10	2
{b} → {d}	6/10	1
{e} → {c}	2/10	4
{c}→{a}	2/10	4

2. Confidence.
Answer:

Rules	Confidence	Rank
{b}→{c}	3/7	3
{a} → {d}	4/5	2
{b} → {d}	6/7	1
{e} → {c}	2/6	5
{c}→{a}	2/5	4

3. Interest. P(X,Y)/P(X) * P(Y)
Answer:

Rules	Interest	Rank
{b}→{c}	(0.3/0.7)*0.5=0.214	3
{a} → {d}	(0.4/0.5)*0.9=0.72	2
{b} → {d}	(0.6/0.7)*0.9=0.771	1
{e} → {c}	(0.2/0.6)*0.5=0.167	5
{c}→{a}	(0.2/0.5)*0.5=0.2	4

4. IS P(X,Y)/√(P(X)*P(Y))
Answer:

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

5. Klosgen. √(P(X,Y))*(P(Y|X)-P(Y))
Answer:

Rules	Klosgen	Rank
{b}→{c}	-0.039	2
{a} → {d}	-0.063	4
{b} → {d}	-0.33	1
{e} → {c}	-0.075	5
{c}→{a}	-0.045	3

6. Odds ratio. P(X,Y) P(X,Y)/P(X,Y)P(X,Y)
Answer:

Rules	Odds ratio	Rank
{b}→{c}	0.375	2
{a} → {d}	0	4
{b} → {d}	0	4
{e} → {c}	0.167	3
{c}→{a}	0.44	1

Q18. Table 6.26 shows a 2 x 2 x 2 contingency table for the binary variables A and B at different values of the control variable C.

Table 6.26. A Contingency Table.

			A	
			1	0
C = 0	B	1	0	15
		0	15	30
C = 1	B	1	5	0
		0	0	15

(a) Compute the phi coefficient for A and B when C=0, C =1, and C= 0 or 1.

Answer:

When C=0, we have

$Phi = P(A,B)-P(A)P(B)/\sqrt{}(P(A)P(B)(1-P(A)(1-P(B))$

$= 0-0.25*0.25/\sqrt{}(0.25*0.25*0.75*0.75)$

$= -0.33$

When C=1, we have

$Phi = P(A,B)-P(A)P(B)/\sqrt{}(P(A)P(B)(1-P(A)(1-P(B))$

$= 0.25-0.0625/\sqrt{}(0.0625*0.5625)$

$= 1$

When C=0 or 1

$Phi = 0$

(b) What conclusions can you draw from the above result?

Answer: If confounding factors are not considered then the result may show discrepancies.

Q19. Consider the contingency tables shown in Table 6.27.

Table 6.27. Contingency tables for Exercise 19.

	B	$\overline{B}$
A	9	1
$\overline{A}$	1	89

(a) Table I.

	B	$\overline{B}$
A	89	1
$\overline{A}$	1	9

(b) Table II.

(a) For table I, compute support, the interest measure, and the correlation coefficient for the association pattern {A, B}. Also, compute the confidence of rules A → B and B → A.
Answer:
s(A) = 0.1, s(B) = 0.9, s(A,B) = 0.09.
I(A,B) = 9, ϕ(A,B) = 0.89.
c(A → B) = 0.9, c(B → A) = 0.9.

(b) For table II, compute support, the interest measure, and the correlation coefficient for the association pattern {A, B}. Also, compute the confidence of rules A → B and B → A.
Answer:
s(A) = 0.9, s(B) = 0.9, s(A,B) = 0.89.
I(A,B) = 1.09, ϕ(A,B) = 0.89.
c(A → B) = 0.98, c(B → A) = 0.98.

(c) What conclusions can you draw from the results of (a) and (b)?
Answer: We can conclude that phi co-efficient is invariant as it takes into consideration both absence and presence of items.

Q20. Consider the relationship between customers who buy high-definition televisions and exercise machines as shown in Tables 6.19 and 6.20.
(a) Compute the odds ratios for both tables.
Answer:
For Table 6.19, odds ratio = 1.4938.
For Table 6.20, the odds ratios are 0.8333 and 0.98.

(b) Compute the coefficient for both tables.
Answer:
For table 6.19, φ = 0.098.
For Table 6.20, the φ-coefficients are -0.0233 and -0.0047.
(c) Compute the interest factor for both tables.
Answer:
For table 6.19, φ = 1.0784.
For Table 6.20, the φ-coefficients are 0.88 and 0.9971

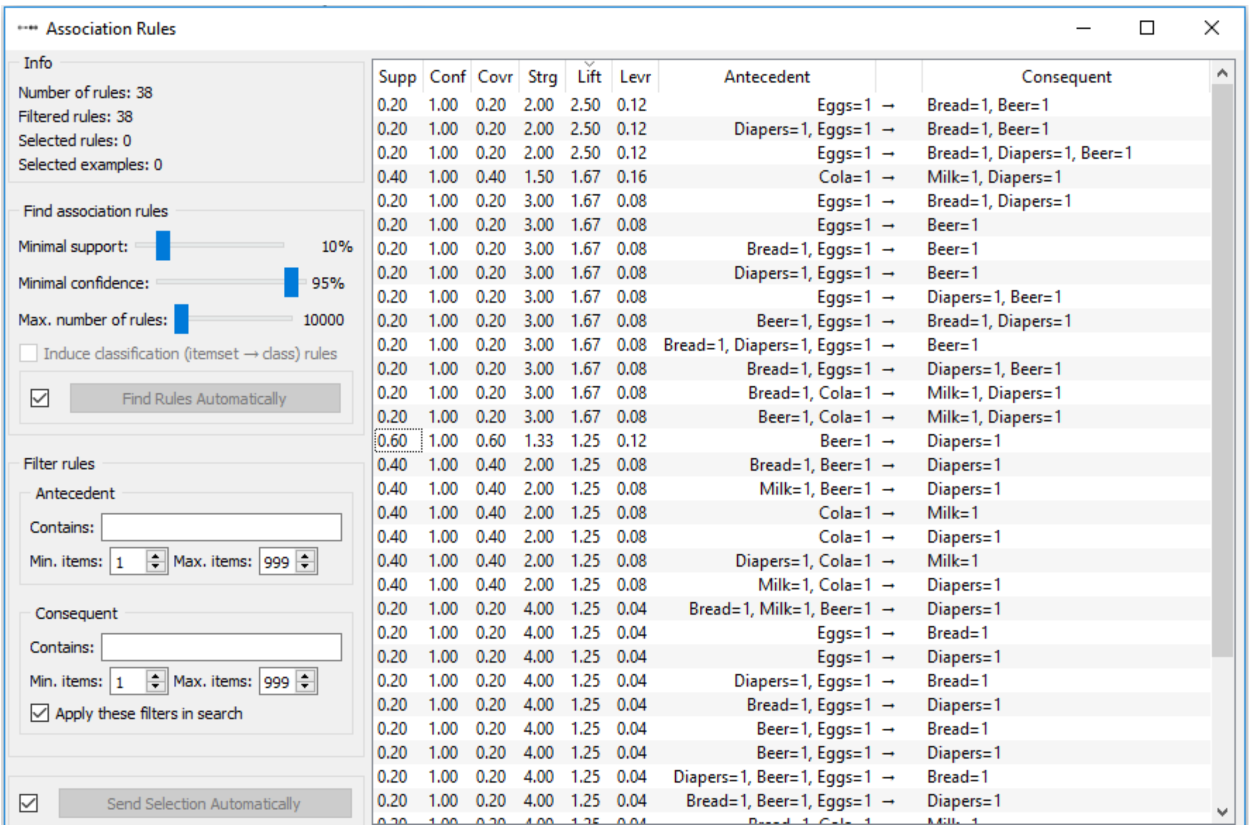
For each of the measures given above, describe how the direction of association changes when data is pooled together instead of being stratified.
Answer:
When data is pooled together, association direction changes from negative to positive.

Practicum Problems:

2.1 Problem 1

Load the market-basket sample dataset into the Orange application, and run both frequent itemset as well as association rule modules. Set the support threshold to 10% and observe the antecedent in the rules with the highest lift. What item is observed to be there, and what is its support? Is this a valuable association rule? Why or why not?

Solution :



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.20	1.00	0.20	2.00	2.50	0.12	Eggs=1	→	Bread=1, Beer=1
0.20	1.00	0.20	2.00	2.50	0.12	Diapers=1, Eggs=1	→	Bread=1, Beer=1
0.20	1.00	0.20	2.00	2.50	0.12	Eggs=1	→	Bread=1, Diapers=1, Beer=1
0.40	1.00	0.40	1.50	1.67	0.16	Cola=1	→	Milk=1, Diapers=1
0.20	1.00	0.20	3.00	1.67	0.08	Eggs=1	→	Bread=1, Diapers=1
0.20	1.00	0.20	3.00	1.67	0.08	Eggs=1	→	Beer=1
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Eggs=1	→	Beer=1
0.20	1.00	0.20	3.00	1.67	0.08	Diapers=1, Eggs=1	→	Beer=1
0.20	1.00	0.20	3.00	1.67	0.08	Eggs=1	→	Diapers=1, Beer=1
0.20	1.00	0.20	3.00	1.67	0.08	Beer=1, Eggs=1	→	Bread=1, Diapers=1
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Diapers=1, Eggs=1	→	Beer=1
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Eggs=1	→	Diapers=1, Beer=1
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Cola=1	→	Milk=1, Diapers=1
0.20	1.00	0.20	3.00	1.67	0.08	Beer=1, Cola=1	→	Milk=1, Diapers=1
0.60	1.00	0.60	1.33	1.25	0.12	Beer=1	→	Diapers=1
0.40	1.00	0.40	2.00	1.25	0.08	Bread=1, Beer=1	→	Diapers=1
0.40	1.00	0.40	2.00	1.25	0.08	Milk=1, Beer=1	→	Diapers=1
0.40	1.00	0.40	2.00	1.25	0.08	Cola=1	→	Milk=1
0.40	1.00	0.40	2.00	1.25	0.08	Cola=1	→	Diapers=1
0.40	1.00	0.40	2.00	1.25	0.08	Diapers=1, Cola=1	→	Milk=1
0.40	1.00	0.40	2.00	1.25	0.08	Milk=1, Cola=1	→	Diapers=1
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Milk=1, Beer=1	→	Diapers=1
0.20	1.00	0.20	4.00	1.25	0.04	Eggs=1	→	Bread=1
0.20	1.00	0.20	4.00	1.25	0.04	Eggs=1	→	Diapers=1
0.20	1.00	0.20	4.00	1.25	0.04	Diapers=1, Eggs=1	→	Bread=1
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Eggs=1	→	Diapers=1
0.20	1.00	0.20	4.00	1.25	0.04	Beer=1, Eggs=1	→	Bread=1
0.20	1.00	0.20	4.00	1.25	0.04	Beer=1, Eggs=1	→	Diapers=1
0.20	1.00	0.20	4.00	1.25	0.04	Diapers=1, Beer=1, Eggs=1	→	Bread=1
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Beer=1, Eggs=1	→	Diapers=1

Here the highest lift is 2.50 and the antecedent for that is Eggs, Diapers. The association rules for lift 2.50 are,

Eggs -> Bread, Beer

Diapers, Eggs -> Bread, Beer

Eggs -> Bread, Diaper, Beer

The support here is 0.20

1. This association rule is valuable because marketing will have to be done only on Eggs or Eggs, Diapers. When the support threshold is set to 10% we have Eggs and Eggs, Diaper as antecedents with highest lift=2.5 and support=20%. Also lift highlights rules that are rare but informative.
2. Since the consequents gained are more the sales overall will be higher which will be the total sales cost of antecedents + consequents.
3. High-confidence rules are sometimes misleading as the confidence measure neglects the support of the itemset which is in the rule consequent. If we apply a metric lift then we can address to this problem.

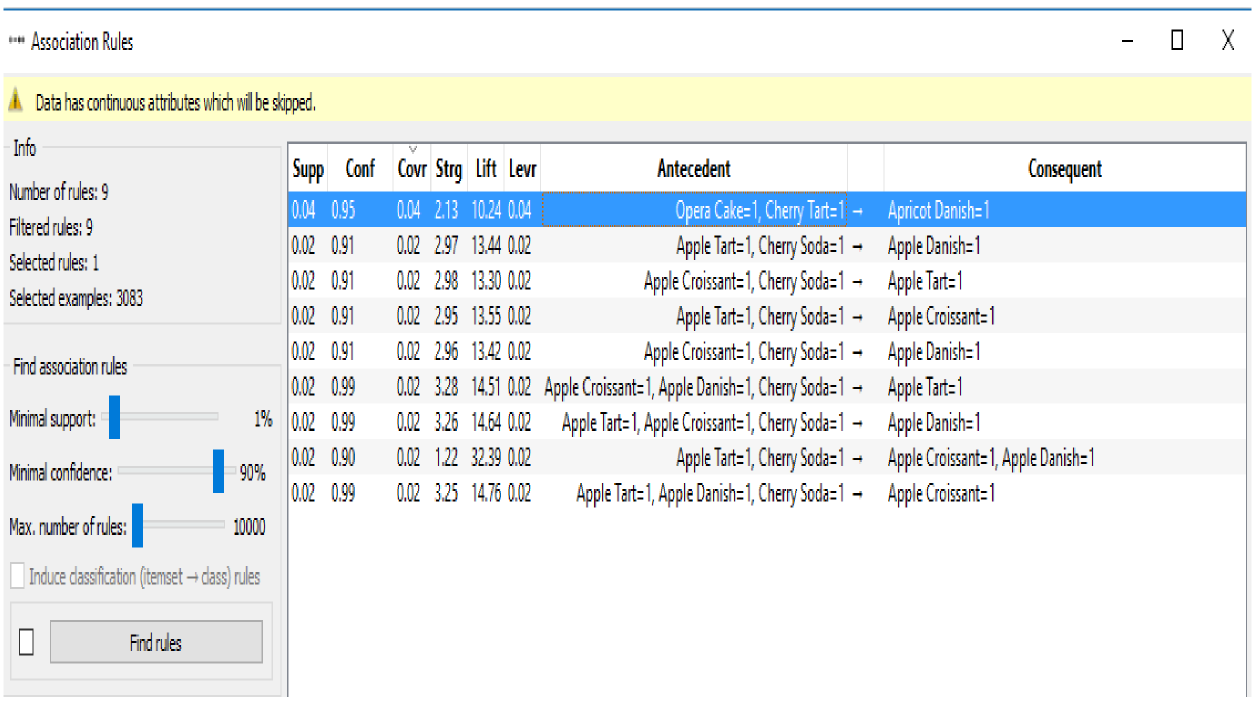
2.2 Problem 2

Load the Extended Bakery dataset (75000-out2-final.csv) into the Orange application, and run both frequent itemset as well as association rule modules. Set the support threshold to 1% and the confidence threshold to 90%. Observe the association rules containing the Cherry Tart item within the antecedent. What other item appears with it? When the confidence threshold is lowered to 45%, does the Cherry Tart item now appear without another item in the antecedent? Is the same consequent observed in both cases? How did lowering the confidence threshold lead to this change? Hint: Reference the Simpson's Paradox section of the text.

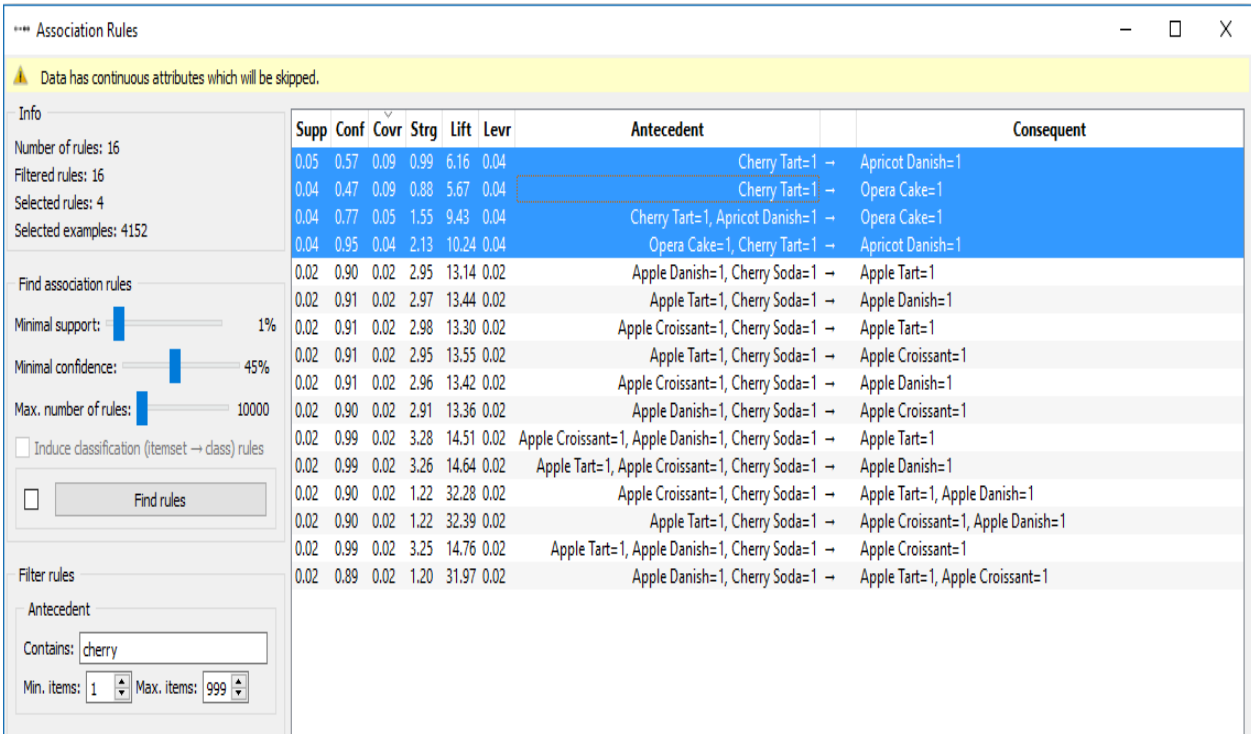
Solution :

CASE 1: The association rules containing the cherry tart item within the antecedent with the confidence threshold of 90% is,

Opera Cake, Cherry Tart(antecedent) ➔Apricot Danish(consequent)



CASE 2: When the confidence threshold has been lowered to 45%:



When the confidence threshold is lowered to 45% the association rules are:

Cherry Tart(consequent) ➔ Apricot Danish(antecedent)

Cherry Tart(consequent) ➔ Opera Cake(antecedent)

Opera Cake, Cherry Tart(consequent) ➔ Apricot Danish(antecedent)

Even after reducing the confidence threshold to 45%, the antecedent has Cherry tart with Opera Cake Opera Cake, Cherry Tart (consequent)➔ Apricot Danish(antecedent)

In both the cases antecedent is different, it varies i.e.

With 90% it is Apricot Danish

With 45% it is Apricot Danish, Opera Cake.

Confidence = frequency (X,Y)/frequency(X)

With respect to our example above, whenever cherry tart is brought at a high frequency in the entire association rule, then the confidence will be lower.

Ultimately, when the frequency of cherry tart is high then there are more products that are brought in case with 45% of threshold. (Apricot Danish, Opera cake was brought and people also brought a combination of Opera cake, Cherry tart and they then brought Apricot Danish).

Secondly, when the frequency of cherry tart was low then there are few products in case of 90% of threshold. (People who brought cherry tart, opera cake also brought Apricot Danish).

Therefore, Cherry Tart, Opera Cake and Apricot Danish are positively correlated in combined data but are negatively correlated in the stratified data. Hence, proper stratification is needed to avoid generating spurious patterns.

2.3 Problem 3

Load the Extended Bakery dataset (75000-out2-binary.csv) into Python using a Pandas dataframe. Calculate the binary correlation coefficient Φ for the Chocolate Coffee and Chocolate Cake items. Show whether the two items are symmetric binary variables via their co-presence and co-absence. Would an association rule between these items as antecedent and consequent have a high confidence level? Why or why not?

Solution :

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        from scipy.stats.stats import pearsonr
```

```
In [2]: df = pd.DataFrame(pd.read_table(r'E:\MS IIT\Data mining\assignment 2\75000-out2-binary.csv',sep = ','))
```

```
In [3]: df1=df['Chocolate Coffee']
        df11=df1.values
```

```
In [4]: df2=df['Chocolate Cake']
        df22=df2.values
```

```
In [5]: pearsonr(df11,df22)
```

```
Out[5]: (0.48556649252787693, 0.0)
```

In [6]:
```
pearsonr(df22,df11)
```

Out[6]: (0.48556649252787693, 0.0)

In [7]:
```
df5 = pd.concat([df1, df2], axis=1)
```

In [8]:
```
df5.corr()
```

Out[8]:

	Chocolate Coffee	Chocolate Cake
Chocolate Coffee	1.000000	0.485566
Chocolate Cake	0.485566	1.000000

In [9]:
```
newdf =(df1==0) & (df2==0)
gb = newdf.groupby(newdf)
Coab = gb.get_group(True)
CoAbsence= Coab.sum()
CoAbsence
```

Out[9]: 65802

In [10]:
```
newdf1 =(df1==1) & (df2==1)
gb = newdf1.groupby(newdf1)
Cop = gb.get_group(True)
CoPresence= Cop.sum()
CoPresence
```

Out[10]: 3303

In [11]:
```
newdf =(df2==0) & (df1==0)
gb = newdf.groupby(newdf)
Coab = gb.get_group(True)
CoAbsence= Coab.sum()
CoAbsence
```

Out[11]: 65802

In [12]:
```
newdf1 =(df2==1) & (df1==1)
gb = newdf1.groupby(newdf1)
Cop = gb.get_group(True)
CoPresence= Cop.sum()
CoPresence
```

Out[12]: 3303

The binary correlation coefficient for this dataset is 0.485566.

Co-presence is calculated for all the occurrence of C.

For chocolate coffee and chocolate cake:

1. Co-efficient correlation
   - Without swap: 0.485566
   - With swap: 0.485566
2. The Co-absence: 65802,
3. The Co-presence: 3303.

Items are symmetric binary variables as the correlation after the swapping is same.

The association rule between chocolate coffee and chocolate cake will have low confidence value, as the confidence increases the association rule does not fall under the threshold.