Homework Assignment - 2

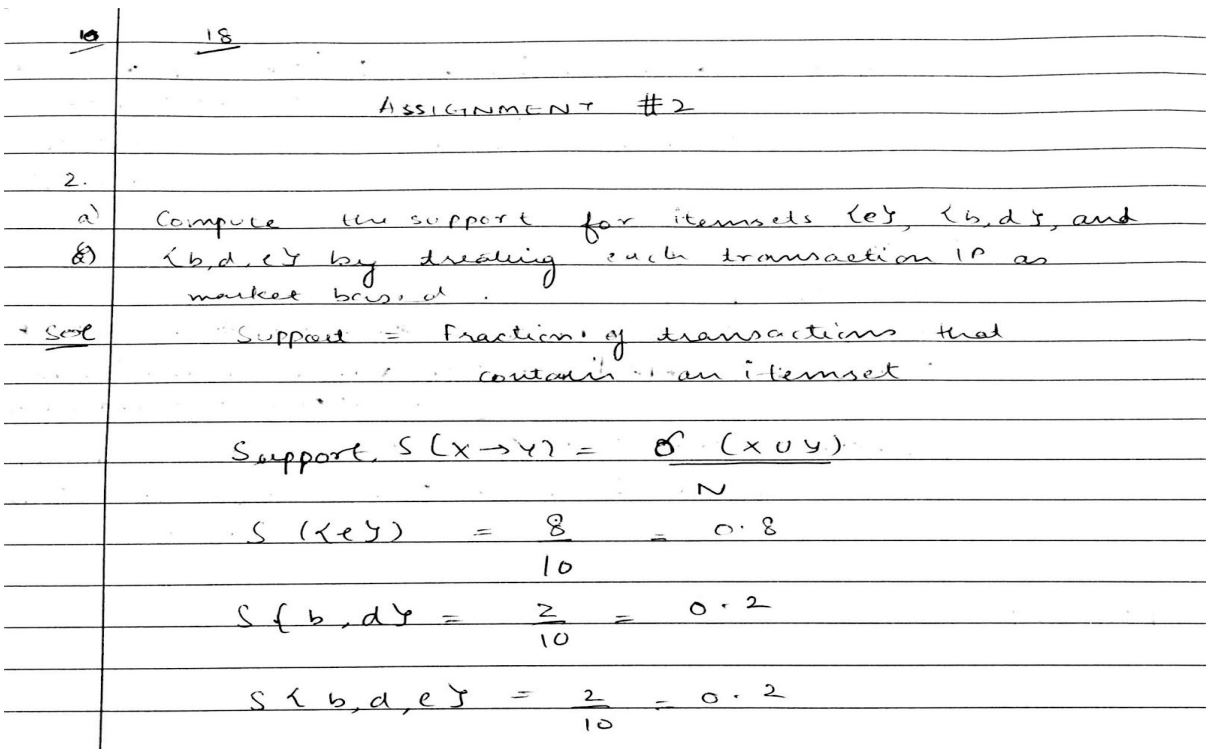
Textbook Problems.

2. 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)



b)

Confidence, $C(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

$C(bd \rightarrow e) = \frac{0.2}{0.2} = 100\%$

As $\sigma(bd) = 0.2$

$C(e \rightarrow bd) = \frac{0.2}{0.8} = 25\%$

Confidence is not symmetric, as the value of it changes.

c)

$S\{e\} = \frac{4}{5} = 0.8$

As in Customer ID – 4, there is no e

$S\{b,d\} = \frac{5}{5} = 1$

$S\{b,d,e\} = \frac{4}{5} = 0.8$

d)

$C(bd \rightarrow e) = \frac{0.8}{1} = 80\%$

$C(e \rightarrow bd) = \frac{0.8}{0.8} = 80\%$

Using this $C(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

e) Values of all the s1, s2, c1, c2 are totally different. There is no relation coming out when we observe the values. So there is no relation between them, which means they are not related to each other.

6. Consider the market basket transactions shown in Table 6.23 and answer the following questions based on it.

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,b)

a) Number of items = 6

Maximum number of rules extracted = $3^d - 2^{(d+1)} + 1$

$= 3^6 - 2^7 + 1$

$= 729 - 128 + 1 = 602$

602 rules can be extracted.

b) As Minsum > 0, transaction number 6 & 9 contains 4 items, the maximum size of the frequent itemset is 4.

c,d)

c) ${}^6C_3$ = 20 candidate 3-itemsets that can be formed.

d) {Bread, Butter} has the largest support

S{Bread, Butter} = 5/10 = 0.5

e)

e) For a pair to have same confidence, it should have same denominator value in both the cases which means same support.

$$C(A \rightarrow B) = \frac{\delta(A \cup B)}{\delta(A)}$$

$$C(B \rightarrow A) = \frac{\delta(A \cup B)}{\delta(B)}$$

To have same confidence the support count should be the same for both so to have same confidence. By analyzing & observing the table, the pairs which comes out are {Bread, Butter}, {Milk Bread}, {Milk, Butter}, {Beer, Cookies}.

7. Consider the following set of frequent 3-itemsets:
{1,2.3},{1,2,4},{1,2,5},{1,3,4},{1,3,5},{2,3,4},{2,3,5},{3,4,5}.

7.
a) {1,2,3}, {1,2,4}, {1,2,5}, {1,3,4}, {2,3,4}, {2,3,5}, {3,4,5}

{1,2,3} → {1,2,3,4}, {1,2,3,5}
{1,2,4} → {1,2,4,5}
{1,3,4} → {1,3,4,5}
{2,3,4} → {2,3,4,5}
Some itemsets cannot be extended as they were already present.

Candidate's after $F_{k-1} \times F_1$ merging are:
{1,2,3,4}, {1,2,3,5}, {1,2,4,5}, {1,3,4,5}, {2,3,4,5}

b) For Apriori Algorithm

{1,2,3} → {1,2,3,4}, {1,2,3,5}
{1,2,4} → {1,2,4,5}
{2,3,4} → {2,3,4,5}
{1,3,4} → {1,3,4,5}
We ignore other steps as they cannot be extended further

{1,2,3,4} will survive after pruning as 5 has least support count so it will be eliminated.

9.

9.

a) Candidates for transaction will be
{1, 3, 4}, {1, 3, 5}, {1, 3, 8}, {1, 4, 5}, {1, 4, 8}, {1, 5, 8}, {3, 4, 5},
{3, 4, 8}, {3, 5, 8}, {4, 5, 8}

given the transactions {1, 3, 4, 5, 8}

It will start from the root node and start iterating from the left. First it will go to the left path from where it will go to the second child. From there it will go to L5. After looking for the candidate it will move toward the right which is L1. After L1 it will go to the middle node L3. After looking at L3 it will go to the right and it will search for 3. So the overall visited path will be:

L1, L3, L5, L9, L11

) The itemset that are contained in the transaction {1, 3, 4, 5, 8} are {1, 4, 5}, {1, 5, 8}, {4, 5, 8}

11.

6

a) if the node is a maximal frequent itemset

For this we will find support of different itemset

TID	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}

- 1 Itemset

{a} ⟶ 5 {c} ⟶ 5

{b} ⟶ 7 {d} ⟶ 9

{e} ⟶ 6

- 2 Itemset

{a, b} → 3, {a, c} → 2, {a, d} → 4, {a, e} → 4

{b, c} → 3, {b, d} → 5, {b, e} → 4

{c, d} → 4, {c, e} → 2

{d, e} ⟶ 6,

3 Itemset

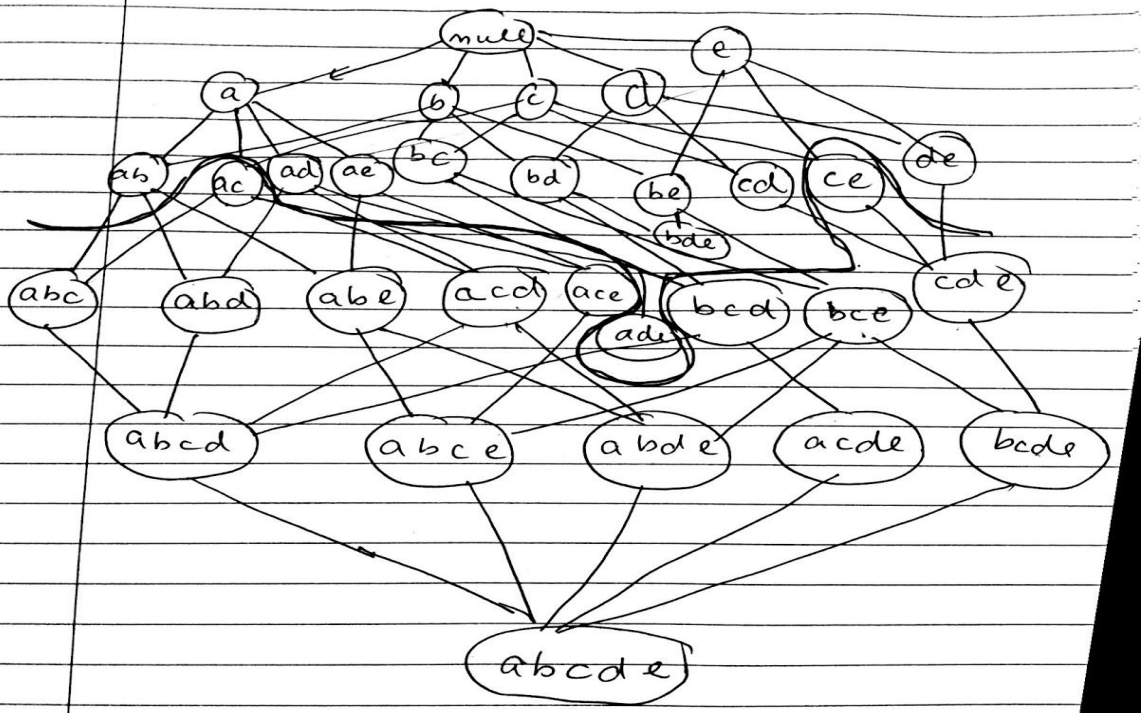
⟨a,b,c⟩ → 1 ⟨a,b,d⟩ → 2 ⟨a,b,e⟩ → 2 ⟨a,c,d⟩ → 1
⟨a,c,e⟩ → 1 ⟨a,d,e⟩ → 4
⟨b,c,d⟩ → 2 ⟨b,c,e⟩ → 1
⟨b,d,e⟩ → 3 ⟨b,d,e⟩ → 2

4 Itemset

⟨a,b,c,d⟩ → 0 , ⟨a,b,c,e⟩ → 0, ⟨a,b,d,e⟩ → 1
⟨a,c,d,e⟩ → 1 , ⟨b,c,d,e⟩ → 1

5 Itemset

⟨a,b,c,d,e⟩ → 0



All the line above the line are frequent item and all the ~~candidate~~ items below the black line in the above diagram will be infrequent item. It is identified by checking the support of the itemset. The itemset with more than 2 count is a frequent.

∴ M = {a, b}, {b, c}, {c, d}, {a, d, e}, {b, d, e}
{a}

Here these are the frequent sets whose all superset belongs to the infrequent item as per the defination.

Now Closed frequent are those candidates which are frequent and its support is not equal to the any of the immediate superset.

C = {a}, {b}, {c}, {d}, {a, b}, {b, c}, {b, d}, {c, d}, {d e}, {a d e}, {b d e}

Now Neither Maximal nor closed & a frequent

N = {e} {a, d} {b, e}, {a, e}

These are all the remaining frequent sets.

I = {a, c}, {a, b, c} {a, c, d} {c, e}. {a, b, d}, {a, c, e} {b, c, d} {b, c, e} {c, d, e} {a, b, c, d}, {a, b, c, e} {a, b, d, e} {a, c, d, e} {b, c, d, e} {a, b, c, d, e}

12.

12. a)

TID	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	{a, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

Table

First creating the contigency table for all the relations.

1) B → C

	C	C⁻
B	3	4
B⁻	2	1

2) A → D

	D	D⁻
A	4	1
A⁻	5	0

3) B → D

	D	D⁻
B	6	1
B⁻	3	0

4) E → C

	C	C⁻
E	2	4
E⁻	3	1

C → A

	A	A⁻
C	2	3
C⁻	3	2

Now support for each rule

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

Confidence

Rules	Confidence	Rank
b → c	3/7 = 0.42	4
a → d	4/5 = 0.8	2
b → d	6/7 = 0.87	1
e → c	2/6 = 0.33	5
c → a	2/5 = 0.4	3

Rules	Interest	Rank
$b \rightarrow c$	$((3/10)(5/10))/7/10 = .214$	3
$a \rightarrow d$	.72	2
$b \rightarrow d$	.771	1
$e \rightarrow c$	.161	5
$c \rightarrow a$	.2	4

IS $(A \rightarrow B) = P(X,Y)/\sqrt{P(X)P(Y)}$

Rules	IS	Rank
$b \rightarrow c$	$3/10/\sqrt{(5/10)(7/10)} = .567$	3
$a \rightarrow d$	.596	2
$b \rightarrow d$	.756	1
$e \rightarrow c$	.365	5
$c \rightarrow a$	.4	4

Rules	Klosgen	Rank
$b \rightarrow c$	-0.039	2
$a \rightarrow d$	-0.063	4
$b \rightarrow d$	-0.033	1
$e \rightarrow c$	-0.075	5
$c \rightarrow a$	-0.045	3

Rules	Odds Ratio	Rank
$b \rightarrow c$	0.375	2
$a \rightarrow d$	0	4
$b \rightarrow d$	0	4
$e \rightarrow c$	0.167	3
$c \rightarrow a$	0.444	1

18.

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

Compute the ϕ coefficient for A and B when C = 0, C = 1, and C = 0 or 1. Note that $\phi(A,B) = \frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$

C = 0, $f_{11} = 0$, $f_{1+} = 15$, $f_{+1} = 15$, $N = 60$

$P(A) = \frac{15}{60} = 0.25$, $P(B) = \frac{5}{20} = 0.25$

$P(A,B) = 0$

$\therefore \phi(A,B) = \frac{0 - 0.25 * 0.25}{\sqrt{(0.25 * 0.25)(1-0.25)(1-0.25)}}$

$= -\frac{1}{3}$

$= -0.33$

C = 1

$f_{11} = 5$, $f_{1+} = 5$, $f_{+1} = 5$, $N = 2$

$P(A) = 5/20 = 0.25$; $P(B) = 0.25$; $P(A,B) =$

$\phi(A,B) = \frac{0.25 - 0.25 * 0.25}{\sqrt{(0.25 * 0.25)(1-0.25)(1-0.25}}$

$= 1$

When C = 0 or 1

New tables will be

		A = 1	A = 0
B	1	5	15
	0	15	45

$f_{11} = 5$; $f_{1+} = 20$; $f_{+1} = 20$

$N = 80$

$P(A) = 20/80 = 0.25$ $P(B) = 20/80 = 0.25$

$P(A, B) = f_{11} / N = 5/80$

$\phi(A, B) = \underline{0}$

What conclusions can you draw from the above result?

- What we are observing here is: when C = 0 and C = 1, here we are getting coefficient correlation otherwise it will be 0 when C = 0 or 1 which means that by not taking some factors in account, we will not get any correlation.

19.

19

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

Table I

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

Table II

a) $N = 100$

$s(A) = \frac{10}{100}$

$\delta(B) = \frac{90}{100} = 0.9$

$S(A,B) = \frac{9}{100} = 0.09$

$$\text{Interest Measure } (I) = \frac{N f_{11}}{(f_{1+} f_{+1})}$$

$N = 100$, $f_{11} = 9$; $f_{1+} = 10$; $f_{+1} = 10$;
$f_{0+} = 90$; $f_{+0} = 90$

$$I = \frac{100 \times 9}{10 \times 10} = 9$$

$$\phi = \frac{N f_{11} - f_{1+} f_{+1}}{\sqrt{f_{1+} f_{+1} f_{0+} f_{+0}}}$$

$\therefore$ Correlation Coefficient $(\phi) = \frac{100 \times 9 - 10 \times 10}{\sqrt{10 \times 10 \times 90 \times 90}}$

$= 0.89$

Confidence $(\{A \rightarrow B\}) = \frac{\delta(A,B)}{\delta(A)} = 9/10 = 0.9$

Confidence $(\{B \rightarrow A\}) = \frac{\delta(A,B)}{\delta(B)} = \frac{9}{10} = 0.$

b) For table 7

$\delta(A) = 90/100 = 0.9$

$\delta(B) = 90/100 = 0.9$

Support of $(A, B) = 89/100$

where $N = 100$

Here $f_{11} = 89$, $f_{1+} = 90$; $f_{+1} = 90$;

$f_{0+} = 10$; $f_{+0} = 10$

$\therefore I = \frac{100 \times 89}{90 \times 90}$

$= 1.097$

$$\phi = \frac{N f_{11} - f_{1+} f_{+1}}{\sqrt{f_{+1} f_{1+} f_{0+} f_{+0}}}$$

$$= \frac{100 \times 89 - 90 \times 90}{\sqrt{100 \times 8100}} = 0.89$$

Confidence $[c(A \rightarrow B)] = \frac{\delta(A,B)}{\delta(A)} = 0.98$

Confidence $[c(B \rightarrow A)] = \frac{\delta(A,B)}{\delta(B)} = 0.98$

It means that it is invariant as it is mentioned in the book.

20.

20

Buy HDTV	Buy exercise Machine: Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

Table 6.19

Customer Group	Buy HDTV	Buy exercise Machine: Yes	No	Total
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

Table 6.20

c) Odd ratios = $\frac{f_{11}\ f_{00}}{f_{10}\ f_{01}}$

$AB = f_{11}$, $\bar{A}\bar{B} = f_{00}$, $\bar{A}B = f_{01}$, $A\bar{B} = f_{10}$

Here $AB = 99$, $\bar{A}\bar{B} = 66$, $A\bar{B} = 81$, $\bar{A}B = 54$

$\therefore$ odd ratio for 1st table $= \frac{99 \times 66}{54 \times 81}$

$= 1.4938$

Table 6.20

College students $AB = 1$, $\bar{A}\bar{B} = 30$, $A\bar{B} = 9$, $\bar{A}B = 4$

$\therefore$ odd ratio $= \frac{30 \times 1}{9 \times 4} = 0.8333$

Working adult $AB = 98$, $\bar{A}\bar{B} = 36$, $A\bar{B} = 72$,

$\therefore$ odd ratio $= \frac{98 \times 36}{72 \times 50} = 0.98$

20 b) ϕ coefficient for table 6.19

$$\phi = \frac{Nf_{11} - f_{1+} * f_{+1}}{\sqrt{f_{1+} * f_{+1} f_{0+} f_{+0}}}$$

Here Ncfar — $f_{11} = 99$, $f_{1+} = 180$, $N = 300$
$f_{+1} = 153$, $f_{+0} = 153$, $f_{+0} = 147$

& a

$$= \frac{300 \times 99 - 180 \quad 153}{\sqrt{180 \times 153 \wedge 120 \times 143}}$$

$$= \frac{29700 - 27540}{\sqrt{472586400}}$$

$$= 0.0993$$

ϕ Coefficient for table 6.20

$$\phi = \frac{Nf_{11} - f_{1+} \quad f_{+1}}{\sqrt{f_{1+} \; f_{+1} \; f_{0+} \; f_{+0}}}$$

College students

$N = 44$, $f_{11} = 1$, $f_{1+} = 10$, $f_{+1} = 5$, $f_{0+} = 34$,
$f_{+0} = 39$

$$\therefore \quad \phi = \frac{44 \times 1 - 10 \times 5}{\sqrt{10 \times 5 \times 34 \times 39}}$$

$$= -0.0233$$

working adults

$$\phi = \frac{Nf_{11} - f_{1+} f_{+1}}{\sqrt{f_{1+} \; f_{+1} \; f_{0+} f_{+0}}}$$

$N = 256$, $f_{11} = 98$, $f_{1+} = 170$, $f_{+1} = 148$
$f_{0+} = 86$, $f_{+0} = 108$

$$= -0.0047$$

Interest factor = $\frac{Nf_{11}}{f_{1r}\ f_{+1}}$

For Table 6.19

$I = \frac{300 \times 98}{153 \times 180} = 1.0784$

For table 6.20

$I = \frac{44 \times 1}{10 \times 5} = 0.88$

$I = \frac{256 \times 98}{170 \times 148} = 0.9971$

Practicum Problems

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?

Answer.

The market-basket sample dataset is loaded in the Orange application. After that some add ons are added to run the frequent itemset and association rule module which is shown in figure 1.1.

Figure 1.1 Frequent itemset and association rule modules are added.

Once the modules are added the support threshold is set to 10% which is shown in figure 1.2 to check the antecedent according to the rule given.

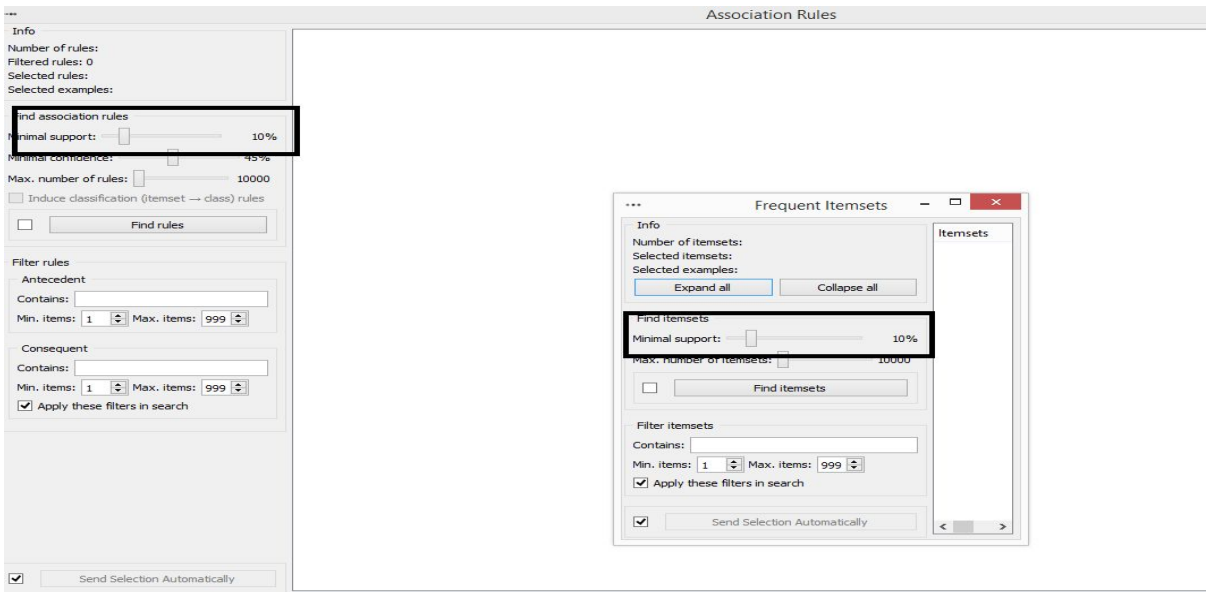


Figure 1.2 Support threshold is set to 10%.

Antecedent in the rule with highest lift value is observed which is shown in figure 1.3.

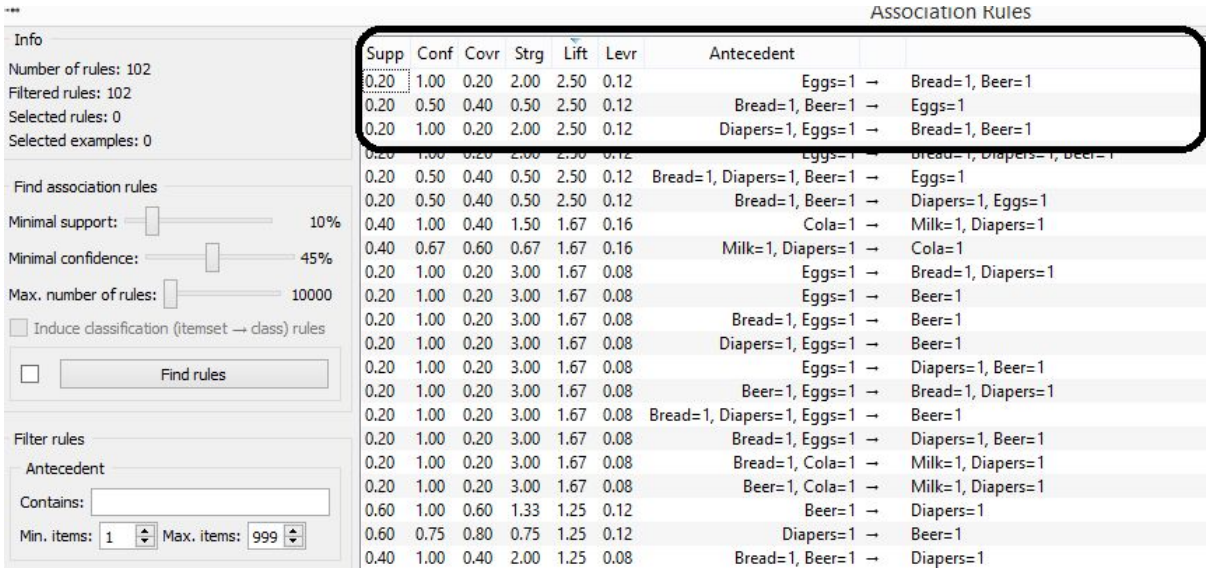


Figure 1.3 Antecedent with the highest lift value.

When lift is sorted with the highest value, what we observe in the Antecedent is that Eggs is coming two times, Bread is here for one time, and beer and diapers are also for a single time. These are the different antecedent which are associated with the highest lift value. The support value for all the antecedent is 0.20 and lift value is 2.50. It is a valuable association rule as the lift value is very high which is a major factor..

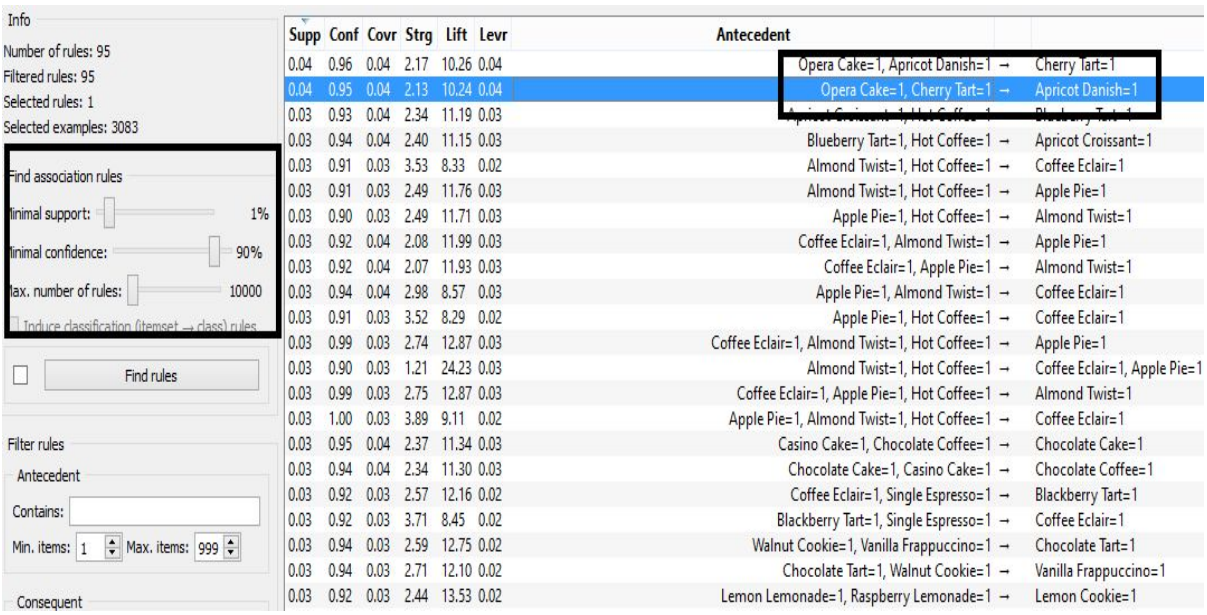
Problem 2.

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.

Answer.

First the extended bakery dataset (75000-out2-final.csv) is loaded into the orange application, and we run frequent itemset as well as association rule module in it. The value of support threshold is set to 1% and the confidence threshold is set to 90% which is shown in figure 2.1. Keeping this value we observe that only one association rule is generated where the antecedent with Cherry Tart is present which is shown in figure 2.1



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.04	0.96	0.04	2.17	10.26	0.04	Opera Cake=1, Apricot Danish=1	→	Cherry Tart=1
0.04	0.95	0.04	2.13	10.24	0.04	Opera Cake=1, Cherry Tart=1	→	Apricot Danish=1
0.03	0.93	0.04	2.34	11.19	0.03	[illegible]	→	[illegible]
0.03	0.94	0.04	2.40	11.15	0.03	Blueberry Tart=1, Hot Coffee=1	→	Apricot Croissant=1
0.03	0.91	0.03	3.53	8.33	0.02	Almond Twist=1, Hot Coffee=1	→	Coffee Eclair=1
0.03	0.91	0.03	2.49	11.76	0.03	Almond Twist=1, Hot Coffee=1	→	Apple Pie=1
0.03	0.90	0.03	2.49	11.71	0.03	Apple Pie=1, Hot Coffee=1	→	Almond Twist=1
0.03	0.92	0.04	2.08	11.99	0.03	Coffee Eclair=1, Almond Twist=1	→	Apple Pie=1
0.03	0.92	0.04	2.07	11.93	0.03	Coffee Eclair=1, Apple Pie=1	→	Almond Twist=1
0.03	0.94	0.04	2.98	8.57	0.03	Apple Pie=1, Almond Twist=1	→	Coffee Eclair=1
0.03	0.91	0.03	3.52	8.29	0.02	Apple Pie=1, Hot Coffee=1	→	Coffee Eclair=1
0.03	0.99	0.03	2.74	12.87	0.03	Coffee Eclair=1, Almond Twist=1, Hot Coffee=1	→	Apple Pie=1
0.03	0.90	0.03	1.21	24.23	0.03	Almond Twist=1, Hot Coffee=1	→	Coffee Eclair=1, Apple Pie=1
0.03	0.99	0.03	2.75	12.87	0.03	Coffee Eclair=1, Apple Pie=1, Hot Coffee=1	→	Almond Twist=1
0.03	1.00	0.03	3.89	9.11	0.02	Apple Pie=1, Almond Twist=1, Hot Coffee=1	→	Coffee Eclair=1
0.03	0.95	0.04	2.37	11.34	0.03	Casino Cake=1, Chocolate Coffee=1	→	Chocolate Cake=1
0.03	0.94	0.04	2.34	11.30	0.03	Chocolate Cake=1, Casino Cake=1	→	Chocolate Coffee=1
0.03	0.92	0.03	2.57	12.16	0.02	Coffee Eclair=1, Single Espresso=1	→	Blackberry Tart=1
0.03	0.92	0.03	3.71	8.45	0.02	Blackberry Tart=1, Single Espresso=1	→	Coffee Eclair=1
0.03	0.94	0.03	2.59	12.75	0.02	Walnut Cookie=1, Vanilla Frappuccino=1	→	Chocolate Tart=1
0.03	0.94	0.03	2.71	12.10	0.02	Chocolate Tart=1, Walnut Cookie=1	→	Vanilla Frappuccino=1
0.03	0.92	0.03	2.44	13.53	0.02	Lemon Lemonade=1, Raspberry Lemonade=1	→	Lemon Cookie=1

Figure 2.1 Minimal support set to 1%, minimal confidence set to 90%.

From the figure it is observed that other items which appear with Cherry tart at 90% minimal confidence is Opera Cake as antecedent. The consequent item is Apricot Danish. Also the support value is very less which is 0.04. Now the minimal confidence is reduced to 45%, it is observed that some more association rules are generated with the cherry tart as antecedent with the existing rule which was present at the minimal confidence as 90%. It is shown in figure 2.2. Now only one item is present in the antecedent which means only cherry tart is present as the antecedent. Also the consequent gets changed for one of the rule, the new consequent is the previous antecedent that is opera cake. By lowering the confidence leads to develope some more association rules which can be explained by the Simpson effect as it

is a paradox in probability and statistics, in which a trend appears in different groups of data but appears or disappears when the groups are combined that is confidence is decreased. Here the results are appeared in the reverse order. Here when the confidence decreases, new hidden association rules are discovered. This is explained by the Simpson's effect.

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.05	0.57	0.09	0.99	6.16	0.04	Cherry Tart=1	→	Apricot Danish=1
0.04	0.52	0.08	1.01	6.25	0.04	Blueberry Tart=1	→	Apricot Croissant=1
0.04	0.47	0.09	0.88	5.67	0.04	Cherry Tart=1	→	Opera Cake=1
0.04	0.77	0.05	1.55	9.43	0.04	Cherry Tart=1, Apricot Danish=1	→	Opera Cake=1
0.04	0.95	0.04	2.13	10.24	0.04	Opera Cake=1, Cherry Tart=1	→	Apricot Danish=1
0.04	0.48	0.08	1.44	4.39	0.03	Blackberry Tart=1	→	Coffee Eclair=1
0.04	0.49	0.08	1.10	5.81	0.03	Lemon Tart=1	→	Lemon Cake=1
0.04	0.49	0.07	1.05	6.30	0.03	Chocolate Tart=1	→	Vanilla Frappuccino=1
0.03	0.94	0.04	2.40	11.15	0.03	Blueberry Tart=1, Hot Coffee=1	→	Apricot Croissant=1
0.03	0.75	0.04	2.36	7.35	0.03	Blueberry Tart=1, Apricot Croissant=1	→	Hot Coffee=1
0.03	0.75	0.04	1.87	10.99	0.02	Coffee Eclair=1, Blackberry Tart=1	→	Single Espresso=1
0.03	0.92	0.03	3.71	8.45	0.02	Blackberry Tart=1, Single Espresso=1	→	Coffee Eclair=1
0.03	0.74	0.04	1.89	10.97	0.02	Chocolate Tart=1, Vanilla Frappuccino=1	→	Walnut Cookie=1
0.03	0.94	0.03	2.71	12.10	0.02	Chocolate Tart=1, Walnut Cookie=1	→	Vanilla Frappuccino=1
0.03	0.92	0.03	2.43	13.53	0.02	Apple Tart=1, Apple Croissant=1	→	Apple Danish=1
0.03	0.92	0.03	2.42	13.61	0.02	Apple Tart=1, Apple Danish=1	→	Apple Croissant=1
0.02	0.74	0.03	2.23	12.01	0.02	Apple Tart=1, Apple Danish=1	→	Cherry Soda=1

Figure 2.2 Changing the minimal confidence to 45%.

Problem 3

Load the Extended bakery dataset (75000-out2-binary.csv) into python using a pandas dataframe. Calculate the binary correlation coefficient 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.

The extended bakery dataset (75000-out-binary.csv) is loaded into the python using the pandas dataframes. The binary correlation coefficient for the chocolate coffee and chocolate cake items were calculated using the corr in the pandas dataframe of python.It calculates the co-presence and co-absence of the coefficients. We will compute the values of correlation and what we observed is that the value for {Chocolate Coffee, Chocolate Cake }, and {Chocolate Cake, Chocolate Coffee } is same which is shown in figure 3.1 and figure 3.2.

```
In [2]: import pandas as pd
        import numpy as np

        dataset = pd.read_csv('C:/Users/Ashu-Palia/Downloads/75000-out2-binary.csv')
        dataset.corr() ["Chocolate Coffee"]

Out[2]: Transaction Number     0.003000
        Chocolate Cake         0.485566
        Lemon Cake            -0.032172
        Casino Cake            0.398779
        Opera Cake            -0.043968
        Strawberry Cake       -0.034929
        Truffle Cake          -0.035656
        Chocolate Eclair      -0.006291
        Coffee Eclair         -0.058087
        Vanilla Eclair        -0.010328
        Napoleon Cake         -0.030326
```

Figure 3.1 Correlation value of chocolate cake.

```
In [3]: import pandas as pd
        import numpy as np

        dataset = pd.read_csv('C:/Users/Ashu-Palia/Downloads/75000-out2-binary.csv')
        dataset.corr() ["Chocolate Cake"]

Out[3]: Transaction Number    -0.002249
        Chocolate Cake         1.000000
        Lemon Cake            -0.030612
        Casino Cake            0.401565
        Opera Cake            -0.042011
        Strawberry Cake       -0.037961
        Truffle Cake          -0.032316
        Chocolate Eclair      -0.010160
        Coffee Eclair         -0.059381
        Vanilla Eclair        -0.009168
        Napoleon Cake         -0.034355
        Almond Tart           -0.006815
```

Figure 3.2 Correlation value of Chocolate Coffee.

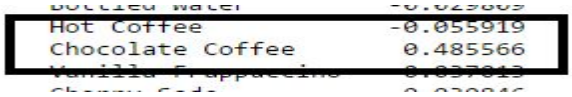


Figure 3.2.

Now we can see that as the value is same for **Chocolate Coffee** and **Chocolate Cake** which is 0.485566, we can say that the two items that is Chocolate Coffee and Chocolate Cake are symmetric binary variables due to their co-presence.

The confidence by the calculation comes out to be 0.52, by plotting the contingency matrix. It shows that the relation does not have a high confidence between the antecedents and consequents.

```
In [8]: import pandas as pd
        import numpy as np

        dataset = pd.read_csv('C:/Users/Ashu-Palia/Downloads/75000-out2-binary.csv')
        dataset.corr() ["Chocolate Cake"]
        pd.crosstab(dataset["Chocolate Coffee"]>0, dataset["Chocolate Cake"]>0)
```

Out[8]:

Chocolate Cake	False	True
Chocolate Coffee		
False	65802	2962
True	2933	3303

Figure3.3 Contingency Matrix.