CS 422

Data Mining

Homework Assignment 2

Ronit Rudra
A20379221
rrudra@hawk.iit.edu

Course Instructor
Prof. Jawahar Panchal

1 Textbook Questions

1.1 Chapter 6

1.1.1 Question 2

We are given the table:

Table 1: Market Basket Transactions

Customer ID	Transaction ID	Item 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) The support count is the number of instances of a particular itemset in all the transactions.

$$\sigma(e) = \frac{8}{10} = 0.8$$
$$\sigma(b,d) = \frac{2}{10} = 0.2$$
$$\sigma(b,d,e) = \frac{2}{10} = 0.2$$

(b) Confidence is the frequency of items Y appearing in transactions containing X.

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{1}$$

$$c(b,d \to e) = \frac{\sigma(b,d,e)}{\sigma(b,d)} = \frac{0.2}{\frac{2}{10}} = 100\%$$
$$c(e \to b,d) = \frac{\sigma(b,d,e)}{\sigma(e)} = \frac{\frac{2}{10}}{\frac{8}{10}} = 25\%$$

Confidence is not a symmetric measure.

(c) There are 5 unique customer IDs. Each item to be included if it appears at least once in the market basket.

$$\sigma(e) = \frac{4}{5} = 0.8$$
$$\sigma(b,d) = \frac{5}{5} = 1$$
$$\sigma(b,d,e) = \frac{4}{5} = 0.8$$

Taking customer ID as the market basket yields quite different results as multiple transactions are coupled into one transaction.

(d) Recalculating confidence of association:

$$c(b, d \rightarrow e) = \frac{\sigma(b, d, e)}{\sigma(b, d)} = \frac{0.8}{1} = 80\%$$
$$c(e \rightarrow b, d) = \frac{\sigma(b, d, e)}{\sigma(e)} = \frac{0.8}{0.8} = 100\%$$

(e) Considering the above calculations, there appears to be no relationship among the support and confidence values of two different attributes.

1.2 Question 6:

We are given the table:

Table 2: Market Basket Transactions

Transaction ID	**Item 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) The itemset is $I = Beer, Bread, Butter, Cookies, Diapers, Milk$ so there are 6 items. The total number of rules that can be extracted from an itemset containing d items is:

$$R = 3^d - 2^{d+1} + 1 \tag{2}$$

Therefore, from the above equation:

$$R = 3^6 - 2^{6+1} + 1 = 729 - 128 + 1 = 602$$

(b) Minimum support is zero. This has no effect on the maximum size of frequent itemsets. Thus the maximum size is 4 as the data has a the largest transaction containing 4 items.

(c) This is just a combination of itemsets i.e choosing 3 items from a set of 6 items.

$$Total = \binom{6}{3} = \frac{6!}{3! \cdot 3!} = 20$$

(d)

Table 3: Itemset Support

Itemset	**Support**
{milk,beer}	1
{milk,diaper}	4
{milk,bread}	3
{milk,butter}	2
{milk,cookies}	1
{beer,diapers}	3
{beer,bread}	0
{beer,butter}	0
{beer,cookies}	2
{diapers,bread}	2
{diapers,butter}	3
{diapers,cookies}	1
{bread,butter}	5
{bread,cookies}	1
{butter,cookies}	1

The Frequent-2 itemsets would have the maximum support hence Frequent-3 and Frequent-4 were not calculated. The itemset {Bread,Butter} has the highest support of 5.

(e) Finding support of each:

$$\sigma(beer) = 0.4$$
$$\sigma(bread) = 0.5$$
$$\sigma(butter) = 0.5$$
$$\sigma(cookies) = 0.4$$
$$\sigma(diapers) = 0.7$$
$$\sigma(milk) = 0.5$$

The items which have the same individual support would have the same confidence as $C(a \rightarrow b) = \frac{a \cap b}{a}$ and $C(b \rightarrow a) = \frac{a \cap b}{b}$

Thus: {Beer, Cookies}, {Bread, Butter}, {Bread, Milk} and {Butter, Milk} have the same confidence for $a \rightarrow b$ and $b \rightarrow a$

1.3 Question 7

We have the following frequent-3 itemset:
{1,2,3},{1,2,4},{1,2,5},{1,3,4},{1,3,5},{2,3,4},{2,3,5},{3,4,5}
There are only 5 items I={1,2,3,4,5}

(a) For the $F_{K-1} \times F_1$ merging strategy, in order to generate a frequent-4 itemset we arrange the frequent-3 itemsets in lexographic order and append one item. Generating, we get: {1,2,3,4}, {1,2,3,5}, {1,2,4,5}, {1,3,4,5}, {2,3,4,5}

(b) Using the Apriori Algorithm, the Frequent-4 Itemsets generated would be the same as in (a) i.e:

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

(c) According to the Apriori Principle, the superset is infrequent if any of its subsets is infrequent 'or' the support of a superset cannot exceed its subset's support. Thus, except {1,2,3,4} all other sets have at least one infrequent subset. Hence, only {1,2,3,4} survives the pruning stage.

1.4 Question 9:

(a) To find the candidates of the transaction {1,3,4,5,8}, all of the leaves of the hash tree containing the subsets of the transaction itemset shall be visited at least once. Since this is a frequent-3 hash tree, the hash function would be $h(p) = p \ mod \ 3$. Hence, this would mean $L1$, $L3$, $L5$, $L9$ and $L11$ will be visited.

(b) For transaction with {1,3,4,5,8} the hash tree candidates would be: {1,4,5}, {1,5,8}, {4,5,8}

1.5 Question 11:

A node is a Maximally Frequent itemset if its immediate supersets are infrequent.

A Closed Itemset is an itemset X whose immediate supersets do not have the same support count as X.

The Support Threshold is 30% or 0.3

The Support Values of the itemsets are given below along with their associated labels according to the question:

Table 4: Support Values

Itemset	**Support**	**Label**
a	0.5	C
b	0.7	C
c	0.5	C
d	0.9	C
e	0.6	N
ab	0.3	M,C
ac	0.2	I
ad	0.4	N
ae	0.4	N
bc	0.3	M
bd	0.6	C
be	0.4	N
cd	0.4	M,C
ce	0.2	I
de	0.6	C
abc	0.1	I
abd	0.2	I
abe	0.2	I
acd	0.1	I
ace	0.1	I
ade	0.4	M,C
bcd	0.2	I
bce	0.1	I
bde	0.4	M,C
cde	0.2	I
abcd	0	I
abce	0	I
abde	0.2	I
acde	0.1	I
bcde	0.1	I
abcde	0	I

1.6 Question 12:

The transaction table is as follows:

Table 5: Market Basket Transactions

Transaction ID	**Item 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) Contingency tables relate which items are present in a particular transaction.

Table 6: $b \rightarrow c$

	c	$\overline{c}$	
b	3	4	7
$\overline{b}$	2	1	3
	5	5	10

Table 7: $a \rightarrow d$

	d	$\overline{d}$	
a	4	1	5
$\overline{a}$	5	0	5
	9	1	10

Table 8: $b \rightarrow d$

	d	$\overline{d}$	
b	6	1	7
$\overline{b}$	3	0	3
	9	1	10

Table 9: $e \rightarrow c$

	c	$\overline{c}$	
e	2	4	6
$\overline{e}$	3	1	4
	5	5	10

Table 10: $c \to a$

	a	$\overline{a}$	
c	2	3	5
$\overline{c}$	3	2	5
	5	5	10

(b) For a 2-way contingency table of variables A and B:

Table 11: Generic Contingency Table

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Except Support (s) and Confidence (c), We have various other Objective measures as defined below. Replacing $P(X) = \frac{f_{1+}}{N}$, $P(Y) = \frac{f_{+1}}{N}$ and $P(X,Y) = \frac{f_{11}}{N}$ in the equations given in the question (**Note: Expression of Interest is incorrect in the question but answers have been calculated according to the expression given**) :

Table 12: Objective Measures

Measure (Symbol)	**Definition**
Odds ratio (α)	$(f_{11}f_{00})(f_{01}f_{10})$
Interest (I)	$\frac{f_{11}f_{+1}}{Nf_{1+}}$
Cosine (IS)	$\frac{f_{11}}{\sqrt{f_{1+}f_{+1}}}$
Klosgen	$\sqrt{\frac{f_{11}}{N}} \times \frac{Nf_{11}-f_{1+}f_{+1}}{Nf_{1+}}$

Calculating measures for contingency matrices in part (a) using the above, we get:

Table 13: Rule Measures

Rule	**Support**	**Confidence**	**Interest**	**Cosine**	**Klosgen**	**Odds Ratio**
$b \to c$	$\frac{3}{10}$	$\frac{3}{7}$	0.214	0.507	-0.039	0.375
$a \to d$	$\frac{4}{10}$	$\frac{3}{5}$	0.720	0.596	-0.063	0
$b \to d$	$\frac{6}{10}$	$\frac{6}{7}$	0.771	0.756	-0.033	0
$e \to c$	$\frac{2}{10}$	$\frac{2}{6}$	0.167	0.365	-0.075	0.167
$c \to a$	$\frac{2}{10}$	$\frac{2}{5}$	0.200	0.400	-0.045	0.444
Rank	{3,2,1,4,4}	{3,2,1,5,4}	{3,2,1,5,4}	{3,2,1,5,4}	{2,4,1,5,3}	{2,4,4,3,1}

1.7 Question 18:

(a) From the contingency table we have: for C=0:
$f_{11} = 0; f_{00} = 30; f_{1+} = 15; f_{0+} = 45; f_{+1} = 15; f_{+0} = 45$ for C=1:
$f_{11} = 5; f_{00} = 15; f_{1+} = 5; f_{0+} = 15; f_{+1} = 5; f_{+0} = 15$ for C=1 or 0:
$f_{11} = 5; f_{00} = 45; f_{1+} = 20; f_{0+} = 60; f_{+1} = 20; f_{+0} = 60$ The Correlation is:

$$C_0 = \frac{60 \times 0 - 15 \cdot 15}{\sqrt{15 \cdot 15 \cdot 45 \cdot 45}} = \frac{-1}{3}$$

$$C_1 = \frac{20 \times 5 - 5 \cdot 5}{\sqrt{5 \cdot 5 \cdot 15 \cdot 15}} = 1$$
$$C_{1,0} = \frac{80 \times 5 - 20 \cdot 20}{\sqrt{20 \cdot 20 \cdot 60 \cdot 60}} = 0$$

(b) When the factor C is not taken into account, the correlation is low and when C is factored in, A and B show a high positive correlation. Thus, ignoring C resulted in loss of interesting relationships as both A and B could have been indirectly related through C. This C is known as the confounding variable as it skews the relationship between A and B as technically we are including A's and B's relationship with C when relating A to B or vice versa. The Reversal or removal of associations between two variables due to a confounding variable is known as Simpson's Paradox

1.8 Question 19:

(a) The Interest Measures for binary relationships can be calculated using expressions from the previous table:

$$s(A) = 0.1 \quad s(B) = 0.9 \qquad s(A \cap B) = 0.09$$
$$I(A \cap B) = 9 \quad \phi(A \cap B) = 0.89$$
$$c(A \to B) = 0.9 \quad c(B \to A) = 0.9$$

(b)

$$s(A) = 0.9 \quad s(B) = 0.9 \qquad s(A \cap B) = 0.89$$
$$I(A \cap B) = 1.09 \quad \phi(A \cap B) = 0.89$$
$$c(A \to B) = 0.98 \quad c(B \to A) = 0.98$$

(c) Correlation and interest are symmetric measures while confidence and support are asymmetric measures. The tables are inversions of each other and under the inversion operation except correlation all these measures are variant. Hence, correlation remains unchanged as it takes into account both absences and presences.

1.9 Question 20:

For Table 6.19:
$N = 300; f_{11} = 99; f_{00} = 66; f_{1+} = 180; f_{0+} = 120; f_{+1} = 153; f_{+0} = 147$ For Table 6.20:
College Student:
$N = 44; f_{11} = 1; f_{00} = 30; f_{1+} = 10; f_{0+} = 34; f_{+1} = 5; f_{+0} = 39$
Working Adult:
$N = 256; f_{11} = 98; f_{00} = 36; f_{1+} = 170; f_{0+} = 86; f_{+1} = 148; f_{+0} = 108$

(a) For Table 6.19, odds ratio is:

$$\alpha = \frac{66 \times 99}{81 \times 54} = 1.4938$$

For Table 6.20, the odd ratios are:

$$\alpha_1 = \frac{1 \times 30}{4 \times 9} = 0.833$$

$$\alpha_2 = \frac{98 \times 36}{50 \times 72} = 0.98$$

(b) For Table 6.19, the correlation is:

$$\phi = \frac{300 \times 99 - 180 \cdot 153}{\sqrt{180 \cdot 153 \cdot 120 \cdot 147}} = 0.0979$$

For Table 6.20, the correlations are:

$$\phi_1 = \frac{44 \times 1 - 10 \cdot 5}{\sqrt{10 \cdot 5 \cdot 39 \cdot 34}} = -0.233$$
$$\phi_2 = \frac{256 \times 99 - 170 \cdot 148}{\sqrt{170 \cdot 148 \cdot 86 \cdot 108}} = -0.0047$$

(c) For Table 6.19, the Interest is:

$$I = \frac{300 \times 99}{180 \cdot 153} = 1.078$$

For Table 6.20, the Interests are:

$$\phi_1 = \frac{44 \times 1}{10 \cdot 5} = 0.88$$
$$\phi_2 = \frac{256 \times 99}{170 \cdot 148} = 0.9971$$

When the data is pooled together, the correlation direction changes. Thus, Customer Group is a confounding variable.