CS 422 – Data Mining

Homework 2

22 February 2017

Mayank Bansal

A20392482

Recitation Problems

Problem 2

a) **Support({e})** = 8/10 = 0.8
Support({b,d}) = 2/10 = 0.2
Support({b,d,e}) = 2/10 = 0.2

b) **Confidence ({b,d}→{e})** $= \frac{Support(\{b,d,e\})}{Support(\{b,d\})}$
= 0.2/0.2 = 1
Confidence ({e}→{b,d}) $= \frac{Support(\{b,d,e\})}{Support(\{e\})}$
= 0.2/0.8 = 0.25
Confidence is not a symmetric measure.

c) **Support({e})** = 4/5 = 0.8
Support({b,d}) = 5/5 = 1
Support({b,d,e}) = 4/5 = 0.8

d) **Confidence ({b,d}→{e})** $= \frac{Support(\{b,d,e\})}{Support(\{b,d\})}$
= 0.8/1 = 0.8
Confidence ({e}→{b,d}) $= \frac{Support(\{b,d,e\})}{Support(\{e\})}$
= 0.8/0.8 = 1

e) No. s_1 and s_2 or c_1 and c_2 have no relations.

Problem 6

a) Number of distinct items = d = 6
Therefore, the number of association rules = $3^d-2^{d+1}+1 = 3^6-2^{(6+1)}+1 = 602$

b) The longest transaction contains 4 items so the largest frequent item set can be of size 4.

c) The maximum size 3-itemsets that can be derived is C(6,3) = 6!/3!*3!= =20

d) The itemset of size two or larger that has the highest support is {Bread,Butter} whose support is 5/10 = 0.5

e) The pair of items such that {a}→{b} and {b}→{a} have the same confidence are (Beer, Cookies) and (Bread, Butter)

Problem 7

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

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

c) {1,2,3,4}, {1,2,3,5} (proper subsets are frequent)

Problem 9

a) L1, L3, L5, L9, L11

b) {145}, {158}, {458}

Problem 11

Null – Frequent, Closed
A – Frequent, Closed
B – Frequent, Closed
C – Frequent, Closed
D – Frequent, Closed
E – Frequent
AB – Frequent, Maximal
AC – Infrequent
AD - Frequent
AE – Frequent
BC – Frequent, Maximal
BD – Frequent, Closed
BE – Frequent
CD – Frequent, Maximal
CE - Infrequent
DE – Frequent, Closed
ABC - Infrequent
ABD - Infrequent
ABE - Infrequent
ACD - Infrequent
ACE - Infrequent
ADE – Frequent, Maximal
BCD - Infrequent
BCE - Infrequent

BDE – Frequent, Maximal
CDE - Infrequent
ABCD - Infrequent
ABCE - Infrequent
ABDE - Infrequent
ACDE - Infrequent
BCDE - Infrequent
ABCDE – Infrequent

Problem 12

a)

	c	**c’**
b	3	4
b’	2	1

	d	**d’**
a	4	1
a’	5	0

	a	**a’**
c	2	3
c’	3	2

	d	**d’**
b	6	1
b’	3	0

	c	**c’**
e	2	4
e’	3	1

b) i) Support

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

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

iii) Interest(X→Y) = $\frac{P(X,Y)*P(Y)}{P(X)}$

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) = $\frac{P(X,Y)}{\sqrt{(P(X)*P(Y))}}$

Rules	**IS**	**Rank**
b→c	.507	3
a→d	.596	2
b→d	.756	1
e→c	.365	5
c→a	.4	4

v) Klosgen(X→Y) = $\sqrt{P(X,Y)} * \ (P(Y|X) - P(Y))$

where $P(Y|X) = \frac{P(X,Y)}{P(X)}$

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) = $\frac{[P(X,Y)* P(X',Y')]}{[P(X',Y)*P(X,Y')]}$

Rules	**Odds Ratio**	**Rank**
b→c	0.0375	2
a→d	0	4
b→d	0	4
e→c	0.167	3

c→a	0.444	1

Problem 18

a) i) C=0, then **φ**(A,B) $= \frac{f11*f00-f10*f01}{\sqrt{(f11+f10)^2*(f00+f01)^2}}$

$= \frac{0*30-15*15}{\sqrt{(0+15)^2*(30+15)^2}}$

$= \frac{-225}{\sqrt{225*2025}}$

= -1/3

ii) C=1, then **φ**(A,B) $= \frac{f11*f00-f10*f01}{\sqrt{((f11+f10)^2*(f00+f01)^2)}}$

$= \frac{5*15-0*0}{\sqrt{(5+0)^2*(15+0)^2}}$

$= \frac{75}{\sqrt{25*225}}$

= 1

iii) C=0 or 1, then **φ**(A,B) $= \frac{f11*f00-f10*f01}{\sqrt{((f11+f10)^2*(f00+f01)^2)}}$

$= \frac{5*45-15*15}{\sqrt{((5+15)^2*(45+15)^2)}}$

$= \frac{0}{\sqrt{(400*3600)}}$

= 0

b) This shows that some relationships are lost when the controlling variable is not accounted for. When C=0 or 1, we lose all correlation data in this case.

Problem 19

a) **Support(A)** = 10/100 = 0.1

Support(B) = 10/100 = 0.1

Support(A,B) = 9/100 = 0.09

Interest(A,B) $= \frac{\left(\frac{9}{100}\right)*\left(\frac{1}{10}\right)}{\frac{1}{10}}$ = 0.09

Φ(A,B) $= \frac{9*89-1*1}{\sqrt{10*10*90*90}}$ = 0.888888

Conf(A→B) = 0.89/0.9= 0.9888

Conf(B→A) = 0.89/0.9 = 0.98888

b) Φ is the only value which is in-variant when the data is transposed. The support, interest and confidences of the association rules all vary. φ uses both presences and absences of an item in a transaction.

Problem 20

a) Table 6.19

Odds Ratio (X→Y)

$$= \frac{[P(X,Y) * P(X',Y')]}{[P(X',Y) * P(X,Y')]}$$

$$= \frac{\left[\left(\frac{99}{300}\right) * \left(\frac{66}{300}\right)\right]}{\left[\left(\frac{54}{300}\right) * \left(\frac{81}{300}\right)\right]}$$

= 1.4938

Table 6.20

College Students

Odds Ratio (X→Y)

$$= \frac{[P(X,Y) * P(X',Y')]}{[P(X',Y) * P(X,Y')]}$$

$$= \frac{\left[\left(\frac{1}{44}\right) * \left(\frac{30}{44}\right)\right]}{\left[\left(\frac{9}{44}\right) * \left(\frac{4}{44}\right)\right]}$$

= 0.83333

College Students

Working Adult (X→Y)

$$= \frac{[P(X,Y) * P(X',Y')]}{[P(X',Y) * P(X,Y')]}$$

$$= \frac{\left[\left(\frac{98}{256}\right) * \left(\frac{36}{256}\right)\right]}{\left[\left(\frac{72}{256}\right) * \left(\frac{50}{256}\right)\right]}$$

= 0.98

b) Table 6.19

Φ(X,Y)

$$= \frac{f11 * f00 - f10 * f01}{\sqrt{(f11+f10)^2 * (f00+f01)^2}}$$

$$= \frac{99*66\text{-}54*81}{\sqrt{(99+81)^2 * (66+54)^2}}$$

$$= \frac{2160}{\sqrt{21600}}$$

= .1

Table 6.20

College Student

Φ(X,Y)

$$= \frac{f11 * f00 - f10 * f01}{\sqrt{(f11+f10)^2 * (f00+f01)^2}}$$

$$= \frac{1*30\text{-}9*4}{\sqrt{(1+9)^2 * (4+30)^2}}$$

$$= \frac{-6}{\sqrt{1256}}$$

= -0.00477

Working Adult

Φ(X,Y)

$$= \frac{f11*f00-f10*f01}{\sqrt{(f11+f10)^2*(f00+f01)^2}}$$

$$= \frac{98*36\text{-}72*50}{\sqrt{(98+72)^2*(50+36)^2}}$$

$$= \frac{-72}{\sqrt{14620}}$$

= -0.00492

c) Table 6.19

Interest(X→Y)

$$= \frac{P(X,Y)*P(Y)}{P(X)}$$

$$= \frac{\left(\frac{99}{300}\right)*\left(\frac{153}{300}\right)}{\frac{180}{300}}$$

= 0.2805

Table 6.20

College Student

Interest(X→Y)

$$= \frac{P(X,Y)*P(Y)}{P(X)}$$

$$= \frac{\left(\frac{1}{44}\right)*\left(\frac{5}{44}\right)}{\frac{10}{44}}$$

= 0.0113

Working Adult

Interest(X→Y)

$$= \frac{P(X,Y)*P(Y)}{P(X)}$$

$$= \frac{\left(\frac{98}{256}\right)*\left(\frac{148}{256}\right)}{\frac{170}{256}}$$

= 0.3332

Practicum Problems

Problem 2.1

Eggs has the highest values for lift. (Support = 0.2)

Since the confidence is significantly high for even one transaction, the association rule is useful.

Problem 2.2

Items with the highest support value:

1. Coffee Éclair - 10.92%
2. Hot Coffee - 10.27%
3. Tuile Cookie - 10.07%

Tuile Cookie does not appear in the association rules when the confidence is 95% and support threshold is 1%. At a confidence threshold of 50% it is found.

Since **(Tuile Cookie→Marzipan Cookie)** has low support, transaction frequency for this association is low.

Tulie cookies occurs with Marzipan Cookies 3819 times. The confidence of the association rule is low because 3819 transactions occurred for people buying both items out of 7556 total transactions where people bought only a Tulie cookie. Given that, the association rule isn't very useful.

Problem 2.3

The variables are binary symmetric variables since the Phi values are almost the same.

The association rule will have high confidence as a high Chi-square value indicates the items are less independent.