Supplementary Notes #1

Data Mining and Data Warehousing

Exercise on Association Rule Mining

You are given the data set below:

Trans ID	Items Purchased
100	Bread, butter, noodles
200	Butter, chips
300	Butter, coke
400	Bread, butter, chips
500	Bread, coke
600	Butter, coke
700	Bread, coke
800	Bread, butter, coke, noodles
900	Bread, butter, coke

1. Mine the above data set for interesting associations:

   a. Use min_sup = 20% and min_conf = 50%.

1.	Bread → Butter	(44.4%, 66.7%)
2.	Bread → Coke	(44.4%, 66.7%)
3.	Butter → Bread	(44.4%, 57.1%)
4.	Butter → Coke	(44.4%, 57.1%)
5.	Noodle → Bread	(22.2%, 100%)
6.	Coke → Bread	(44.4%, 66.7%)
7.	Noodle → Butter	(22.2%, 100%)
8.	Chips → Butter	(22.2%, 100%)
9.	Coke → Butter	(44.4%, 66.7%)
10.	Noodle → Bread, Butter	(22.2%, 100%)
11.	Butter, Noodle → Bread	(22.2%, 100%)
12.	Butter, Coke → Bread	(22.2%, 50%)
13.	Bread, Noodle → Butter	(22.2%, 100%)
14.	Bread, Coke → Butter	(22.2%, 50%)
15.	Bread, Butter → Noodle	(22.2%, 50%)
16.	Bread, Butter → Coke	(22.2%, 50%)

b. How many times do you have to go through the data set to discover all interesting association?

Given *n* items, one can discover 1-itemsets, 2-itemsets, …, *n*-itemsets. The number of itemsets, *N*, is therefore given by:

$$N = \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{n}.$$

For each *k*-itemset, one can form association rules with consequents composed of 1-itemsets, 2-itemsets, …, (*k* – 1)-itemsets. The number of association rules, M_k, is therefore given by:

$$M_k = \binom{k}{1} + \binom{k}{2} + \ldots + \binom{k}{k-1}.$$

Consequently, the total number of association rules is calculated by:

$$\binom{n}{1} M_1 + \binom{n}{2} M_2 + \ldots + \binom{n}{n} M_n.$$

Hence, for 5 items, we need to consider: 5 + 20 + 60 + 70 + 30 = 185.

c. Repeat a using Apriori Algorithm.

L1

Bread	A	6
Butter	B	7
Noodle	C	2
Chips	D	2
Coke	E	6

C2

AB	4
AC	2
AD	1
AE	4
BC	2
BD	2
BE	4
CD	0
CE	1
DE	1

L2

AB
AC
AE
BC
BD
BE

C3

ABC	2
ABD	1
ABE	2

L3

ABC
ABE

X	→	Y	P(X)		P(Y)		P(X,Y)		CONFIDENCE
A	→	B	6	66.7%	7	77.8%	4	44.4%	66.7%
A	→	C	6	66.7%	2	22.2%	2	22.2%	33.3%
A	→	E	6	66.7%	6	66.7%	4	44.4%	66.7%
B	→	C	7	77.8%	2	22.2%	2	22.2%	28.6%
B	→	D	7	77.8%	2	22.2%	2	22.2%	28.6%
B	→	E	7	77.8%	6	66.7%	4	44.4%	57.1%
B	→	A	7	77.8%	6	66.7%	4	44.4%	57.1%
C	→	A	2	22.2%	6	66.7%	2	22.2%	100.0%
E	→	A	6	66.7%	6	66.7%	4	44.4%	66.7%
C	→	B	2	22.2%	7	77.8%	2	22.2%	100.0%
D	→	B	2	22.2%	7	77.8%	2	22.2%	100.0%
E	→	B	6	66.7%	7	77.8%	4	44.4%	66.7%
A	→	BC	6	66.7%	2	22.2%	2	22.2%	33.3%
B	→	AC	7	77.8%	2	22.2%	2	22.2%	28.6%
C	→	AB	2	22.2%	4	44.4%	2	22.2%	100.0%
A	→	BE	6	66.7%	4	44.4%	2	22.2%	33.3%
B	→	AE	7	77.8%	4	44.4%	2	22.2%	28.6%
E	→	AB	6	66.7%	4	44.4%	2	22.2%	33.3%
BC	→	A	2	22.2%	6	66.7%	2	22.2%	100.0%
AC	→	B	2	22.2%	7	77.8%	2	22.2%	100.0%
AB	→	C	4	44.4%	2	22.2%	2	22.2%	50.0%
BE	→	A	4	44.4%	6	66.7%	2	22.2%	50.0%
AE	→	B	4	44.4%	7	77.8%	2	22.2%	50.0%
AB	→	E	4	44.4%	6	66.7%	2	22.2%	50.0%

2. Find the followings:
   a. The lift ratio of the rule “Coke → Butter” is ___.

Coke → Butter, Lift ratio = Confidence(Coke, Butter)/p(Butter) = p(Coke, Butter)/p(Butter)p(Coke) = 0.444 / (0.778 * 0.667) = 0.86

   b. People who buy butter are ___ times more likely to also buy noodles.

Butter → Noodle, Lift ratio = 1.3

   c. People who buy ___ are at least 2 times more likely to also buy ___.

The following rules with Lift ratio = 2.25

1. Noodle → Bread & Butter
2. Bread & Butter → Noodle
3. Bread & Butter → Noodle & Coke
4. Noodle & Coke → Bread & Butter
5. Noodle → Bread, Butter & Coke
6. Bread, Butter & Coke → Noodle