Homework 5 Data Mining

Jan C. Bierowiec

Question 1

1. (25 points) Consider a dataset for frequent set mining as shown in the table below where we have 6 binary features and each row represents a transaction.

0	0	1	0	1	0
0	1	1	1	0	1
1	0	0	0	1	0
1	1	1	0	0	0
0	0	0	1	0	0
1	0	0	1	0	1
0	0	1	1	1	1
1	0	1	0	1	0
1	0	0	1	0	0
0	1	1	0	0	1

a) Illustrate the first three levels of the Apriori algorithm (set sizes 1, 2, and 3) for support threshold of 3 transactions, by identifying candidate sets and calculating their support. What are the maximal frequent sets discovered in the first 3 levels?

Scan Iteration 1

Candidate Set 1 (C1):

Transaction	Items
T1	{I3, I5}
T2	{I2, I3, I4, I6}
T3	{I1, I5}
T4	{I1, I2, I3}
T5	{I4}
T6	{I1, I4, I6}
T7	{I3, I4, I5, I6}
T8	{I1, I3, I5}
T9	{I1, I4}
T10	{I2, I3, I6}

Items	Support Count
I1	5
I2	3
I3	6
I4	5
I5	4
I6	4

From Candidate Set 1 (C1) to Large Set 1 (L1):

Scan Iteration 2

Candidate Set 2 (C2):

Item Set	Support Count
{I1, I2}	1
{I1, I3}	2
{I1, I4}	2
{I1, I5}	2
{I1, I6}	1
{I2, I3}	3
{I2, I4}	1
{I2, I5}	0
{I2, I6}	2
{I3, I4}	2
{I3, I5}	3
{I3, I6}	3
{I4, I5}	1
{I4, I6}	3
{I5, I6}	1

From Candidate Set 2 (C2) to Large Set 2 (L2):

Item Set	Support Count
{I2, I3}	3
{I3, I5}	3
{I3, I6}	3
{I4, I6}	3

Candidate Set 3 (C3):

Scan Iteration 3

No items in Large Set 3 (L3) since no item set meets the minimum support threshold of 3.

Item Set	Support Count
{I3, I5, I6}	1

b) Pick one of the maximal sets and check if any of its subsets are association rules with frequency at least 0.3 and confidence at least 0.6. Pleas explain your answer and show your work.

If we pick one of the maximal set {I2, I3} from L2, the association rules are as follows:

Let X = {I2, I3}: 3 Non-empty subsets of X are {I2}: 3 and {I3}: 6

So, we get the following association rules for {I2, I3}

1. $\{I2\} \Rightarrow \{I3\} = \frac{\text{Support count of } \{I2,I3\}}{\text{Support count of } \{I2\}} = \frac{3}{3} = 1.0$
2. $\{I3\} \Rightarrow \{I2\} = \frac{\text{Support count of } \{I2,I3\}}{\text{Support count of } \{I3\}} = \frac{3}{6} = 0.5$

Since the given confidence cut-off is 0.6, we would report only the first rule which is {I2} ⇒ {I3}, confidence = 1.0

Question 2

2. (25 points) Table 1 shows how many transactions containing beer and/or nuts among 10000 transactions. Answer the following questions.

	Beer	No Beer	Total
Nuts	50	800	850
No Nuts	150	9000	9150
Total	200	9800	10000

Table 1: Transactions Involving Beer and Nuts

a) How many possible association rules can be generated based on Table 1?

There are two unique items, those being Beer & No Beer. There are also two transactions that are related to the unique items, those being Nuts & No Nuts.

The number of possible association rules would be

$$3^2 - 2^{(2+1)} + 1 = 3^2 + 2^3 + 1 = 9 - 8 + 1 = 2$$

b) Calculate support, confidence, lift for each of the rules in (a).

	Beer & Nuts	No Beer & Nuts
Support	$\frac{50}{10000} = 0.005$	$\frac{800}{10000} = 0.08$
Confidence	$\frac{50}{200} = 0.25$	$\frac{800}{9800} = 0.0816$
Lift	$\frac{0.25}{\frac{850}{10000}} = 2.941$	$\frac{0.0816}{\frac{850}{10000}} = 0.096$

c) What are your conclusions of the relationship between buying/not buying beer and buying/not buying nuts, based on the above measures?

There is both a positive & negative association. There is a positive association between buying beer and buying nuts given that the lift value is greater than 1, 2.941 > 1. There is a negative association between not buying beer and buying nuts given that the lift value is less than 1, 0.096 < 1.

The conclusion that can be made here is that customers who buy more beer, are likely to buy more nuts, while those who do not buy beer are less likely to buy nuts.

Question 3

3. (25 points) In the GSP algorithm, suppose we have the length-3 frequent pattern set L_3 as follows:

```
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {5} {3 4} >
```

Generate length-4 candidates set C_4. Show your work by writing down the details of the join and prune steps.

Frequent Sequences L3
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {5} {3 4} >

Step 1 is to join the generation of candidates set C4:

< ~~{2}~~ {3} {4} >	< {2} {3} ~~{4}~~ >
	< {2 5} ~~{3}~~ >
	< {3} {4} ~~{5}~~ >
	< {1} {2} ~~{3}~~ >
	< {1} {2 ~~5}~~ >
	< {1} {5} ~~{3}~~ >
	< {5} {3 ~~4}~~ >

< ~~{2~~ 5} {3} >	< {2} {3} ~~{4}~~ >
	< {2 5} ~~{3}~~ >
	< {3} {4} ~~{5}~~ >
	< {1} {2} ~~{3}~~ >
	< {1} {2 ~~5}~~ >
	< {1} {5} ~~{3}~~ >
	< {5} {3 ~~4}~~ >

< ~~{3}~~ {4} {5} >	< {2} {3} ~~{4}~~ >
	< {2 5} ~~{3}~~ >
	< {3} {4} ~~{5}~~ >
	< {1} {2} ~~{3}~~ >
	< {1} {2 ~~5}~~ >
	< {1} {5} ~~{3}~~ >
	< {5} {3 ~~4}~~ >

< ~~{1}~~ {2} {3} >	< {2} {3} ~~{4}~~ >
	< {2 5} ~~{3}~~ >
	< {3} {4} ~~{5}~~ >
	< {1} {2} ~~{3}~~ >
	< {1} {2 ~~5}~~ >
	< {1} {5} ~~{3}~~ >
	< {5} {3 ~~4}~~ >

< ~~{1}~~ {2 5} >	< {2} {3} ~~{4}~~ >
	< {2 5} ~~{3}~~ >
	< {3} {4} ~~{5}~~ >
	< {1} {2} ~~{3}~~ >
	< {1} {2 ~~5}~~ >
	< {1} {5} ~~{3}~~ >
	< {5} {3 ~~4}~~ >

< ~~{1}~~ {5} {3} >	< {2} {3} ~~{4}~~ >
	< {2 5} ~~{3}~~ >
	< {3} {4} ~~{5}~~ >
	< {1} {2} ~~{3}~~ >
	< {1} {2 ~~5}~~ >
	< {1} {5} ~~{3}~~ >
	< {5} {3 ~~4}~~ >

< ~~{5}~~ {3 4} >	< {2} {3} ~~{4}~~ >
	< {2 5} ~~{3}~~ >
	< {3} {4} ~~{5}~~ >
	< {1} {2} ~~{3}~~ >
	< {1} {2 ~~5}~~ >
	< {1} {5} ~~{3}~~ >
	< {5} {3 ~~4}~~ >

Candidate Generation
< {2} {3} {4} {5} >
< {2 5} {3 4} >
< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3} {4} >
< {5} {3 4} {5} >

Step 2: Pruning: Candidate Pruning To Check if all k-1 length's subsequences of a candidates is in L_{k-1}

< {2} {3} {4} {5} >	< {3} {4} {5} >
	< {2} {4} {5} >
	< {2} {3} {5} >
	< {2} {3} {4} >

< {2 5} {3 4} >	< {5} {3 4} >
	< {2} {3 4} >
	< {2 5} {4} >
	< {2 5} {3} >

< {1} {2} {3} {4} >	< {2} {3} {4} >
	< {1} {3} {4} >
	< {1} {2} {4} >
	< {1} {2} {3} >

< {1} {2 5} {3} >	< {2 5} {3} >
	< {1} {5} {3} >
	< {1} {2} {3} >
	< {1} {2 5} >

< {1} {5} {3} {4} >	< {5} {3} {4} >
	< {1} {3} {4} >
	< {1} {5} {4} >
	< {1} {5} {3} >

< {5} {3 4} {5} >	< {3 4} {5} >
	< {5} {4} {5} >
	< {5} {3} {5} >
	< {5} {3 4} >

Length 3 Frequent sequences L3
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {5} {3 4} >

Candidate Pruning
< {1} {2 5} {3} >

Final Results after Prunning:

Frequent sequences L3
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {5} {3 4} >

Candidate Generation C4
< {2} {3} {4} {5} >
< {2 5} {3 4} >
< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3} {4} >
< {5} {3 4} {5} >

Candidate Pruning L4
< {1} {2 5} {3} >

Question 4

4. (25 points) For the following two time series:

$$X = [32, 36, 27, 37, 35, 40, 34, 33, 25, 29]$$
$$Y = [31, 32, 32, 30, 37, 39, 29, 34, 25, 26]$$

Calculate the DTW distance between X and Y and point out the optimal warping path. The local cost function is defined as the absolute difference of the two values, e.g., $c(x_1, y_1) = d(32 - 31) = 1$.

The cost matrix for the given two time series:

26	6	10	1	11	9	14	8	7	1	3
25	7	11	2	12	10	15	9	8	0	4
34	2	4	7	3	1	6	0	1	9	5
29	3	7	2	8	6	11	5	4	4	0
39	7	3	12	2	4	1	5	6	14	10
37	5	1	10	0	2	3	3	4	12	8
30	2	6	3	7	5	10	4	3	5	1
32	0	4	5	5	3	8	2	1	7	3
32	0	4	5	5	3	8	2	1	7	3
31	1	5	4	6	4	9	3	2	6	2
	32	36	27	37	35	40	34	33	25	29

Highlighted is the optimal warping path
DTW Distance between X,Y is 20.

26	33	37	19	29	31	36	33	31	18	20
25	27	27	18	24	22	28	25	24	17	21
34	20	16	16	12	13	19	16	17	26	29
29	18	14	9	17	16	22	16	20	24	24
39	15	7	16	10	12	11	16	22	34	42
37	8	4	14	8	10	13	16	20	32	40
30	3	7	8	15	20	28	30	31	34	35
32	1	5	10	15	18	26	28	29	36	39
32	1	5	10	15	18	26	28	29	36	39
31	1	6	10	16	20	29	32	34	40	42
	32	36	27	37	35	40	34	33	25	29