Computer Science 145, Homework 5

Michael Wu
UID: 404751542

March 6th, 2019

Problem 1

a) In the first iteration we obtain the following table.

Item Set	Support
$\{a\}$	6
$\{b\}$	8
$\{c\}$	6
$\{d\}$	4
$\{e\}$	2
$\{i\}$	1
$\{j\}$	1
$\{k\}$	1

We can ignore $\{i\}$, $\{j\}$, and $\{k\}$ since they are not frequent. Then we have the following length two candidates and supports.

Item Set	Support
$\{a, b\}$	4
$\{a, c\}$	4
$\{a, d\}$	2
$\{a, e\}$	2
$\{b, c\}$	4
$\{b, d\}$	4
$\{b, e\}$	2
$\{c, d\}$	1
$\{c, e\}$	1
$\{d, e\}$	0

We can ignore $\{c, d\}$, $\{c, e\}$, and $\{d, e\}$ since they are not frequent. Then we generate the length three candidates by looking at the length two frequent patterns and combining the ones that only differ in the last element. During candidate generation we can prune $\{a, c, d\}$, $\{a, c, e\}$, $\{a, d, e\}$, $\{b, c, d\}$, $\{b, c, e\}$, and $\{b, d, e\}$ since they are supersets of the previous length two sets that are not frequent. We then have the following length three candidates and supports.

Item Set	Support
$\{a, b, c\}$	2
$\{a, b, d\}$	2
$\{a, b, e\}$	2

During generation of length four candidates, we will prune away all the candidates because they contain one of $\{c, d\}$, $\{c, e\}$, or $\{d, e\}$ which are not frequent. Thus our frequent patterns are

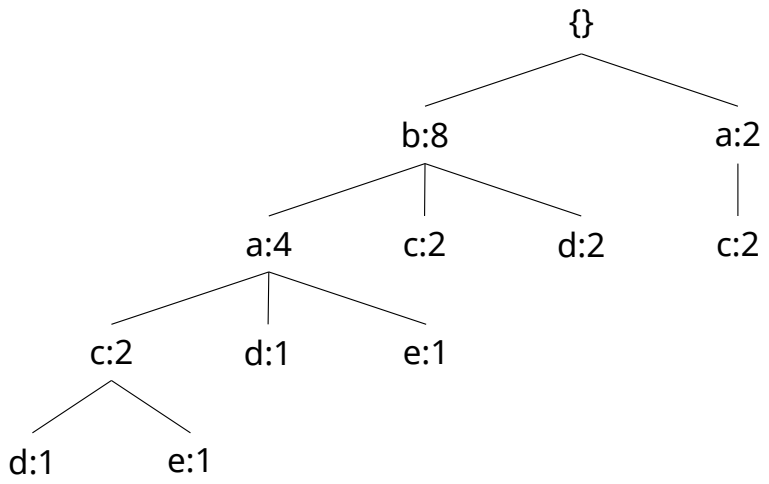
$$\{a, b, c\}, \{a, b, d\}, \{a, b, e\}$$

and any subset of these item sets.

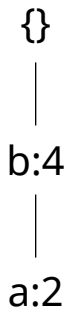
b) Our initial F-list contains a to e. If we arrange by frequency the list becomes $\{b, a, c, d, e\}$. The ordered frequent items in the database are shown in the following table.

TID	Ordered Frequent Items
1	{b,c}
2	{b,a,d}
3	{a,c}
4	{b,d}
5	{b,a,c,e}
6	{b,c}
7	{a,c}
8	{b,a,e}
9	{b,d}
10	{b,a,c,d}

The FP-tree is shown below.



c) The conditional pattern base for d is $\{bac : 1, ba : 1, b : 2\}$. We can see that only b and a are frequent here, so the conditional FP-tree is the one shown below.



d) Based on the conditional FP-tree we have the frequent patterns $\{bad : 2, ad : 2, bd : 4, d : 4\}$.

Problem 2

a)

b)

c)

Problem 3

a)

$$\text{confidence}(\text{Beer} \Rightarrow \text{Nuts}) = P(\text{Nuts}|\text{Beer}) = \frac{150}{500} = 0.3$$
$$\text{confidence}(\text{Nuts} \Rightarrow \text{Beer}) = P(\text{Beer}|\text{Nuts}) = \frac{150}{850} = 0.1765$$
$$\text{all confidence}(\text{Beer}, \text{Nuts}) = 0.1765$$
$$\text{lift}(\text{Beer}, \text{Nuts}) = \frac{150 \times 10000}{500 \times 850} = 3.5294$$
$$\chi^2 = \frac{(150 - 42.5)^2}{42.5} + \frac{(700 - 807.5)^2}{807.5} + \frac{(350 - 457.5)^2}{457.5} + \frac{(8800 - 8692.5)^2}{8692.5} = 312.81$$

b) Considering that the overall probability of buying nuts is 0.085 and the overall probability of buying beer is 0.05, I would say that the confidence values indicate that buying beer increases the chance of buying nuts and vice versa. The lift and the χ^2 statistics reveal that these two variables are correlated as well. The lift is greater than 1, so these variables are positively correlated. The χ^2 is very high, which means that these values are not independent. If we were to take the p-value of this using one degree of freedom, we would find that it would be very close to zero so we would reject the null hypothesis that these two events are independent.

Problem 4

a) This contains six elements. The length of s is eight. It contains $2^6 - 1 = 63$ non-empty subsequences.

b) In order to perform the join step, we must look for items that match when removing the first item from one and the last item from another. Removing the first items yields the following sequences.

$$\{\langle ce\rangle, \langle (cd)\rangle, \langle ce\rangle, \langle (cd)\rangle, \langle bd\rangle, \langle bc\rangle\}$$

Removing the last items yields the following sequences.

$$\{\langle (ac)\rangle, \langle bc\rangle, \langle bc\rangle, \langle ac\rangle, \langle (ab)\rangle, \langle (ab)\rangle\}$$

The only matching sequence is $\langle bc\rangle$, so joining the corresponding sequences yields the following sequences.

$$\{\langle (ab)(cd)\rangle, \langle (ab)ce\rangle\}$$

Then in order to prune, we must check that all the subsequences of the joined sequences are in L_3. For the first sequence $\langle b(cd)\rangle$, $\langle a(cd)\rangle$, $\langle (ab)d\rangle$, and $\langle (ab)c\rangle$ are all in L_3 so it can stay. For the next sequence $\langle ace\rangle$ and $\langle (ab)e\rangle$ are not in L_3 so we can remove it. Finally we have the following set of candidate 4-sequences C_4.

$$C_4 = \{\langle (ab)(cd)\rangle\}$$