Northeastern University, Data Mining Techniques - CS6220
Fall 2017

Solutions to Homework 1

Nakul Camasamudram

September 28, 2017

4. The Lift Measure (10 pts)

Solution:
Given,

$$\begin{aligned} lift(X \rightarrow Y) &= 1 \\ \frac{c(X \rightarrow Y)}{s(Y)} &= 1 \\ \frac{s(X \cup Y)}{s(X) \cdot s(Y)} &= 1 \\ s(X \text{and} Y) &= s(X) \cdot s(Y) \end{aligned}$$

Given some transaction A, $s(A) = \frac{\sigma(A)}{N}$ is the probability of transaction A in a given dataset D. Hence, in the above equations, we see that the probability of occurrence of the antecedent(X) and that of the consequent(Y) is the product of the individual occurrences of the antecedent and consequent, which is the requirement for independence.

5. Apriori (10 pts)

Solution:

a. **Pass 1:**

Candidate set of size = 1	Support
{bananas}	0.3
{carrots}	0.6
{figs}	0.4
{apples}	0.5
{donuts}	0.5
{eggs}	0.4

Frequent 1-itemset =
{bananas},{carrots},{figs},{apples},{donuts},{eggs}

Pass 2:

Candidate Itemset Size = 2	Support
{bananas, carrots}	0.3
{bananas, figs}	0.2
{bananas, apples}	0.1
{bananas, donuts}	0.1
{bananas, eggs}	0.0
{carrots, figs}	0.3
{carrots, apples}	0.2
{carrots, donuts}	0.2
{carrots, eggs}	0.3
{figs, apples}	0.1
{figs, donuts}	0.3
{figs, eggs}	0.1
{apples, donuts}	0.2
{apples, eggs}	0.2
{donuts, eggs}	0.1

Frequent 2-itemset =
{bananas, carrots}, {carrots, figs}, {carrots, eggs}, {figs, donuts}

Pass 3:

Candidate Itemset Size = 2	Support
{bananas, carrots, donuts}	0.1
{bananas, carrots, figs}	0.2
{bananas, carrots, eggs}	0.0
{bananas, eggs, figs}	0.0
{bananas, donuts, figs}	0.1
{carrots, donuts, eggs}	0.1
{carrots, donuts, figs}	0.2
{carrots, eggs, figs}	0.1
{donuts, eggs, figs}	0.1
{apples, bananas, carrots}	0.1
{apples, carrots, figs}	0.0
{apples, carrots, eggs}	0.1
{apples, donuts, figs}	0.1

Frequent 3-itemset – {}

b. A maximal frequent itemset is a frequent itemset for which none of its immediate supersets are frequent. These are the maximal frequent sets discovered in the first three phases: {apples}, {bananas, carrots}, {carrots, figs}, {carrots, eggs}, {figs, donuts}

c. Among the above maximal frequent sets, the association rule $bananas \rightarrow carrots$ has support of at least 0.3 and confidence of at least 0.7. The proof for support being greater than 0.3 is shown in 5a. The proof of confidence being greater that 0.7 is as follows:

$$\begin{aligned} \frac{s(bananas, carrots)}{s(bananas)} &= \frac{0.3}{0.3} \\ &= 1 \end{aligned}$$

6. FP-Growth (10 pts)

Solution:

a.

Original Transactions

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D
- B, D, E
- B, C, F
- A, B, C
- A, B, D, E
- A, C, E

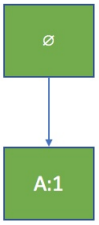
Item	Frequency
A	8
B	6
C	6
D	4
E	4
F	1

Filtered & Sorted

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D
- B, D, E
- B, C
- A, B, C
- A, B, D, E
- A, C, E

Transactions:

- A

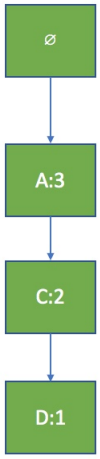


Transactions:

- A
- A, C

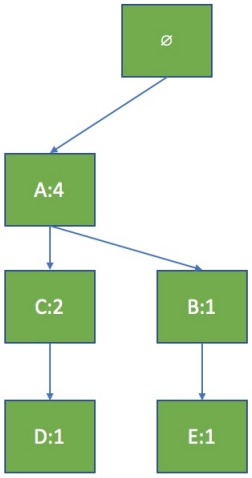
Transactions:

- A
- A, C
- A, C, D



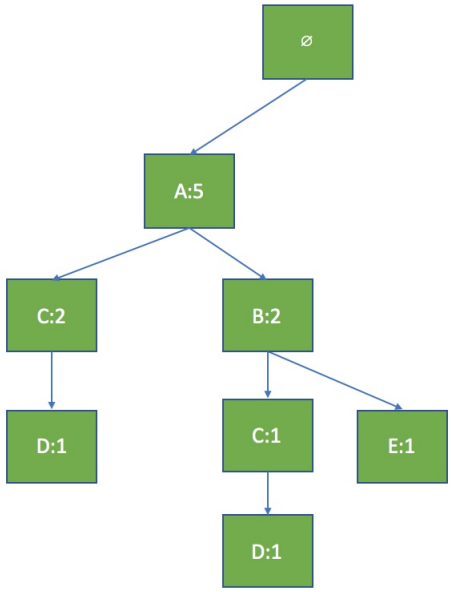
Transactions:

- A
- A, C
- A, C, D
- A, B, E



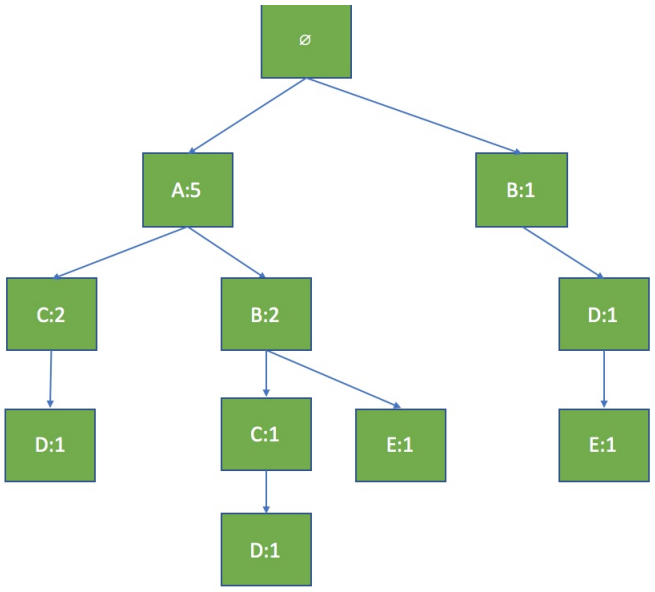
Transactions:

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D



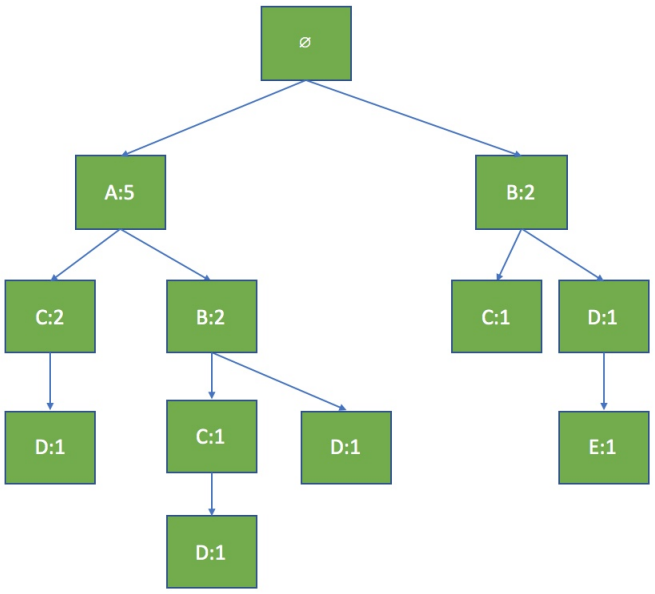
Transactions:

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D
- B, D, E



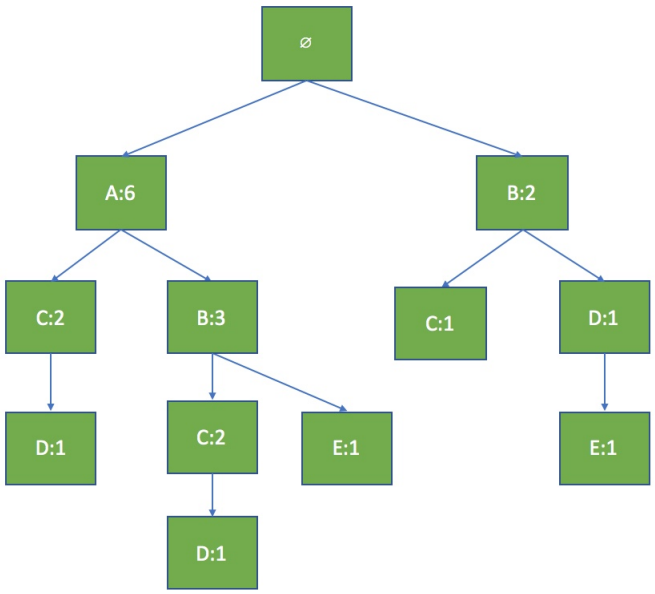
Transactions:

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D
- B, D, E
- B, C



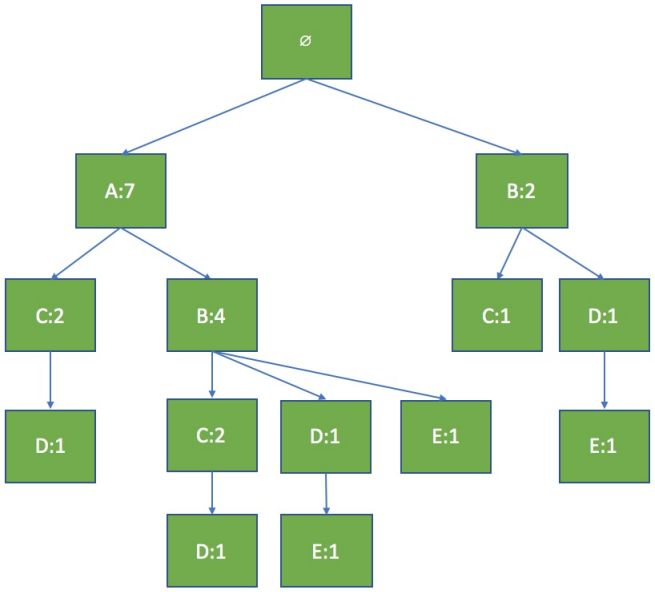
Transactions:

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D
- B, D, E
- B, C
- A, B, C



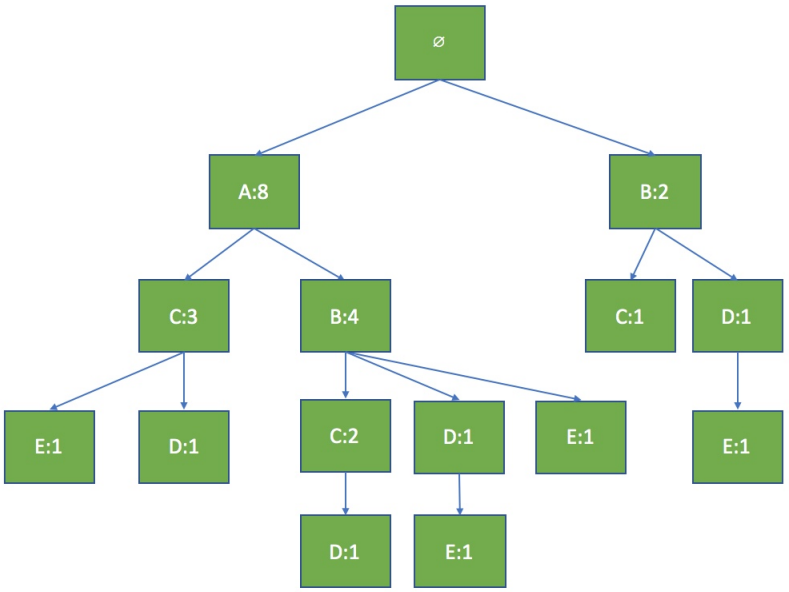
Transactions:

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D
- B, D, E
- B, C
- A, B, C
- A, B, D, E



Transactions:

- A
- A, C
- A, C, D
- A, B, E
- A, B, C, D
- B, D, E
- B, C
- A, B, C
- A, B, D, E
- A, C, E



7. A Scaled FP-Growth Pipeline (20 pts)

Solution:

b. Run "./itemsets2sparsearff.py kosarak.dat ¿ kosarak.arff". The program took 27.738s to run.

c. It took 50s to load "kosarak.arff" into Weka.

d. The below figure shows the two rules discovered by the FPGrowth algorithm

```
1. [i11=1, i218=1, i148=1]: 50098 ==> [i6=1]: 49866   <conf:(1)> lift:(1.64) lev:(0.02) conv:(84.4)
2. [i11=1, i148=1]: 55759 ==> [i6=1]: 55230   <conf:(0.99)> lift:(1.63) lev:(0.02) conv:(41.3)
```

e. Run 1: 0.01s; Run 2: 0.00s; Run 3: 0.01s; Run 4: 0.01s; Run 5: 0.00s. Therefore the average FP-growth runtime in Weka is 0.01s as compared to 77.738s for dataset conversion and loading.