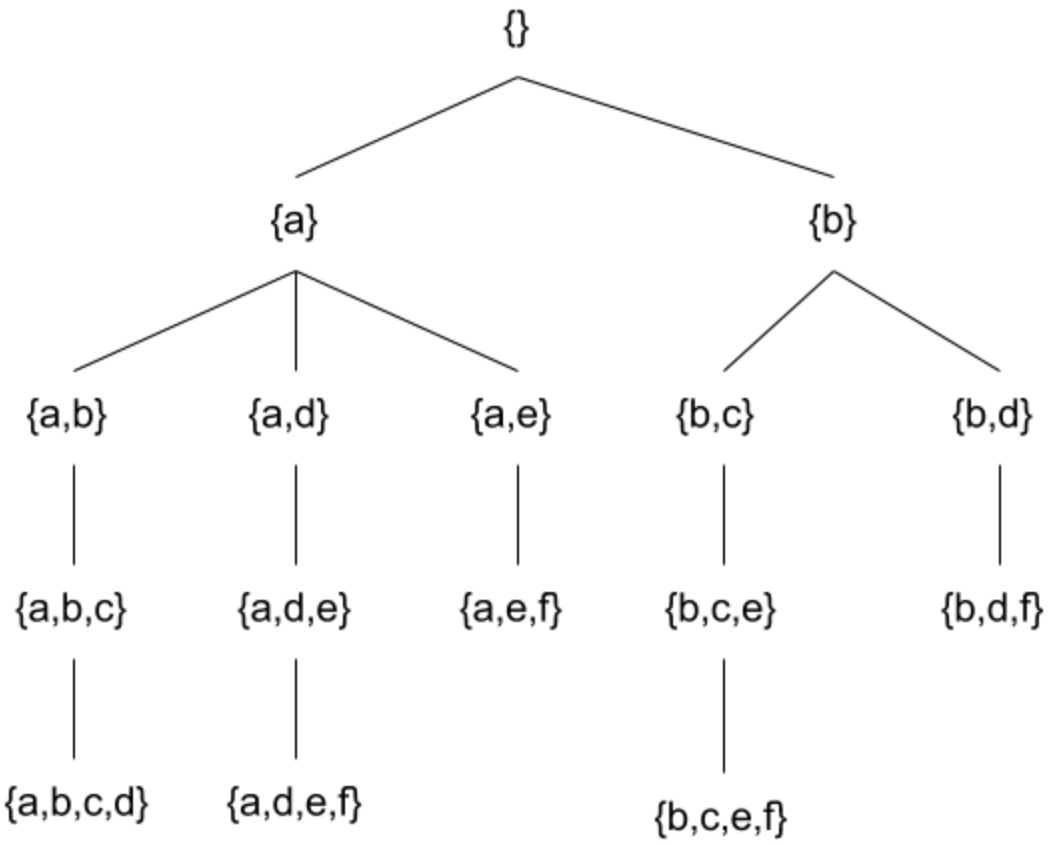
Roy Chan
chanroyc@sfu.ca
301202770

CMPT 459 Fall 2017
Data Mining
Martin Ester
TA: Zhilin Zhang

Assignment 5

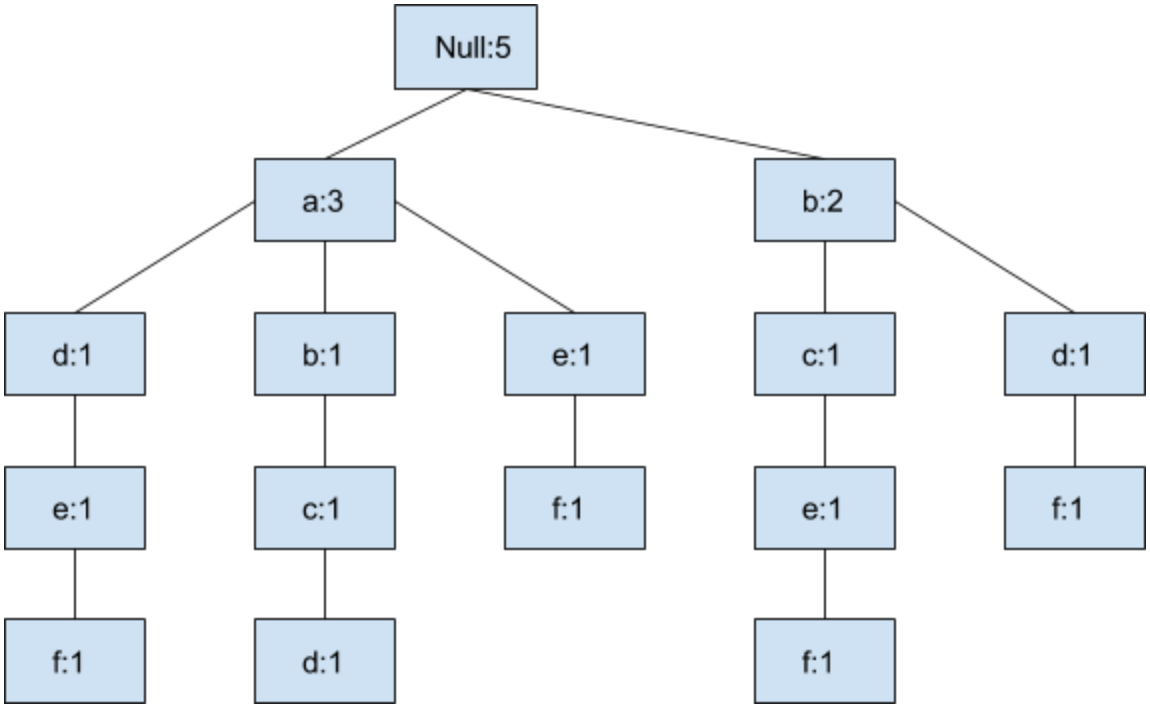
Assignment 5.1

a) Enumeration tree of the frequent itemsets with the lexicographical ordering a,b,c,d,e,f.

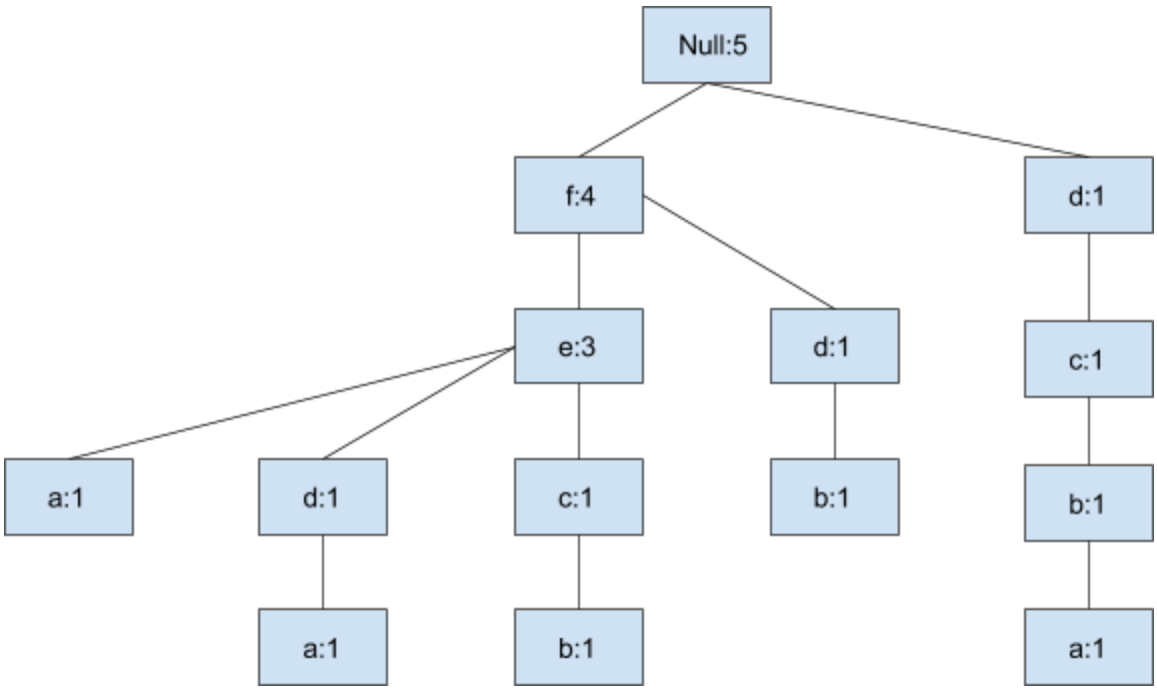


Roy Chan
chanroyc@sfu.ca
301202770

b) Prefixed-based FP-tree with lexicographic ordering a,b,c,d,e,f.



c) FP-tree with lexicographic ordering f,e,d,c,b,a



Roy Chan
chanroyc@sfu.ca
301202770

d) The second FP-tree is smaller, i.e. has fewer nodes because the lexicographic ordering of the second tree prioritizes the most frequent item *f*.

Assignment 5.2

a) Using the apriori algorithm principles, the following modified algorithm will generate only frequent itemsets which correspond to classification rules.

i) Start with itemsets containing just a single item.

ii) Keep only the itemsets that meet the minimum support/confidence threshold and remove itemsets that do not.

iii) Use itemsets from ii) and generate all possible itemset configurations

iv) repeat steps ii) and iii) until there are no more new itemsets

b) i) Determine all rules to be applied using the modified algorithm from a). Define weights for the classification rules, giving heavier weights to rules with higher support/confidence values.

ii) Based on the weights obtained in step i), sort them by class item and then add up all the weights to determine which class item is most frequent and aggregate their predictions to predict class label of the test record.

Assignment 5.3

a) Point 6 and 14 have the highest outlier score of 4 using the knn-distance based algorithm for outlier detection with k=2.

|1-2| = 1	|2-1| = 1	|6-1| = 5	|8-1| = 7	|10-1| = 9	|12-1| = 11
|1-2| = 1	|2-2| = 0	**|6-2| = 4**	|8-2| = 6	|10-2| = 8	|12-2| = 10
|1-2| = 1	**|2-2| = 0**	|6-2| = 4	|8-2| = 6	|10-2| = 8	|12-2| = 10
|1-2| = 1	|2-2| = 0	|6-2| = 4	|8-2| = 6	|10-2| = 8	|12-2| = 10
|1-2| = 1	|2-2| = 0	|6-2| = 4	|8-2| = 6	|10-2| = 8	|12-2| = 10
|1-6| = 5	|2-6| = 4	|6-2| = 4	|8-2| = 6	|10-2| = 8	|12-2| = 10
|1-8| = 7	|2-8| = 6	|6-8| = 2	|8-6| = 2	|10-6| = 4	|12-6| = 6
|1-10| = 9	|2-10| = 8	|6-10| = 4	**|8-10| = 2**	|10-8| = 2	|12-8| = 4
|1-12| = 11	|2-12| = 10	|6-12| = 6	|8-12| = 4	**|10-12| = 2**	|12-10| = 2
|1-14| = 13	|2-14| = 12	|6-14| = 8	|8-14| = 6	|10-14| = 4	**|12-14| = 2**

Roy Chan
chanroyc@sfu.ca
301202770

|14-1| = 13
|14-2| = 12
|14-2| = 12
|14-2| = 12
|14-2| = 12
|14-2| = 12
|14-6| = 8
|14-8| = 6
|14-10| = 4
|14-12| = 2

b) Point 14 has the highest outlier score of 1.5 using the LOF algorithm for outlier detection with k=2.

ARk(1): 2/(1+1) = 1
ARk(2): 2/(0+0) = 0
ARk(2): 2/(0+0) = 0
ARk(2): 2/(0+0) = 0
ARk(2): 2/(0+0) = 0
ARk(2): 2/(0+0) = 0
ARk(6): 2/(2+4) = 0.33
ARk(8): 2/(2+2) = 0.5
ARk(10):2/(2+2) = 0.5
ARk(12): 2/(2+2) = 0.5
ARk(14): 2/(2+4) = 0.33
LOF(1): (0+0)*(1+1)/4 = 0
LOF(2): (0+0)*(0+0)/4 = 0
LOF(2): (0+0)*(0+0)/4 = 0
LOF(2): (0+0)*(0+0)/4 = 0
LOF(2): (0+0)*(0+0)/4 = 0
LOF(2): (0+0)*(0+0)/4 = 0
LOF(6): (0.5+0)*(2+4)/4 = 0.75
LOF(8): (0.33+0.5)*(2+2)/4 = 0.83
LOF(10): (0.5+0.5)*(2+2)/4 = 1
LOF(12): (0.33+0.5)*(2+2)/4 =0.83
LOF(14): (0.5+0.5)*(2+4)/4 = 1.5

c) The difference between the results of the knn-distance based algorithm and the results of the LOF algorithm is that the LOF algorithm is an extension of the knn-distance based algorithm which provides more accurate results. LOF does not only use the average reachability distance ARk(p) for its outlier score, but it also factors in the average reachability distance of p's neighbours to their knns, which is where the algorithm's name comes from (Local Outlier Factor).

Roy Chan
chanroyc@sfu.ca
301202770