CS 6220 Assignment 4

Boyuan Yu
Boyuan-Yu
yu.boyu@northeastern.edu

Please see .ipynb file for Questions 1 and 2

Question 3

3a.

{1,2,3,4}, {1,2,3,5}, {1,2,4,5}, {1,3,4,5}, {2,3,4,5}

3b.

{1,2,3,4}, {1,2,3,5}, {1,2,4,5}, {1,3,4,5}, {2,3,4,5}

3c.

We need to find the 4-itemset which have all subset in 3-itemsets.
{1,2,3,4}: all (k-1)-subsets are frequent: {1,2,3}, {1,2,4}, {1,3,4}, {2,3,4}.

Question 4

4a.

There are 7 items in the dataset. Therefore the total number of rules is
2^(7) - 2 = 128 - 2 = 126

4b.

To calculate the confidence of the rule {Milk, Diapers} ⇒ {Butter}, we need to find the support for the itemset {Milk, Diapers, Butter} and the support for the itemset {Milk, Diapers}.
The support for {Milk, Diapers, Butter} is 2 since it appears in transactions 2 and 7.
The support for {Milk, Diapers} is 4 since it appears in transactions 2, 3, 5, and 7.
The confidence of the rule {Milk, Diapers} ⇒ {Butter} is: 2 / 4 = 0.5

4c.

The support for the rule {Milk, Diapers} ⇒ {Butter} is the number of transactions that contain all three items, divided by the total number of transactions in the dataset. 2 / 10 = 0.2

4d.

True. Since {a,b} is a subset of {a,b,c,d}, it follows that the support count for {a,b} must be greater than or equal to the support count for {a,b,c,d}. If {a,b,c,d} is a frequent itemset, it means that it appears frequently enough in the dataset to meet the minimum support threshold. Therefore, the support count for {a,b} is greater than or equal to the minimum support count.

4e.

True. Since {a,b,c} contains all the items in each of these itemsets, it means that {a,b,c} is a superset of each of these itemsets. It would survive the candidate pruning step of the Apriori algorithm.

4f.

False. {b} is a subset of {a.b} and {b,c}, so the support of {b} is greater or equal than both 20 and 30. Therefore it cannot be smaller than 30.

4g.

False. The number of size-2 frequent itemsets is C(5,2) = 10.

4h.