Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 4

Due: March 1, 2023 (100 points)

Shu Yi Wang
shuyi0220
wang.shuyi@northeastern.edu

K-Means

The normalized automobile distributor timing speed and ignition coil gaps for production F-150 trucks over the years of 1996, 1999, 2006, 2015, and 2022. We have stripped out the labels for the five years of data.

Each sample in the dataset is two-dimensional, i.e. $x_i \in \mathbb{R}^2$(one dimension for timing speed and the other for coil gaps), and there are N = 5000 instances in the data.

Question 1 [20 pts total]

Please see Colab .ipynb file.

Question 2 [30 pts total]

Please see Colab .ipynb file.

Market Basket Analysis and Algorithms

Consider F_3 as the following set of frequent 3-itemsets:

$\{1,2,3\},\ \{1,2,4\},\ \{1,2,5\},\ \{1,3,4\}$
$\{2,3,4\},\ \{2,3,5\},\ \{3,4,5\}\,.$

Assume that there are only five items in the data set.

Question 3 [25 pts total]

[10 pts] Question 3a.) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

$F1 = \{1\},\{2\},\{3\},\{4\},\{5\}$, $F4$ candidates can be generated by $F3 \times F1$ where all the 3-itemsets in F3 merge with a frequent item in F1:

$\{1,2,3\} \times F1$ we get $\{1,2,3\ ,4\}$ and $\{1,2,3,5\}$ with 4 unique items in sets.

$\{1,2,4\} \times F1$ with new 4-itemsets $\{1,2,4,5\}$ for F4 candidates.

$\{1,2,5\} \times F1$ with no new 4-itemsets for F4 candidates.

$\{1,3,4\} \times F1$ with new 4-itemsets $\{1,3,4,5\}$ for F4 candidates.

$\{2,3,4\} \times F1$ with new 4-itemsets $\{2,3,4,5\}$ for F4 candidates.

$\{2,3,5\} \times F1$ with no new 4-itemsets for F4 candidates.

$\{3,4,5\} \times F1$ with no new 4-itemsets for F4 candidates.

Therefore,

$F4\ candidates = \{1,2,3,4\},\{1,2,3,5\},\{1,2,4,5\},\{1,3,4,5,\},\{2,3,4,5\}$

[10 pts] Question 3b.) List all candidate 4-itemsets obtained by the candidate generation procedure in A Priori, using $F_{k-1} \times F_1$.

Using candidate generation procedure in Apriori that we attempt to merge two frequent subsets in F3 if their first 2 items are identical.

$$F3 = \{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}$$

$$Merge(\{1,2,3\}, \{1,2,4\}) = \{1,2,3,4\}$$

$$Merge(\{1,2,3\}, \{1,2,5\}) = \{1,2,3,5\}$$

$$Merge(\{1,2,4\}, \{1,2,5\}) = \{1,2,4,5\}$$

$$Merge(\{2,3,4\}, \{2,3,5\}) = \{2,3,4,5\}$$

Therefore,

$$F4\ candidates = \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{2,3,4,5\}$$

[5 pts] Question 3c.) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

From Question 3b) using Apriori algorithm we get the candidates:

$$F4\ candidates = \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{2,3,4,5\}$$

And as we known
$$F3 = \{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}$$

Then we want to prune out F4 candidates using the rule that all the 3-itemsets inside candidates exist in F3.

We don't prune $\{1,2,3,4\}$ because $\{1,2,3\}$, $\{1,2,4\}, \{1,3,4\}$, $\{2,3,4\}$ are all frequent (in F3).

We prune $\{1,2,3,5\}$ because $\{1,3,5\}$ is infrequent (not in F3).

We prune $\{1,2,4,5\}$, $and\, \{2,3,4,5\}$ because $\{2,4,5\}$ is infrequent (not in F3).

Therefore, after pruning:

$F4 = \{1, 2, 3, 4\}$

Question 4 [25 pts total]

Consider the following table for questions 4a) to 4c):

Transaction ID	Items
1	**{Beer, Diapers}**
2	**{Milk, Diapers, Bread, Butter}**
3	**{Milk, Diapers, Cookies}**
4	**{Bread, Butter, Cookies}**
5	**{Milk, Beer, Diapers, Eggs}**
6	**{Beer, Cookies, Diapers}**
7	**{Milk, Diapers, Bread, Butter}**
8	**{Bread, Butter, Diapers}**
9	**{Bread, Butter, Milk}**
10	**{Beer, Butter, Cookies}**

[3 pts] Question 4a.) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

To apply association rules, we first need to figure out the numbers of unique items we have. Go through all the transcations, we can find there are 7 unique items which are $\{Beer, Diapers, Milk, Bread, Butter, Cookies, Eggs\}$.

So the maximum numbers of association rules can be extracted from data

$= 3^{UItems} - 2^{UItems+1} + 1$

$= 3^7 - 2^{7+1} + 1 = 1932$

[3 pts] Question 4b.) What is the confidence of the rule {Milk, Diapers} ⇒ {Butter}?

To find out the confidence of {Milk, Diapers} ⇒ {Butter}, we need to find out how often {Milk, Diapers} appear in transactions that contain {Butter}.

Numbers of transactions {Milk, Diapers} include: 4 (Transaction ID: 2,3,5,7)

Numbers of transactions amongTransaction ID: 2,3,5,7 include {Butter}: 2 (Transaction ID:2,7)

$Confidence = 2 / 4 = 0.5$

[3 pts] Question 4c.) What is the support for the rule {Milk, Diapers} ⇒ {Butter}?

To find out the support of {Milk, Diapers} ⇒ {Butter}, we need to find out how often {Milk, Diapers, Butter} appear in transactions.

Numbers of transactions {Milk, Diapers, Butter} include: 2 (Transaction ID: 2,7)

Total of transactions: 10

$Support = 2 / 10 = 0.2$

[3 pts] Question 4d.) True or False with an explanation: Given that $\{a, b, c, d\}$ is a frequent itemset, $\{a, b\}$is always a frequent itemset.

True. If $\{a, b, c, d\}$ is a frequent itemset. The subsets of $\{a, b, c, d\}$ are always frequent itemset too. Since for all transactions that include itemset $\{a, b, c, d\}$, they should all include their subset and those transactions that do not include $\{a, b, c, d\}$ might also include the subset. Therefore the support of any subset of $\{a, b, c, d\}$ should be larger or equal to the support of $\{a, b, c, d\}$. And so the subsets are all frequent itemset.

[3 pts] Question 4e.) True or False with an explanation: Given that $\{a, b\}\{b, c\}$and $\{a, c\}$ are frequent itemsets, $\{a, b, c\}$is always frequent.

False. It is not necessary that if $\{a, b\}$,$\{b, c\}$ and $\{a, c\}$ are frequent, $\{a, b, c\}$ is frequent too. Since $\{a, b, c\}$ might not appear in the same transactions as any one of the subset. Therefore the support of $\{a, b, c\}$ should be less or equal to the support of its subset. And so t$\{a, b, c\}$ is not always frequent itemset.

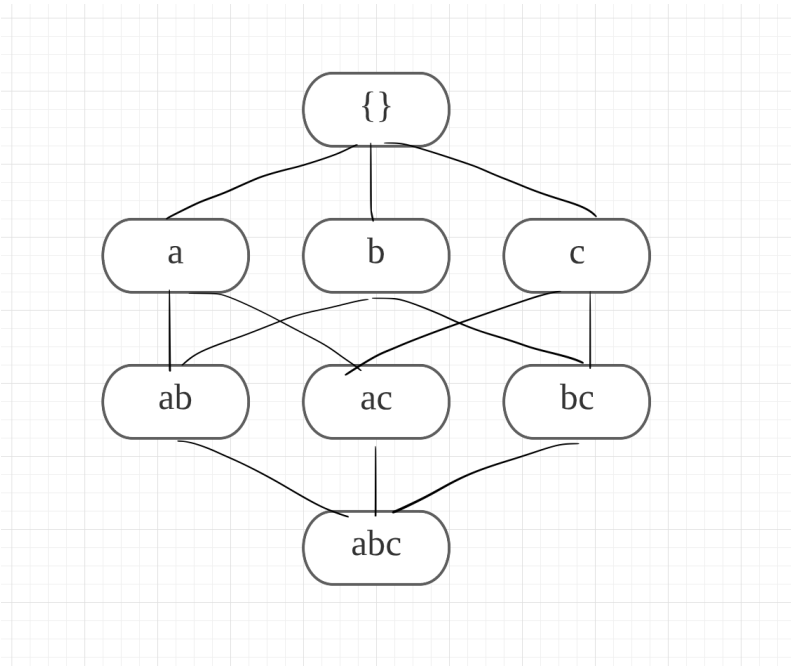
[3 pts] Question 4f.) True or False with an explanation: Given that the support of $\{a, b\}$ is 20 and the support of $\{b, c\}$ is 30, the support of $\{b\}$ is larger than 20 but smaller than 30.

False. The support of $\{b\}$ is how often b appears in transactions. So b should at least appears in 30 transactions based on the question (e.g. 20 transactions are $\{a, b, c\}$and 10 transactions are $\{b, c\}$, $20 + 10 = 30$ transactions). And it very likely b appears in more transactions, so the support of $\{b\}$ is always larger than or equal to 30.

[3 pts] Question 4g.) True or False with an explanation: In a dataset that has 5 items, the maximum number of size-2 frequent itemsets that can be extracted (assuming minsup > 0) is 20.

False. The size-2 frequent itemsets candidates are to choose 2 items from a total 5 items, so $C(5, 2) = 10$ as candidates and do pruning. So, we will have maximum 10 size-2 frequent itemsets.

[4 pts] Question 4h.) Draw the itemset lattice for the set of unique items I = {a, b, c}.


{}
a
b
c
ab
ac
bc
abc