CSIT5210 Data Mining and Knowledge Discovery (Fall Semester 2021)
Homework 1
Deadline: 5 Oct, 2021 3pm
(Please hand in during lecture.)
Full Mark: 100 Marks

Coupon Instructions:

1. You can use a coupon to waive any question you want and obtain full marks for this question.
2. You can waive at most one question in each assignment.
3. You can also answer the question you will waive. We will also mark it but will give full marks to this question.
4. The coupon is non-transferrable. That is, the coupon with a unique ID can be used only by the student who obtained it in class.
5. Please staple the coupon to the submitted assignment.
6. Please write down the question no. you want to waive on the coupon.

Q1 [20 Marks]

(a) We are given two customers, namely X and Y. The following shows 5 transactions for these two customers. Each transaction contains three kinds of information: (1) customer ID (e.g., X and Y), (2) the time that this transaction occurred, and (3) all the items involved in this transaction.

Customer X, time 1, items A, B, C
Customer Y, time 2, items A, F
Customer X, time 3, items D, E
Customer X, time 4, item G
Customer Y, time 5, items D, E, G

For example, the first transaction corresponds to that customer X bought item A, item B and item C at time 1, while the last transaction corresponds to that customer Y bought item D, item E and item G at time 5.

A sequence is defined to be a series of itemsets in form of $<S_1, S_2, S_3, \ldots, S_m>$ where S_i is an itemset for i = 1, 2, …, m. The above transactions can be transformed into two sequences as follows.

X: <{A, B, C}, {D, E}, {G}>
Y: <{A, F}, {D, E, G}>

After this transformation, each customer is associated with a sequence.

Given a sequence S in form of $<S_1, S_2, S_3, \ldots, S_m>$ and another sequence S' in form of $<S_1', S_2', S_3', \ldots, S_n'>$, S is said to be a subsequence of S' if $m \le n$ and there exist m integers, namely $i_1, i_2, \ldots, i_m$, such that (i) $1 \le i_1 < i_2 < \ldots < i_m \le n$, and (2) $S_j \subseteq S_{i_j}'$ for j = 1, 2, …, m. If S is a subsequence of S', then S' is defined to be a super-sequence of S.

The support of a sequence S is defined to be the total number of customers which sequences are super-sequences of S.

Given a positive integer k, a sequence in form of $<S_1, S_2, S_3, \ldots, S_m>$ is said to be a k-sequence if $\sum_{i=1}^{m} |S_i| = k$.

Can the Apriori algorithm be adapted to mining all k-sequences with support at least 2 where k = 2, 3, 4, …. ? If yes, please write down the proposed method using the concept of the Apriori algorithm and illustrate your algorithm with the above example. If no, please explain the reason.

(b) We want to study the same problem described in (a). However, the support of a sequence is defined in another way. Now, the support of a sequence S is re-defined to be the total number of all possible occurrences of S in all customers divided by the total number of customers which sequences are super-sequences of S. An occurrence of a sequence S (in form of $<S_1, S_2, S_3, \ldots, S_m>$) in a customer who has his/her sequence S' ($<S_1', S_2', S_3', \ldots, S_n'>$) corresponds to one possible set of m integers, namely $i_1, i_2, \ldots, i_m$, such that (i) $1 \leq i_1 < i_2 < \ldots < i_m \leq n$, and (2) $S_j \subseteq S_{i_j}'$ for j = 1, 2, …, m. Note that if S is a subsequence of S', it is possible that there are multiple possible sets of m integers (or multiple possible occurrences). For example, suppose that there is a sequence S' for a customer as <{A}, {B}, {B}, {C}>. Consider a sequence S = <{A}, {B}, {C}>. There are two possible occurrences of S in this customer (and the corresponding two possible sets of integers are {1, 2, 4} and {1, 3, 4}).

Can the Apriori algorithm be adapted to mining all k-sequences with support at least 2 where k = 2, 3, 4, …. with this new definition of support? If yes, please write down the proposed method using the concept of the Apriori algorithm and illustrate your algorithm with the above example. If no, please explain the reason.

Q2 [20 Marks]

Given a positive integer K, we denote S_K to be a set of K-itemsets with support at least 1.
Given a positive integer K and a positive integer l, we define a set $S_{K,l}$ which is a subset of S_K such that each K-itemset in $S_{K,l}$ has its support at least s_l where s_l is the l-th greatest value in the multi-set of the supports of all K-itemsets in S_K. For example, the second greatest value in a multi-set of {4, 4, 3, 2} is 4 while the second greatest value of another multi-set of {4, 3, 3, 2} is 3.

We are given six items, namely A, B, C, D, E and F.

Suppose l is fixed and is set to 2.
We want to find $S_{K,l}$ for K = 1, 2 and 3.

The following shows four transactions with six items. Each row corresponds to a transaction where 1 corresponds to a presence of an item and 0 corresponds to an absence.

A	B	C	D	E	F
0	0	1	1	0	0
0	1	0	0	1	1
1	0	1	1	0	0
1	0	1	1	0	0

(a) (i) What is $S_{1,2}$?
(ii) What is $S_{2,2}$?
(iii) What is $S_{3,2}$?
(b) Can algorithm FP-growth be adapted to finding $S_{1,2}$, $S_{2,2}$ and $S_{3,2}$. If yes, please write down how to adapt algorithm FP-growth and illustrate the adapted algorithm with the above example. If no, please explain the reason.
(c) There are two parameters of finding $S_{K,l}$. They are K and l. In the traditional problem of finding frequent itemsets, we need to provide only one parameter, a support threshold.

It seems that it is troublesome to set one more parameter in the problem of finding $S_{K,l}$ (compared with the traditional frequent itemset mining you learnt). What are the advantages of the problem of finding $S_{K,l}$ compared with the traditional problem?

Q3 [20 Marks]

(a) Consider the following eight two-dimensional data points:

x_1:(55, 50), x_2: (43, 50), x_3: (55, 52), x_4: (43, 54), x_5: (58, 53), x_6: (41, 47), x_7: (50, 41), x_8: (50, 70)

Consider algorithm k-means.

Please answer the following questions. You are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster) as the output of the algorithm. You can consider writing a program for this part but you are not required to submit the program.

(i) If k = 2 and the initial means are (50, 41) and (50, 70), what is the output of the algorithm?

(ii) If k = 2 and the initial means are (43, 50) and (55, 50), what is the output of the algorithm?

(iii) If k = 3 and the initial means are (50, 41), (50, 70) and (43, 50), what is the output of the algorithm?

(iv) If k = 4 and the initial means are (50, 41), (50, 70), (43, 50) and (55, 50), what is the output of the algorithm?

(b) In class, we learnt "original k-means clustering" and "sequential k-means clustering". We knew that "sequential k-means clustering" is used when we acquire data points over a period of time, and we want to start clustering before we have seen all data points. Assume that each of these two clustering methods start with the same initial means. Let O_1 be the output (i.e., the clustering result) of "original k-means clustering" and O_2 be the output (i.e., the clustering result) of "sequential k-means clustering". Is it always true that O_1 is equal to O_2? If yes, please elaborate it briefly. If no, please give an example showing that O_1 is not equal to O_2 and illustrate it.

Q4 [20 Marks]

Consider eight data points.

The following matrix shows the pairwise distances between any two points.

$$\begin{array}{cc} & \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} & \begin{pmatrix} 0 & & & & & & & \\ 11 & 0 & & & & & & \\ 5 & 13 & 0 & & & & & \\ 12 & 2 & 14 & 0 & & & & \\ 7 & 17 & 1 & 18 & 0 & & & \\ 13 & 4 & 15 & 5 & 20 & 0 & & \\ 9 & 15 & 12 & 16 & 15 & 19 & 0 & \\ 11 & 20 & 12 & 21 & 17 & 22 & 30 & 0 \end{pmatrix} \end{array}$$

(a) Please use the agglomerative approach to group these points with distance group average linkage. Draw the corresponding dendrogram for the clustering. You are required to specify the distance metric in the dendrogram.

(b) Suppose that we want to find 4 clusters. According to the dendrogram in (a), please state the 4 clusters. For each cluster, please include all data points involved.

(c) (i) What is the greatest possible number of data points in a cluster containing data 1 and data 5?

(ii) What is the smallest possible number of data points in a cluster containing data 1 and data 5?

(d) Suppose that data points satisfy the triangle inequality. That is, for any three data points, a, b and c, we have |a, b| + |b, c| ≥ |a, c| where |a, b| denotes the pairwise distance between a and b, and |b, c| and |a, c| have similar meanings. Does the triangle inequality enhance the agglomerative approach? If yes, please elaborate it. If no, please give the reason.

Q5 [20 Marks]

Consider the eight data points and the distance matrix in Q4.

Given a point p and a non-negative real number ε, the *ε-neighborhood* of point p, denoted by N(p), is the set of points q (including point p itself) such that the distance between p and q is within ε.
According to ε-neighborhood of point p, we classify all points into three types, namely *core points*, *border points* and *noise points*.

- Given a point p and a non-negative integer MinPts, if the size of N(p) is at least MinPts, then p is said to be a *core* point.
- Given a point p, p is said to be a *border point* if it is not a core point but N(p) contains at least one core point.
- Given a point p, p is said to be a *noise point* if it is neither a core point nor a border point.

Assume that we cluster the data points with the following principles. We call this clustering technique the *density-based clustering*.

- **Principle 1:** Each cluster contains at least one core point.
- **Principle 2:** Given any two core points p and q, if N(p) contains q (or N(q) contains p), then p and q are in the same cluster.
- **Principle 3:** Consider a border point p to be assigned to one of the clusters formed by Principle 1 and Principle 2. Suppose N(p) contains multiple core points. A border point p is assigned arbitrarily to one of the clusters containing these core points (formed by Principle 1 and Principle 2).
- **Principle 4:** All noise points do not belong to any clusters.

(a) Please write down a pseudo-code in order to perform clustering according to the above principles.

(b) Suppose ε = 10 and MinPts = 3. Please also illustrate your proposed algorithm with the above example in Q4.

(c) Please analyze the complexity of your proposed algorithm.

(d) Please state the differences between the density-based clustering and the hierarchical clustering.