Fall 2018 Data Mining Homework 3 Due: Nov 07, 2018

1. **[20 points]** For the following sets of two-dimensional points, (1) draw a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 1 matches the corresponding part of this question, e.g., Figure 1(a) goes with part (a).

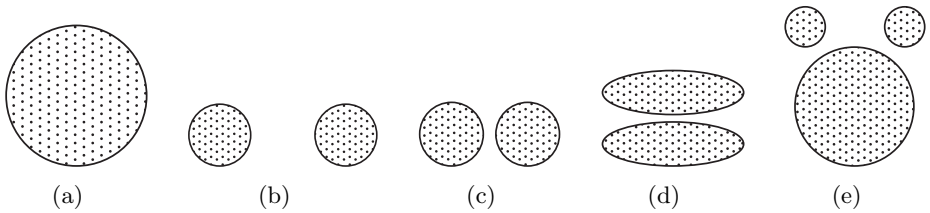


Figure 1: Diagrams for Question 1.

   (a) $K = 2$. Assuming that the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids? (Again, you don't need to provide exact centroid locations, just a qualitative description.)

   (b) $K = 3$. The distance between the edges of the circles is slightly greater than the radii of the circles.

   (c) $K = 3$. The distance between the edges of the circles is much less than the radii of the circles.

   (d) $K = 2$.

   (e) $K = 3$. Hint: Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.

2. **[20 points]** Explain and compare the advantages and disadvantages of DBSCAN over the K-means clustering algorithm.

3. **[20 points]** Traditional K-means has a number of limitations, such as sensitivity to outliers and difficulty in handling clusters of different sizes and densities, or with non-globular shapes. Comment on the ability of fuzzy c-means to handle these situations.

4. **[20 points]** Answer the following questions for the data set in Table 1.

   (a) What is the maximum number of association rules that can be extracted from this data set (including rules that have zero support)?

   (b) What is the maximum size of frequent itemsets that can be extracted (assuming $minsup > 0$)?

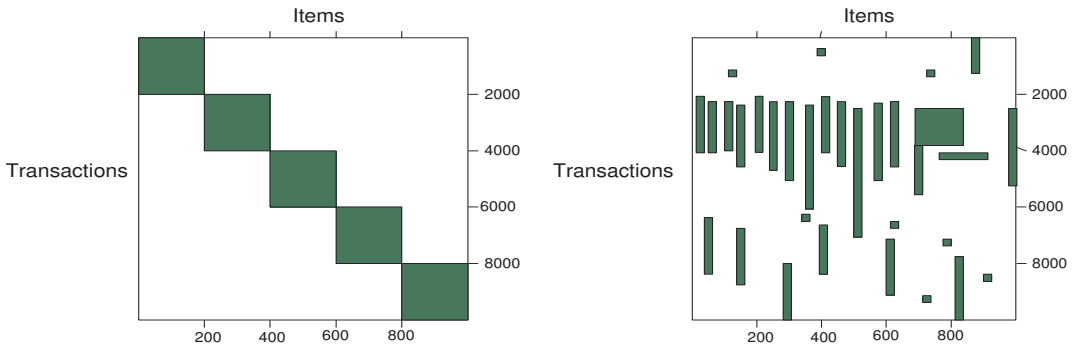
   (c) Calculate the maximum number of size-3 itemsets that can be derived from this data set.

Table 1: Data set of market-basket transactions.

Transaction ID	Items Bought
1	$\{A, B, D, E\}$
2	$\{B, C, D\}$
3	$\{A, B, D, E\}$
4	$\{A, C, D, E\}$
5	$\{B, C, D, E\}$
6	$\{B, D, E\}$
7	$\{C, D\}$
8	$\{A, B, C\}$
9	$\{A, D, E\}$
10	$\{B, D\}$

(d) Find an itemset (of size 2 or larger) that has the largest support.

(e) Find a pair of items, say x and y, such that the rules $\{x\} \longrightarrow \{y\}$ and $\{y\} \longrightarrow \{x\}$ have the same confidence.



(a) Synthetic data set 1.

(b) Synthetic data set 2.

Figure 2: Data sets for Question 5.

5. **[20 points]** Answer the following questions based on the data sets shown in Figure 2. Note that each data set contains 1000 items and 10000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent item sets with $minsup = 10\%$ (i.e, itemsets must be contain at least 1000 transactions).

   (a) Which data set will produce the most number of frequent item sets?

   (b) Assume that the minimum support threshold is equal to 10%. How many closed frequent itemsets will be discovered from data set 1?

   (c) Which data set will produce the longest frequent item set?

   (d) Which data set will produce frequent itemset with high support?

   (e) Which data set will produced the most number of closed frequent itemsets?

6. **Extra Credit [10 points]** For the definition of SNN similarity provided by Algorithm 9.10, the calculation of SNN distance does not take into account the weights of the edges connecting the two points to their shared neighbors. In other words, it might be more desirable to give higher similarity to two points that are connected to their shared neighbors by edges having higher weights, as compared to two points that are connected to their shared neighbors by edges having lower weights.

   (a) Describe how you might modify the definition of SNN similarity to give higher similarity to points whose shared neighbors are connected to them by edges having higher weights.

   (b) Discuss the advantages and disadvantages of such a modification.