Homework#3

ACS577 Knowledge Discovery and Data Mining, Summer II 2023

Due: July 22

- For your homework3 submission, prepare a single file, *YourLastName_FirstName*_ACS575_HW3.zip
- Organize the submission file with Part I and Part II and clearly number your answer with the question number.
- For each problem solving question, give your answer and also show the steps of computation to get the answer, if any.

Part I. Problem Solving

1. Suppose that the data mining task is to cluster the following eight data points with two attributes into three clusters, where the data points are

p1(2, 10), p2 (2, 5), p3(8, 4), p4 (5, 8), p5(7, 5), p6(6, 4), p7(1, 2), p8(4, 9)

Here we name each data point with p1, p2, p3, etc. for convenience.

We do the cluster analysis using the ***k-means*** algorithm with Euclidean distance (L_2-norm). Here k=3. Suppose we selected (2,10), (5,8) and (1,2) for three initial representative (centroid) points. Use the k-means algorithm for clustering

(1) Show the clustering result with the representative (i.e., centroid) of each cluster and its members (data points). Show the intermediate computing steps as well as the final result.

(2) Compute the SSE (Sum of the Squared Error) of each cluster and the total SSE of the clustering.

2. The table below is a proximity matrix with the similarity values of all pairs of data points p1, p2, p3, p4, and p5. Note that the matrix shows similarity values, not distance (dissimilarity) values.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

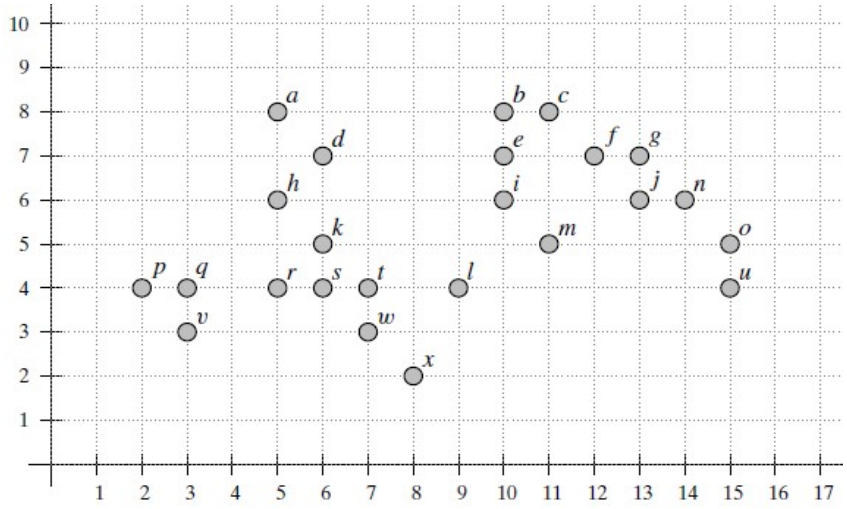
Table: Similarity matrix of data points

Suppose we perform ***a single link agglomerative hierarchical clustering***.

Show the result of the final clustering using its **dendogram**. Assume that smaller indexes are merged first in cases of tie-breaks.

The dendrogram should clearly show the order in which the points are merged.

3. The figure below shows a two-dimensional data set.



Suppose we perform a ***density-based clustering*** with the data set, assuming that we use the Euclidean distance between points, and that $\epsilon = 2$ and $minpts = 3$.

(1) List all the core points.

(2) Show the density-based clusters and the noise points if there are noise points.

~~**4.** Consider the following dada points of 1-dimensional for Q1 – Q4.~~ ← This question moves from HW#3 to HW#4 due to the class schedule.

~~{1, 3, 2, 1, 3, 2, 75, 1, 3, 2, 2, 1, 2, 3, 2, 1}~~

~~**(1)** Compute the Z-value of each data point. Which of these values can be considered the most extreme value?~~

~~**(2)** Determine the k nearest neighbor (k=1) of each data point. Which data point is the largest value of the nearest neighbor distance? Is this the correct outlier?~~

~~**(3)** Apply a k-means clustering algorithm (k=2) to the data set. Which data point lies furthest from the centroids of the clusters? Is this the correct outlier?~~

Part II. Hands-on practice

P1. The purpose of this practice is to get familiar with R scripts for clustering analysis. Follow "Ch 9. Clustering" tutorial from Zhao, "R and Data Mining". The tutorial document is attached.

Submit: (1) Your program codes, e.g., the R script codes in the tutorial document
(2) A proof to show the successful execution of your codes, e.g., screen shots on running, and the output.

Alternatively, you can show the same results using Python or any other programming language you prefer.

(P2 – P3) Download a customer data (named with **Wholesale customers data.csv**). For the detail of the dataset, refer to http://archive.ics.uci.edu/ml/datasets/Wholesale+customers# . The data includes the annual spending in monetary units on diverse product categories.

The goal of this practice problem is to segment the clients of a wholesale distributor based on their annual spending on diverse product categories.

P2. For the following tasks, you can use any programming language (R, Python, etc.) and APIs or a data mining tool.

(1) (Data Preparation and Data Transformation) The dataset has 8 attributes. Two attributes, CHANNEL and REGION, are nominal, and the others are continuous. For this study, we will use only numeric attributes except CHANNEL and REGION. The numerical attributes show different magnitude of values. To bring all the attributes to the same magnitude, standardize the attributes. Prepare a data file (named *trans_data.csv*) with these attribute values.

(2) (Cluster tendency) Compute the *Hopkins statistics* of the *trans_data.* Determine whether the warehouse customer data shows useful clustering tendencies using the Hopkins statistics value.

(3) (Optimal number of clusters) Before the actual clustering, identify the optimal number of clusters (k) for the data with the *trans_data* data using *Elbow method* or *Shilhouette method*. Write an optimal number of clusters you found. Briefly explain it.

P3. For the following tasks, you can use any programming language (R, Python, etc.) and APIs or a data mining tool.

(1) (Representative-based clustering) Perform the cluster analysis using *k-means* with k=6. For the input data, use the *trans_data* data from P2 (2). Report the clustering result. For that, you may prepare a result file (named *k-means-result.csv*) which is similar with the input data file but includes the name of cluster (e.g., Cluster1) a data point belongs to.

(2) (Hierarchal clustering) Perform the cluster analysis using *a complete link agglomerative hierarchical clustering algorithm*. For the input data, use the *trans_data* data. Show the dendogram for the clustering result.

(3) (Density-based clustering) Perform the cluster analysis using DBSCAN with $\epsilon = 0.5$ and $minpts = 15$. For the input data, use the *trans_data* data. Show the clustering result. For that, you may prepare a result file (named *DBSCAN-result.csv*) which is similar with the input data file but includes the name of cluster a data point belongs to.