CS 418: Introduction to Data Science
Homework Assignment 04
Fall 2019

Instructions

This assignment is due Tuesday, October 29, at 11:59PM (Central Time).

Answers for this assignment must be entered on *Blackboard*. You have 1 attempt to submit the assignment. This attempt is not timed. The assignment will close automatically at the due date and no late submissions will be accepted.

This assignment is individual. Offering or receiving any kind of unauthorized or unacknowledged assistance is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.

Part I: Clustering (100 pt.)

Problem 1 (10 pt.)

Match each statement with the corresponding clustering technique:

- **Returns a hierarchy of clusters.**
- **Returns a partial clustering.**
- **Computationally efficient in terms of both time and space.**
- **Computationally expensive in terms of both time and space.**
- **Has difficulty detecting clusters with non-spherical shapes.**
- **Can detect clusters with different shapes, but has difficulty detecting clusters with different densities.**

a. Hierarchical Clustering

b. K-Means Clustering

c. DBSCAN

Problem 2 (15 pt.)

Given the following distance matrix for observations $p1$, $p2$, $p3$, $p4$, and $p5$:

	$p1$	$p2$	$p3$	$p4$	$p5$
$p1$	0.00	0.10	0.41	0.55	0.35
$p2$	0.10	0.00	0.64	0.47	0.98
$p3$	0.41	0.64	0.00	0.44	0.85
$p4$	0.55	0.47	0.44	0.00	0.76
$p5$	0.35	0.98	0.85	0.76	0.00

Suppose that the observations are assigned to the following clusters: $C_1 = \{p1, p2\}$ and $C_2 = \{p3, p4, p5\}$. Answer the following questions:

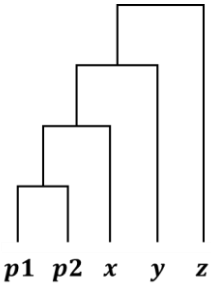
a. **(10 pts.) Compute the cohesion and the separation of the clusters.**

b. **(5 pts.) Compute the silhouette coefficient of observation $p1$.**

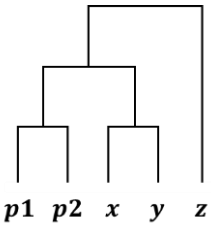
Problem 3 (25 pt.)

Given the distance matrix in Problem 2, answer the following questions.

a. **(10 pts.) Perform hierarchical clustering with the single linkage method on this dataset and complete the dendrogram below.**



b. **(10 pts.) Perform hierarchical clustering with the complete linkage method on this dataset and complete the dendrogram below.**



c. **(5 pts.) Suppose that you are performing hierarchical clustering with the average linkage method on this dataset. What is the proximity between clusters $\{p1, p2\}$ and $\{p3, p4\}$?**

Problem 4 (25 pt.)

Given the following dataset with observations $p1$, $p2$, $p3$, $p4$, and $p5$:

	X	Y
$p1$	1.5	2.0
$p2$	3.0	1.0
$p3$	3.5	2.5
$p4$	1.0	0.5
$p5$	2.5	2.0

Suppose that you cluster this dataset using K-means clustering with $K = 2$ and observations $p2$ and $p5$ as the initial centroids. Answer the following questions:

a. **(10 pts.) Assuming that observation $p2$ is assigned to cluster 1 and observation $p5$ is assigned to cluster 2, indicate the cluster of each observation.**

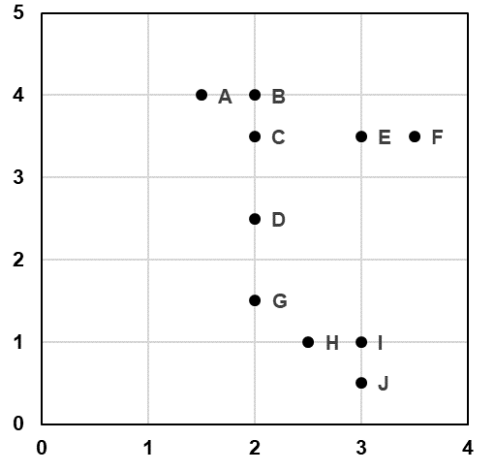
b. **(5 pts.) Recompute the centroids using the observations in each cluster. What are the new centroids?**

c. **(5 pts.) What is the sum of squared errors for these clusters (using the new centroids)?**

d. **(5 pts.) Do these clusters represent a stable solution? A solution is stable if the clusters do not change in the next iteration of the algorithm.**

Problem 5 (25 pt.)

Given the following dataset:



Answer the following questions:

a. **(10 pts.) Given $Eps = 1$ and $MinPts = 4$, label each observation as a core point, a border point, or a noise point.**

Suppose that you cluster this dataset using DBSCAN.

b. **(5 pts.) How many clusters are found by the algorithm?**

c. **(10 pts.) Assuming that observation A is assigned to cluster 1, indicate the cluster of each observation. Observations that do not belong to any cluster should be labeled as -1.**