CS245: Big Data Analytics (Fall 2023) October 28, 2023

Assignment 2: Clustering

Due Date: **November 13, 2023**

Instructions

- Submit your answer on Gradescope as a PDF file. Both typed and scanned handwritten answers are acceptable.
- Late submissions will not be accepted. Exceptionally, each student may request a *one-day* extension for one of the three assignments, provided they contact the instructor and TA *before the deadline*.
- Cite all resources used. Plagiarism will be reported.

Problems

Problem 1: KMeans (45 points)

Suppose we run the KMeans algorithm with $K = 2$ on the following set of 2D data points: {(9, 5),(5, 10),(10, 11),(4, 9),(14, 2),(4, 1),(9, 9),(3, 12),(19, 6),(17, 13),(13, 15)}, starting with initial centroids at $(10, 13)$ and $(7, 10)$. It is highly recommended to use a script to solve this question (*hint: for Python scripts, use package* ***sklearn***.)

1. **(5 points)** Predict the centroids of the clusters and classify each point into its respective cluster.
2. **(10 points)** The KMeans algorithm aims to minimize the Within-Cluster Sum of Squares (WCSS). Explain the reasoning behind this objective function. What are the potential limitations or drawbacks of this approach?
3. **(30 points)** Investigate the impact of initial centroid selection in the KMeans algorithm.
   (a) Analyze at least two common strategies, discussing their pros and cons.
   (b) Use of the two common strategies identified in part (a) on a dataset of your choice: (1) perform KMeans on two dimensions of this dataset and show the outcome of each cluster center initialization, and (2) discuss the observed effect of cluster center initialization and what information the clustering provides.

Problem 2: Hierarchical Clustering (15 points)

Run agglomerative hierarchical clustering on the following 1D dataset: {10, 2, 8, 12, 13, 15}.

1. **(5 points)** If we use single-linkage as the cluster distance, perform the hierarchical clustering step-by-step and show the resulting dendrogram.
2. **(5 points)** What clusters will we get if we cut the dendrogram at a height of 3?
3. **(5 points)** Discuss the differences between single-linkage and complete-linkage methods. How might the choice of linkage method impact the final clusters?

Problem 3: DBSCAN (40 points)

Consider the following 2D dataset:

Observation	Data Point
1	$(10, -20)$
2	$(5, 12)$
3	$(24, 22)$
4	$(-2, -7)$
5	$(-3, -10)$
6	$(12, -20)$
7	$(8, 17)$
8	$(9, 11)$
9	$(18, 20)$
10	$(8, 2)$
11	$(20, -10)$

1. **(10 points)** Apply the DBSCAN algorithm to the given dataset using Euclidean distance with the following parameters: $\varepsilon = 1$ and `MinPts` $= 2$. Identify the core points, border points, and noise points in the final clustering result.
2. **(5 points)** Discuss the effect of changing the ε value in DBSCAN. What happens when its value is increased or decreased? How does it affect the resulting clusters and noise points?
3. **(10 points)** Discuss the advantages and limitations of DBSCAN as compared to KMeans and hierarchical clustering.
4. **(15 points)** Read the paper *LOF: Identifying Density-based Local Outliers* by Breunig et al. (2000). Apply the Local Outlier Factor (LOF) algorithm for outlier detection to the same dataset. Discuss the differences you observe between the two methods in terms of identifying outliers. How do the parameter settings of LOF (e.g., number of neighbors) affect the points that are identified as outliers?