Janet, Arthur and Dr. Eick

COSC 3337
ProblemSet2 Task 5 Fall 2024
Clustering

Deadline; Task 5 will be due on November 13 end of the day in MS Teams
Last updated: November 5, 9a

Task 5: Clustering with K-Means and DBSCAN
Individual Task

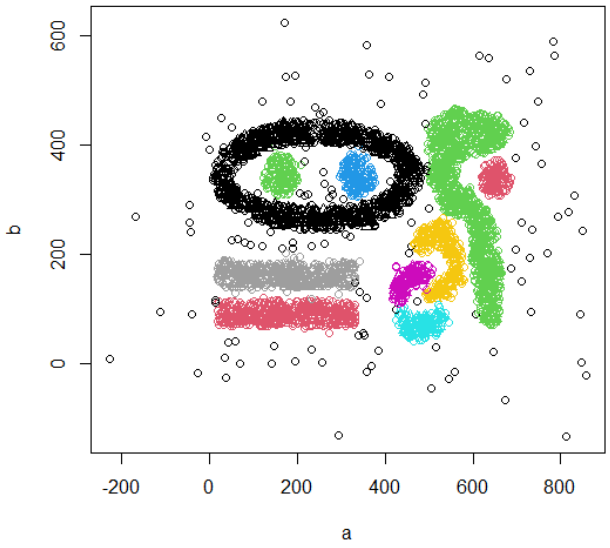

Fig. 1: Complex9_gn8 DBSCAN Clustering Result

Available Points: 34 points

Learning Objectives:
1. Learn to use popular clustering algorithms, namely K-means and DBSCAN
2. Learn how to summarize and interpret clustering results
3. Learn to write analysis and evaluation functions which operate on the top of clustering algorithms and clustering results
4. Learning how to interpret unsupervised data mining results

Datasets: In Task you will be using the Complex9 dataset with 8% Gaussian noise(Complex9_gn8), which is a 2D spatial dataset that can be found at https://www2.cs.uh.edu/~ceick/UDM/complex9_gn8.txt. The last attribute of the dataset denotes a class variable which should be ignored when clustering the data sets—however, the class variable will be used in the post analysis of the clusters generated by running K-means and DBSCAN.

Task 5 Subtasks:

a. Write an function[1] purity(a,b,outliers=FALSE) that computes the purity of a clustering result based on an apriori given set of class lables, where *a* gives the assignment of objects in O to clusters, and *b* is the "ground truth".

Purity is defined as follows:
Let O be a dataset
$X=\{C_1,\ldots,C_k\}$ be a clustering of O with $C_i \subseteq O$ (for i=1,…,k), $C_1 \cup \ldots \cup C_k \subseteq O$ and $C_i \cap C_j = \emptyset$ (for $i \neq j$)

PUR(X)= (number_of_majority_class_examples(X)/(total_number_examples_in_clusters(X))

If the used clustering algorithm supports outliers, outliers should be ignored in purity computations; if you use R-clustering algorithms, you can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent "true" clusters. If the parameter outliers is set to FALSE, the function just returns a floting point number of the observed purity, if parameter outliers is set to T the function returns a vector: (<purity>,<percentage_of_outliers); e.g. if the function returns (0.98, 0.2) this would indicate that the purity is 98%, but 20% of the objects in dataset O have been classified as outliers. **2p**

b. Develop a search procedure that looks for the "best" clustering by exploring different settings for the (MinPoints, epsilon) parameters of DBSCAN for the Complex9_gn8 dataset. The procedure should find hyperparameter settings which maximize the putiry of the obtained clustering, subject to the following constraints:
   i. There should be between 2 and 18 clusters
   ii. The percentage of outliers should be 10% or less.

The procedure returns the "best" DBSCAN clustering found and the accomplished purity as its result[2]; please limit the number of tested (MinPoints, epsilon)-pairs tested to 5000 in your implementation! Explain how your automated parameter selection method works and demonstrate your automated procedure using an example! Report, interpret and visualize the best clustering you found. *****
Alternatively, you could manually search for the "best" clustering and report and visualize the best clustering; however, we will lose some points if you do not create a search procedure. ** **20p**

[1] This function could be an R-function, a Python function or any other function. You might find some implementation of this function online; it is okay to use those implementations, as long as you acknowledge in your report what you use, and not all software you find on the internet is running properly.
[2] It should report the number of clusters obtained and the percentage of outliers as well.

c. Run K-means for k=9 and k=13 for the Complex9_gn8 dataset. Report the cluster centroids, the SSE, and compute the purity of the obtained two clustering results. Visualize the two clustering clustering result. Finally, discuss the clusters what where found and compare them with those found with DBSCAN, and assess which clustering algorithm did a better job with the dataset. **** **12p**

Deliverables for Task 5:

A. A Report[3] which contains all deliverables for the 3 subtasks of Task 5.
B. An Appendix which contains the software/code you developed as part of Task 5.

Task 3 Submission Guidelines:

1. Name your python/R files to **COSC3337F24-PS2T5-Firstname-Lastname.ipynb** or any other appropriate extension.
2. Name the pdf copy of your report **COSC3337F24-PS2T5-Report-Firstname-Lastname.pdf** carefully.
3. Create a folder and name it **COSC3337F24-PS2T5-Firstname-Lastname**.The folder should contain both python/R file and pdf copy of your report named correctly. Compress (zip) the folder and submit it to MS TEAMS.
4. Upload the zipped folder to the Assignment tab in MS Teams **before the deadline**.

[3] Single-spaced; please use an 11-point or 12-point font!