COMP 4331: Spring 2020
Assignment 3

Deadline: 23:59 4th of May, 2020

An amendment to the assignment has been made. All changes have been highlighted in red.

1 Submission Guidelines

- Assignments (including all attachments) should be emailed to comp4331spring2020@gmail.com before the deadline. All submissions after the deadline will incur a penalty of 20% of your marks for the first day (24 hours) and a further 10% for each hour it is late beyond the first day.

- Your final submission, titled A3_stuid.zip (where stuid is your itsc student id i.e. atzhou [**Note: this is not your student number**]) should be zipped file composed of two files:

  1. **A3_stuid_report.pdf** or **A3_stuid_report.docx**: Your report detailing your environment and experimental outcomes. This file **must** be either a pdf or docx file.

  2. **A3_stuid_code.zip**: A zipped file containing all source code used to complete this assignment. All code should be compilable (without the requirement to install additional libraries) and well-commented. Please provide a separate file for each technique as mentioned in section 2.

- By submitting any work, you are acknowledging that you are the sole contributor of the work unless specified and properly referenced. **All plagiarism will not be tolerated and incur a mark of 0.**

- Your work will be graded on correctness, efficiency and clarity of communication.

- To inquire about any questions you may have regarding the assignment, please email atzhou@connect.ust.hk or jfangak@connect.ust.hk

2 Clustering Implementation

Using the dataset (a3dataset.txt) found “Files → Assignment 3”, you are to implement the following clustering models. For each implementation you are to visualise your clusters on a graph.

Hint: To visualise clusters, plot each cluster separately onto a scatter plot [use matplotlib]. A basic guide as to how to do this can be found here: https://www.science-emergence.com/Articles/How-to-create-a-scatter-plot-with-several-colors-in-matplotlib-/

2.1 K-Means Clustering - 30 Marks

You are to implement K-Means Clustering in Python 2/3. Using your implementation, perform K-Means Clustering over the a3dataset for the following K values:

- K = 3
- K = 6
- K = 9

Please select a suitable method of initialising your means and describe why you selected it in your report.

Report a scatter plot (detailing the clusters) for each value of K as well as the Sum of Squared Error of the entire graph compared to each points respective mean. Based on these resources, report which value of K (if any) you believe to be the best of the available choices and why.

2.2 DBScan - 30 Marks

You are to implement DBScan in Python 2/3. Using your implementation, please find the clusters in the a3dataset with the parameters:

- $\epsilon = 5$ and MinPoints = 10.
- $\epsilon = 5$ and MinPoints = 4.
- $\epsilon = 1$ and MinPoints = 4

Report a scatter plot (detailing the clusters) for your DBScan Implementation. Colour all outliers uniformly as black. Based on these resources, report which parameters (if any) you believe to be the best of the available choices and why.

2.3 Report - 40 Marks

You are required to report the **running time** of each method. Also, the **environment** you use (System, CPU, RAM, etc) should be provided. Additionally, provide all resources required by the previous sections.

Discuss which method (K-Means or DBScan) you believe produced the best clusters for the given dataset. Additionally, explain why your selected method produced a superior final product to the alternatives. For the method you did not select, describe when it may perform equal or better to the method you did select.

2.4 Data Description

The dataset has 8000 data points, with two attributes each of doubles. Each cluster in the dataset does not contain a uniform shape. Additionally, note there is no header in the dataset file we provide.