MSC-BDT5002, Spring 2020
Knowledge Discovery and Data Mining
Assignment 4
Deadline: May 13th, 2020 11:59pm

Submission Guidelines

1. Assignments should be submitted to **mscbd5002spring20@gmail.com** as attachments.

2. Attachments should be named in the format of: **A4_itsc_stuid.zip** which includes
   - A4_itsc_stuid_report.pdf/.docx: Please put all your reports in this file. (Attachments should be original .pdf or .docx, NOT compressed)
   - A4_itsc_stuid_code.zip: The zip file contains all your source codes for the assignment.
   - A4_itsc_stuid_Q1_code: this is a folder that should contain all your source code for Q1.
   - A4_itsc_stuid_Q2_code: same as above.
3. TA will check your source code carefully, so your code MUST be runnable, your result MUST be reproducible.

4. For programming language, in principle, python is preferred.

5. Your grade will be based on the correctness, efficiency and clarity.

6. Please check carefully before submitting to avoid multiple submissions.

7. Submissions after the deadline or not following the rules above are **NOT** accepted.

8. The email for Q&A: hlicg@connect.ust.hk.

9. **Plagiarism and No Report/Code will lead to zero points**.

(Please read the guidelines carefully)

1 Fuzzy Clustering using EM (50 marks)

Given the training data *EM_Points.mat*, you should implement the Fuzzy Clustering using EM algorithm for clustering.

1.1 Data Description

The dataset contains 400 2D points totally with 2 clusters. Each point is in the format of [X-coordinate, Y-coordinate, label].

1.2 Implementation

You are required to implement Fuzzy Clustering using the EM algorithm.

1. You are **NOT** allowed to use any existing EM library. You need to implement it manually and submit your code.
2. Report the updated centers and SSE for the first two iterations. (If you set any hyper parameter when computing SSE, please write it clearly in the report.)
3. Report the final converged centers for each cluster.
4. In your report, draw the clustering results of your implemented algorithm and compare it with the original labels in the dataset. You need to discuss the result briefly.

Hint: For terminate condition, you can consider the change of parameters or the max iterations.

2 DBSCAN (50 marks)

Given the dataset *DBSCAN.mat* with 500 2D points, you should apply DBSCAN algorithm to cluster the dataset and find outliers as the following settings:

2.1 Parameter Setting

1. Set $\epsilon = 5$, Minpoints=5.
2. Set $\epsilon = 5$, Minpoints=10
3. Set $\epsilon = 10$, Minpoints=5.
4. Set $\epsilon = 10$, Minpoints=10.

2.2 Implementation

1. Draw a picture for your cluster results and outliers in each parameter setting in your report. For clearly, in each picture, the color of outliers should be **BLUE**.
2. Add a table to report how many clusters and outliers you find in each parameter setting in your report.

3. Discuss the results of different parameter settings, and report the best setting that you think and write your reason clearly.

4. Note that you are **NOT** allowed to use any existing DBSCAN library. You need to submit your code.

3 Note

One way to draw the clustering results is shown as below.

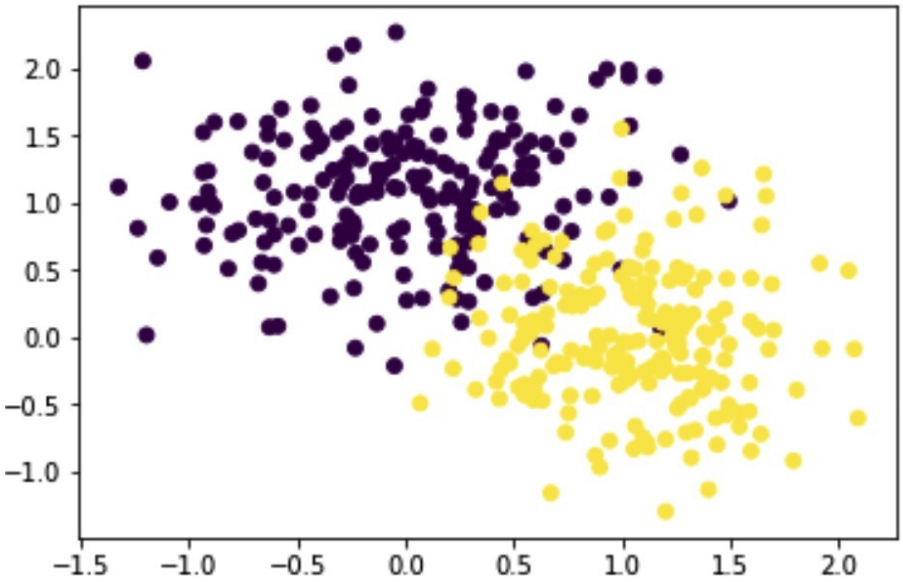

Figure 1: One example of drawing the clustering results where points are assigned with colors according to the corresponding clusters.