MSC-BDT5002, Fall 2021

Knowledge Discovery and Data Mining

Assignment 4

Deadline: Nov. 26th, 11:59pm, 2021

Submission Guidelines

1. Assignments should be submitted via Canvas.
2. All assignment files should be packed into one .zip file **named in the format of**: Ax_itsc_studentID.zip.
   e.g., for a student with **ITSC account**: zxuav, student ID: 20181234, the 4th assignment should be named as: A4_zxuav_20181234.zip.
3. You need to zip the following files together:
   1) A4_itsc_ studentID_answer**.pdf**: please put all your answers in this document including output answers for Q1, Q2 and Q3.
   2) A4_itsc_ studentID_Q1_code: this is a **folder** that should contain all your source code for Q1.
   3) A4_itsc_ studentID_Q2_code: this is a **folder** that should contain all your source code for Q2.
   2) A4_itsc_ studentID_Q3_code: this is a **folder** that should contain all your source code for Q3.
4. For programming language, in principle, **python** is preferred.
5. TA will check your source code carefully, so your code MUST be **runnable**, your result MUST be **reproducible**.
6. Keep your code clean and comment on it clearly. Missing the **necessary comments** will be deducted a certain score.
7. Your grade will be scored based on correctness and clarity.
8. Please check carefully before submitting to avoid multiple submissions.
9. Submissions after the deadline or not following the rules above are **NOT** accepted.
10. **Plagiarism will lead to zero points.**
11. If you have any questions, please **email to us with the subject "Inquiry for A4"**.

(Please read the guidelines carefully)

Q1. Fuzzy Clustering using EM (30 marks)

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

1.1 Data Description

The dataset contains 500 2D points totally with 3 clusters.

Format of the dataset *EM_Points.txt*
line 1-500: $x_i\ y_i\ label_i$ (i.e., the coordinates of each point and its 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 own code (but basic library is okay, e.g., Numpy).

2. Report the updated centers and SSE for the first two iterations (If you set any hyperparameter when computing SSE, please explain it clearly in the report).

3. Report the final converged centers for each cluster.

4. In your report, draw the clustering results of your algorithm and compare it with the original labels in the dataset. You need to discuss the result briefly.

Q2. DBSCAN (40 marks)

Given the dataset *DBSCAN_Points.txt* with 1212 points on a 2D Euclidean space, you should apply DBSCAN algorithm to cluster the dataset and find outliers as the following settings:

2.1 Parameter Setting

1. Set ϵ = 3, Minpoints = 5.
2. Set ϵ = 2.5, Minpoints = 20
3. Set ϵ = 2, Minpoints = 20.
4. Set ϵ = 2, Minpoints = 15.

Format of the dataset *DBSCAN_ Points.txt*
line 1-1212: x_i y_i (i.e., the coordinates of each point)

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 BLACK and the color of each cluster should be UNIQUE and different from BLACK.

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. To explain the reason, you may also need to draw a picture with your selected parameters.

4. Note that you are NOT allowed to use any existing DBSCAN library. You need to write your own code of the DBSCAN (but basic library is okay, e.g., Numpy).

Q3. Outlier Detection by Nested-Loop (30 marks)

Given the dataset *Nested_Points.txt* with 327 points on a 2D Euclidean space, you should apply the Nested-Loop algorithm to find the outliers.

3.1 Data Description

The dataset contains 327 points on a 2D Euclidean space.

Format of the dataset *Nested_Points.txt*
line 1-327: x_i y_i (i.e., the coordinates of each point)

3.2 Implementation

1. You are NOT allowed to use any existing Nested-Loop library. You need to write your own code of the Nested-Loop (but basic library is okay, e.g., Numpy).

2. In your report, please explain your parameter settings (e.g., distance threshold) and write down the coordinates of the outliers detected by your algorithm and your parameters. You also need to draw a picture of the result by marking the outliers by RED. You only need to report one group of proper parameter settings.