COMP 4331 Data Mining, Spring 2018

Assignment 3

Deadline: 23:59pm Apr 13th, 2018

1 Submission Guidelines

- Assignments should be sumbmitted to comp4331spring18@gmail.com as attachments.
- You need to zip the following two files together:
  - A3_itsc_stuid_report.pdf/.docx: Please put all your reports in this file. (Attachments should be original .pdf or .docx, NOT compressed)
  - A3_itsc_stuid_code.zip: The zip file contains all your source codes for this assignment.
- All attachments, including report and code, should be named in the format of: Ax_itsc_stuid.zip. E.g. for a student with itsc account: sdiaa, student id: 20171234, the 3rd assignment can be named as: A3_sdiaa_20171234.zip.
- Submissions after the deadline or not following the rules above are NOT accepted.
- Your grade will be based on the correctness, efficiency and clarity.
- The email for **Q&A**: zyanad@connect.ust.hk
- **Plagiarism will lead to zero mark.**

2 DBSCAN via Python (30 points)

Given the dataset (`https://github.com/ZengqiangYan/COMP4331/tree/master/Assignment3-Dataset/DBSCAN-Points.mat`), implement the DBSCAN algorithm for clustering.

2.1 Dataset Description

The dataset contains 500 2D points totally.

2.2 DBSCAN Implementation

You are required to implement the DBSCAN clustering algorithm:

- You are not allowed to use any existing DBSCAN method.
- Run your implemented DBSCAN on the dataset by setting $\epsilon = 0.12$ and $MinPts = 3$.
- Use the euclidean distance as measurement.
- Draw the clustering results and compare your results with the corresponding results generated by the DBSCAN model in *scikit-learn* library.
- Adjust the parameters 3-5 times, draw the corresponding results and analyze the influence of the parameters.

3 EM-GMM via Python (20 points)

Given the training data (`https://github.com/ZengqiangYan/COMP4331/tree/master/Assignment3-Dataset/GMM-Points.mat`), implement the GMM algorithm for clustering.

3.1 Dataset Description

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

3.2 EM-GMM Implementation

You are required to implement the GMM clustering method by using the EM algorithm (reference to slides No. 29-41):

- You are not allowed to use any existing EM-GMM method.
- Run your implemented GMM on the dataset.
- In your report, draw the clustering results of your implemented algorithm and compare with the original labels in the dataset.

Hint: For simplification, during the M step, you can directly calculate the mean and the std of points assigned to each cluster for updating.

4 Note

One exemplar way to draw the clustering results on the DBSCAN dataset is shown as below.

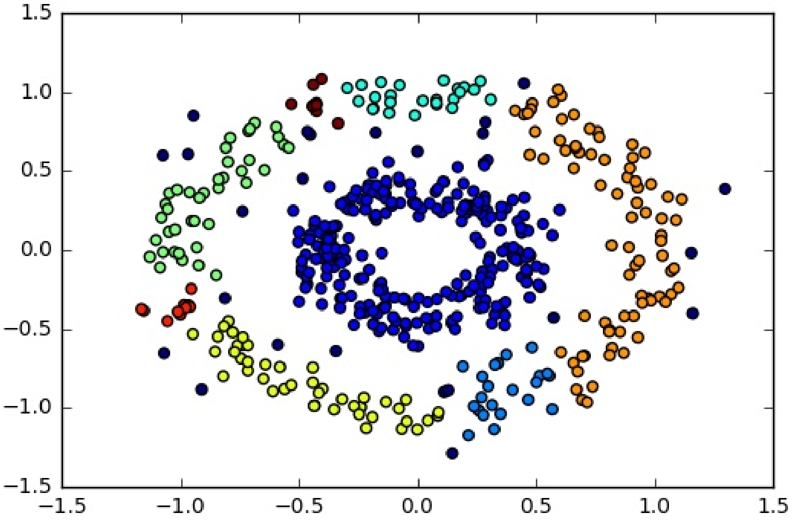

Figure 1: One exemplar way to draw the clustering results where points are assigned with colors according to the corresponding clusters.