DSCI 552: Programming Assignment 2

Part 1: Implementation [7 points]

Implement the K-means algorithm AND the Expectation Maximization algorithm for clustering using a Gaussian Mixture Model (GMM). Run your algorithms on the data file "clusters.txt" using K, the number of clusters, set to 3. Report the centroid of each cluster in K-means; and report the mean, amplitude and covariance matrix of each Gaussian in GMM. Compare the results of the two algorithms. The data file contains 150 2D points. Each row in the file contains the coordinates of a single point.

You can write your program in any programming language. However, you will have to implement the algorithms yourself instead of using high-level library functions. Please provide a description of the data structures you use, any code-level optimizations you perform, any challenges you face, and of course, the requested output.

Part 2: Software Familiarization [Optional - No Credit]

Do your own research and find out about library functions that offer good implementations of the two algorithms. Learn how to use them. Compare them against your implementations and suggest some ideas for how you can improve your code. Describe all this in your report.

Part 3: Applications [Optional - No Credit]

Do your own research and describe some interesting applications of the two algorithms.

Submission Guidelines

In your report, please include the names of all group members and mention their individual contributions. The maximum number of the members in a team is 2. The report should be a PDF file. Your submission should include the code as well as the report and is due before **02/16**, **11:59pm** in an archive in a zip, tar.gz or tar.xz format. Your source code should have a comment line that contains the names of all group members. Only one submission is required for each group by one of the group members. Please submit your homework assignment on **D2L** (do NOT email the homework to the instructor or the TA).