Faculty of Information Engineering & Technology
The Networking Department
Course: Analysis and Design of Algorithms [CSEN 707]
Dr. Tallal El Shabrawy
Eng. Nahla Afifi

German University in Cairo

Project

Deadline 18th of December 2023

Suppose you are working in the AI department for a company that is manufacturing T-shirts. Your colleagues from the data collection team gathered information about N persons.
For each person they got 5 features:

1. The Height
2. The Weight
3. The body mass index
4. The length between the shoulders
5. The length of the arms

They provided you with data in the form of an Nx5 matrix. Your manager tasked you with clustering the data into K clusters to determine K sizes for manufacturing T-shirts. For instance, if K=3, you should cluster the N samples into three sizes: small, medium, and large. If K=5, the data should be clustered into XS, S, M, L, and XL sizes. Here's what you need to do for Gaussian Mixture Model (GMM) clustering:

1. Normalize each feature so they are on the same scale. Note: A sequence of numbers is normalized by subtracting its mean and then dividing by the standard deviation.
2. Implement a Gaussian Mixture Model (GMM) clustering algorithm to cluster your normalized N samples into K groups.
3. Compute the complexity of the GMM algorithm.
4. Run the code for K=3 and K=5. Increase N and rerun the code several times for each K and different N. Plot the runtime curve versus N. Evaluate the complexity from the graph. Does it match your answer in 3?

After completing the previous milestone, you recognized redundancy in the features that can be reduced while preserving the same information. Therefore, you need to:

1. From the raw data, normalize each feature so they are on the same scale. Note: A sequence of numbers is normalized by subtracting its mean and then dividing by the standard deviation.
2. Implement Principal Component Analysis (PCA) to reduce the features from 5 dimensions to 2.
3. Implement a Gaussian Mixture Model (GMM) clustering algorithm to cluster your N samples into K groups using the 2 dimensions obtained from PCA.
4. Plot the samples on the x-y coordinates (2D features) colored according to their cluster.
5. Calculate the percentage of each cluster from the total N samples.
6. Run your code twice for K=3 and 5.

- Please note that you need to implement the GMM clustering algorithm by yourself not by calling already existing functions. You can use whatever language you prefer.
- You will be provided with a script to generate the data.
- The project can be accomplished in teams of max 4 members.
- Cheating cases will get a zero.
- The code and the results should be zipped in a file and sent to the following form along with the team members https://forms.gle/BoyyWv1y2g5gAEUn7 by maximum the 18th of December.
- Dataset can be found in the following link: https://drive.google.com/file/d/1xJzp5vsJCpIhDJoLzlBmc8yZjjme5BLH/view?usp=share_link
- Evaluation slots will be announced later.