Assignment on PCA and EM Algorithm

CSE472 (Machine Learning Sessional)

July 2023 Term at CSE, BUET

Introduction

In this assignment, you will implement the Principal Component Analysis (PCA) technique for reducing the dimensions of sample data. Besides, you will implement the Expectation-Maximization (EM) algorithm to estimate the Gaussian Mixture Model (GMM). PCA is a useful tool used in machine learning and data science to simplify the complexity of high dimensional data while retaining its trends and patterns. It does so by transforming the data into fewer dimensions, which act as summaries of features. On the other hand, GMM is a widely used unsupervised clustering method to group a set of sample data into clusters.

Task Summary

In this assignment, you will be given a dataset with *N* data points, each of which has *M* feature values. These data points are generated from a mixture of *K* unknown Gaussian distributions. And these data points are referred to as the **Gaussian Mixture Model** since they belong to a mixture of Gaussian distributions. Your task will be to reduce the dimension *M* of the data points into two dimensions using PCA **if the data points originally had more than two dimensions.** Then you will estimate the parameters of *K* unknown Gaussian distributions with the EM Algorithm.

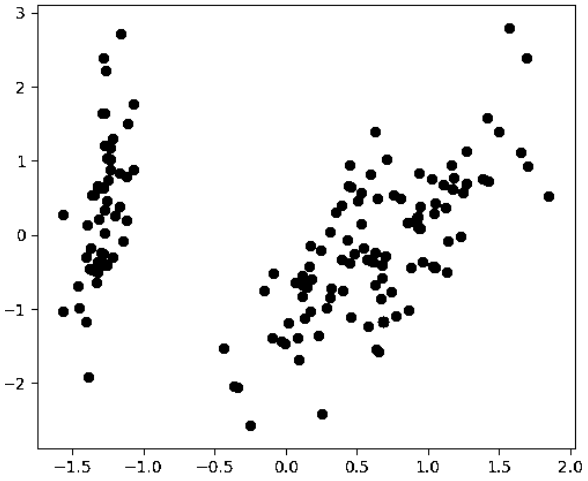
Dataset

You will find **here** some sample datasets that you will use in this assignment. Please note that *M* values in the feature vector of each data point are *comma* separated in sample dataset files.

Task 1: Principal Component Analysis

Take a sample dataset file as input. Assume that *D* is a matrix of shape *N x M*, where *N* denotes the number of sample data points in a dataset and *M* denotes the feature dimension. Here, *D* represents a particular dataset. Perform the PCA of *D* using the Singular Value Decomposition (SVD) technique that you learned and used in Assignment-1. **Keep in mind that you have to perform PCA only when *M* > 2.** You can and most likely should call the library functions for this purpose. Project the sample data points along the two most prominent principal axes. **You should never call the library function to perform the entire PCA at once.**

You now should have *N* sample data points, each having two dimensions. Plot the data points along two principal axes (if PCA is applied) or along two feature dimensions (if PCA is not applied). This plot may look like the following one. **You should store the plot as an image file.**



Task 2: GMM Estimation with Expectation-Maximization Algorithm

Estimate the unknown parameters of the Gaussian Mixture Model by applying the EM Algorithm as follows. **You are not allowed to use any library function to estimate GMM directly.**

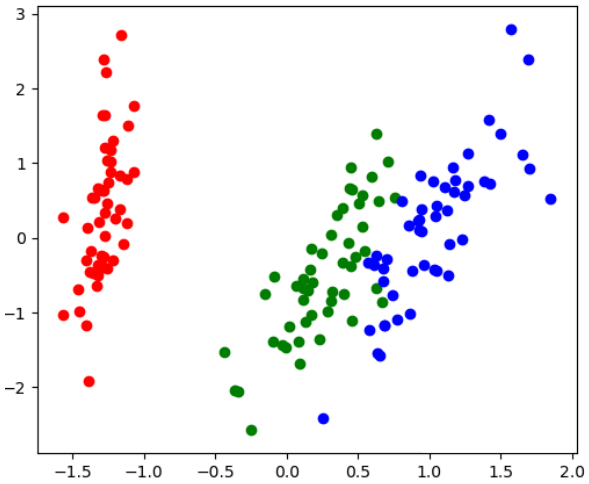
1. Assume a range for the value of *K* (for example, from 3 to 8) since the number of components (the number of distinct Gaussian distributions) in a GMM is usually unknown beforehand.

2. For each value of *K*, do the following.

    a. Apply the EM Algorithm to estimate the GMM. You should run the algorithm five times (preferably), each time with different random initialization of parameters.

    b. Pick and keep track of the best value of log-likelihood of the convergence from trials in previous step.

Now, plot the best value of convergence log-likelihood against the value of *K* in a 2D plot. Also, choose an appropriate value of *K* (denoted by *K'*) based on the convergence log-likelihood. And plot the estimated GMM for *K'* by showing sample data points and Gaussian distributions in a 2D plot. This plot may look like the following one. **You should store these two plots as image files.**



Kindly refer to the corresponding class materials for the details of this algorithm.

Bonus Task

This is an optional task that will carry bonus marks. You are highly encouraged to attempt this task, not for the sake of some additional marks but for a better understanding of how the EM Algorithm works in action. Make the necessary modifications to your EM Algorithm code so that a plot of the estimated GMM with sample data points and Gaussian distributions in a 2D plot is shown after each iteration (one E-step followed by one M-step). **You should not store the plots as image files.** The plot should be updated as the algorithm progresses (something similar to **this**).

Notes

- Take all user inputs (such as sample dataset file name and range of *K*, to name a few) from the command line so that you need not modify even a single line in your code to run the program for different sets of user inputs.
- You will be given a new sample dataset file during the assignment evaluation. Therefore, follow the preceding note to ease your evaluation.
- You are allowed to use Python libraries - including **NumPy**, **Pandas**, **Matplotlib**, and **Seaborn** - in this assignment.
- This assignment is probably going to be the last coding assignment of your undergraduate study at CSE, BUET. **Therefore, do not adopt any unfair means.** As a wise man once said, "All is well that ends well."

Submission Guiedeline

Move all your code into a single Python file. And submit your assignment, following the below folder structure, in Moodle **by 10 p.m. on January 27th, 2024 (Saturday). Keep in mind that it is a strict deadline.**

```
1805ABC <Folder>
|
|-- 1805ABC.py
```

Zip the `1805ABC` folder and submit the zipped `1805ABC.zip` file.

Food for Thought

- How is the Gaussian Mixture Model as a clustering method different from K-means Clustering and Hierarchical Clustering?
- How is the PCA as a dimensionality reduction method different from t-SNE and UMAP?