Comparison of Two Clustering Algorithms: EM-GMM vs. K-Means

Week 4 - Session 2

In this problem, you will be exploring two clustering methods, K-Means and GMM. Your task is to apply the algorithms to 1) the same synthetic dataset provided for comparison, 2) a real-world data set, MIMIC-III, and 3) your own data.

Data You are given a artificial dataset "EMGMM_Kmeans_dataset.npy", which has the same format (two columns of data X) with the dataset used in the workshop.

Model You are given "EMGMM_Kmeans_demo.py", which consists of the functions that can be used to draw a plot for K-Means and GMM result.

- *def plot_scatter()*: A Function for drawing a scatter plot for given dataset.
- *def plot_gmm()*: A function for drawing a scatter plot for the GMM result.
- *def plot_kmeans()*: A function for drawing a scatter plot for the K-Means result.

For the clustering methods, we use the *Scikit-learn* packages, *sklearn.cluster.KMeanas* and *sklearn.mixture.GaussianMixture*.

Report Start experimenting your models by implementing and running the followings. Include the result from each question in your report.

- Load the artificial dataset (EMGMM_Kmeans_dataset.npy") and report the shape of data. Then, draw a scatter plot for describing the distribution of data-points. How many clusters can you observe from the plot?
- Compare K-Means with GMM algorithms on the data with the number of clusters you observed from the previous question. Draw a scatter plot of the final cluster for each algorithm.
- Briefly describe how the clusters are different(or the same) and why, then choose one method.
- Explore K-Means and GMM using any two features from an imputed MIMIC-III data, which we provided in the last week: "mimic_shock.csv" and "mimic_nonshock.csv"

- Explore K-Means and GMM with your own data. If you do not have any data, you may artificially generate a data set, using a sample generator, "make_blobs", shown in the given code.

Demo: Jupyter Notebook file "EMGMM Kmeans demo.ipynb" is provided.