Expectation Maximization
with Gaussian Mixture Models

Week 4 - Session 1

EM-GMM

In this problem, you will practice implementing the EM algorithm for a Gaussian Mixture Model. Your task is to apply the provided Python code to the given dataset and conduct an experiment.

Data You are given a synthetic dataset(“W4S1_EMGMM_dataset.npy”) in a .npy format, which is a form of Python Numpy array. The two columns in the dataset indicate sample datapoints(X).

Model You will be using *Scikit-learn* package, *GaussianMixture*, to implement the GMM, where we can customize the parameters needed for the model, including the number of iterations.

- *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.

Your task is to understand and explore the code, focusing on the following action items.

Report Start experimenting your model by implementing and running the following components. Include the result of each part in your report.

- Load the dataset 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?
- Run your GMM algorithm with the number of clusters from the previous question. Draw a scatter plot for the final cluster. How many clusters are resulted from the model?
- Draw a scatter plot for iteration 7, 15, 25 and briefly describe how each Gaussian distribution changes with respect to its mean(center) point.

- Explore 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 GMM with your own data. If you do not have any data, you may artificially generate a data set, using “make_blobs”, shown in the given code.

Demo: A demo file “EMGMM_demo.ipynb” is provided.