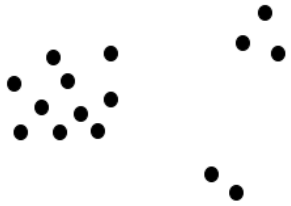
Week 4-Seminar 2

Q1: GMM Theory

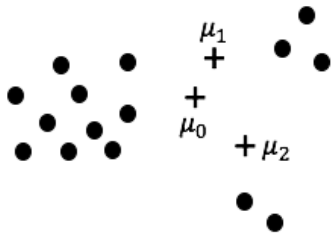
Consider the set of training data in the graph below, let's assume it contains three clusters. For GMM, the means and variances of three Gaussians are μ_0 and σ_0, μ_1 and σ_1, and μ_2 and σ_2, respectively. Additionally, we have π_0, π_1, π_2 to denote the mixture proportions of the three Gaussians (i.e., $p(x) = \pi_0 N(\mu_0, \sigma_0 I) + \pi_1 N(\mu_1, \sigma_1 I) + \pi_2 N(\mu_2, \sigma_2 I)$), where I is the identity matrix and $\pi_0 + \pi_1 + \pi_2 = 1$. We will also use θ to refer to the entire collection of parameters $(\mu_0, \mu_1, \mu_2, \sigma_0, \sigma_1, \sigma_2, \pi_0, \pi_1, \pi_2)$ defining the mixture model $p(x)$.



(a) (10 points) Would K-Means ($K = 3$) and our 3-cluster GMM trained using EM produce the same cluster centers (means) for this data set above? Justify your answer. (Answer without any justification will get zero point.)

Solution: Either algorithm will find the clusters just fine. But the difference lies in that k-means uses hard assignment of each point to a single cluster, whereas GMM uses soft assignment, where every point has non-zero (though possibly small) probability of being in each cluster. So in k-means, the means of the clusters are determined by an average of the points assigned to that cluster, but in GMM the means of each cluster are (differently) weighted averages of all points. This has the effect of skewing the center of the left cluster to the right, the center of the right cluster to the left, the center of the bottom cluster to the top.

(b) (10 points) In the following, we apply EM to train our 3-cluster GMM on the data below. The '+' points indicate the current means μ_0, μ_1, and μ_2 of the three Gaussians after the kth iteration of EM.



(b.1) On the figure, draw the directions in which μ_0, μ_1 and μ_2 will move in the next EM iteration.

Solution: μ_0 moves to the left, μ_1 moves to the right, and μ_2 moves to the bottom.

(b.2) Will the marginal likelihood of the data, $\prod_j P(x^j|\theta)$ increase or decrease on the next EM iteration? Explain your reasoning.

Solution: Increase. Each iteration of the EM algorithm increases to likelihood of the data, unless you happen to be exactly at a local optimum.

(b.3) Will the estimate of π_0 increase or decrease on the next EM iteration?Explain your reasoning.

Solution: Yes, π_0 will increase. π_0 is determined by adding the probabilities of all points that they are in cluster 1. When μ_0 moves to left, the probabilities of points in cluster 1 will increase but cluster 2 and 3 will decrease. Because more points are becoming closer to μ_0 and the increase is faster than decrease.

EM & GMM Programming

In this problem, you will be exploring two clustering methods, K-Means and GMM. Your task is to implement the algorithms and apply them to the same synthetic dataset provided for comparison. Read the following description and start implementing your models based on the instruction.

Data You are given a artificial dataset "w3_EMGMM_HW_dataset.npy", which has the same format (two columns of data X) with the dataset used in the workshop.

Model You are given "w3_EMGMM_HW_starter.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, you may 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 ("w3_EMGMM_HW_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?
- Implement and run both K-Means and 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.

Solution: Jupyter Notebook file "w3_EMGMM_HW_solution.ipynb" is provided.