# Report

Inyeob Kim

[Initialization]

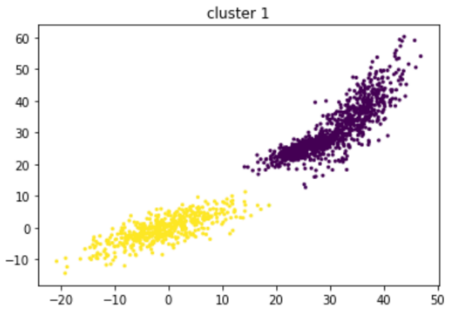
Initialization of mean, sigma and phi can affect the result of training, since it can be stuck at local optima instead of global optima. Convergence in Gaussian Mixture Model is not guaranteed to be to global optima, so we may have to start from several initializations and use the loglikelihood to select the best parameters. First, k number of random data points are chosen for the means. Then, with these random means, covariance matrices are created. Since there are k number of clusters, equal probabilities of phi = 1 / k are allocated to each cluster (If there are 3 clusters, each phi will be 1/3). All the means, covariance matrices, and phis are stored in separate lists.

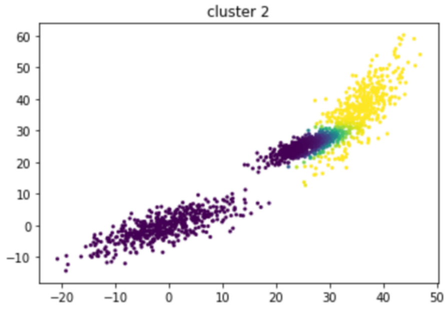
[GMM vs Single Gaussian]

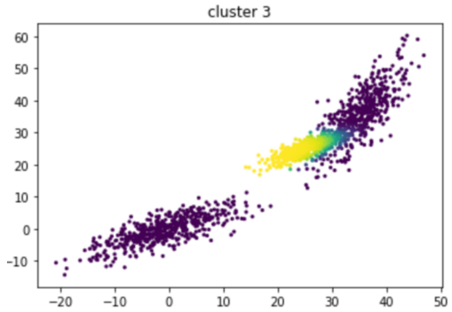
One of the advantages of Gaussian mixture model is that it accommodates mixed membership. A data point in GMM belongs to each cluster with different degrees. The degree is based on the probability of the point being generated from each cluster's Gaussian distribution, with each unique means and covariances. Some points may have attributes of more than one clusters, so in the case of using Single Gaussian will not be able represent points well enough. Even with random initialization of mean, sigma, and phi, Single Gaussian eventually yields same updated parameters eventually.

[2D scatter plot for Gaussian clustering with corresponding p(x) value]

Number of clusters = 3, iteration = 100







[Strengths and limitations of using GMM]

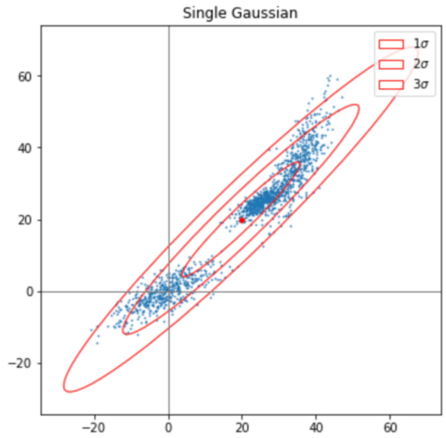
We had a problem clustering with K-means algorithm when clusters overlap each other. Then, it is very hard to tell which cluster is the right one, and we may have to try to remain uncertain. K-means is a hard clustering, which means a data point cannot belong to multiple clusters. One of the advantages of using Gaussian mixture model instead of K-means is that it is a soft clustering, which means that a data point can belong to more than one clusters. It is represented with probabilities. We calculated probabilities using the equation below.

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{\sum_{k=1}^{K} p(x|C_k)p(C_k)}$$

If the responsibility is small, it means that cluster k is not a good explanation for a data point. If the responsibility is close to 1, it means cluster k is a good explanation for the data point. Likewise, a data point can belong to multiple clusters in GMM. Another good advantage is that GMM does not assume clusters to be of any geometry, and it works well with non-linear geometric distributions. Additionally, it does not bias the cluster sizes to have specific structures. However, there are some limitations of GMM. One of the disadvantages of using GMM is that it uses all the components it has access to. This means if the dimensionality of data is high, the initialization of clusters will be very difficult. Furthermore, when there is a mixed data set, even if you cluster the data, you do not know anything about the data itself (It is just clustering). Also, it is challenging to interpret the result.

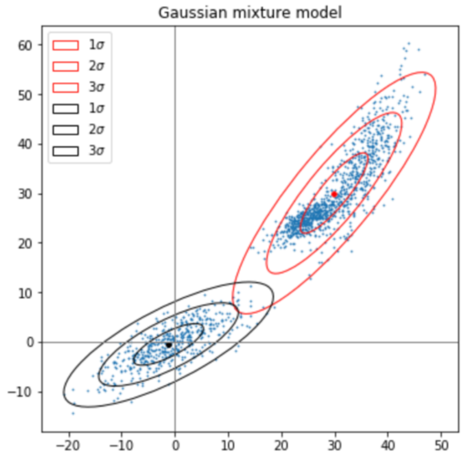
[GMM clustering with different number of Gaussians]

1) Number of clusters = 1, iteration = 50



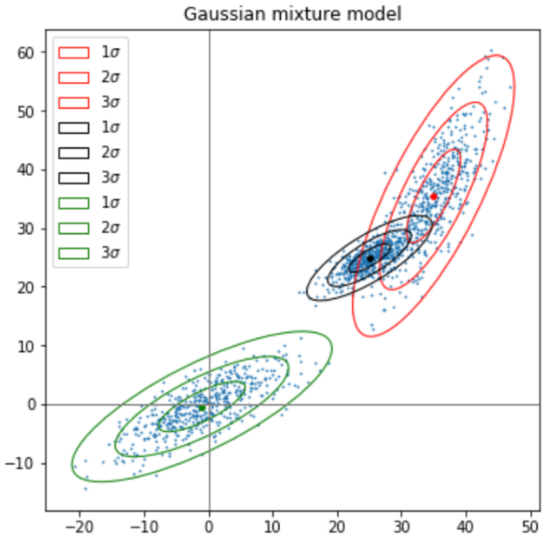
It simply uses mean, sigma and phi value of the entire data set.

2) Number of clusters = 2, iteration = 50



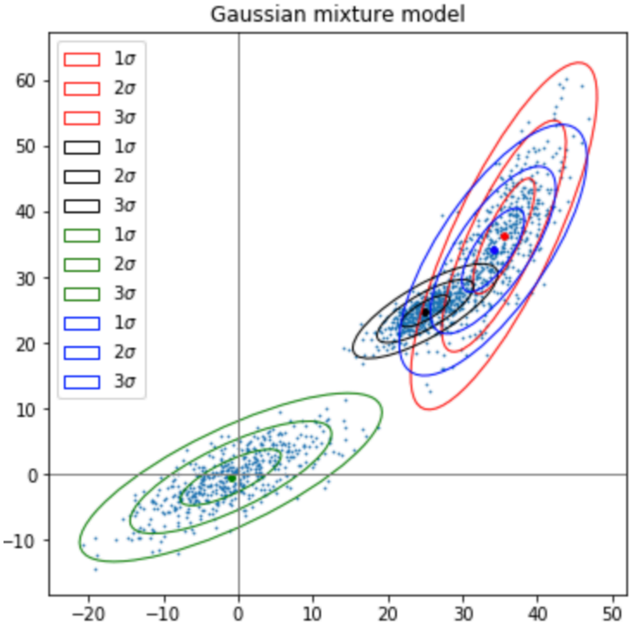
It seems the two clusters are fit where the data is most gathered.

3) Number of clusters = 3, iteration = 50



Data points fit into 3 clusters the best, since the data is a mixture of three distributions.

4) Number of clusters = 4, Iteration = 100



The 4th cluster overlaps with the other one, which seems that 4 is not a reasonable cluster number for the data set.