# 1 Lab 6

Name: Pin Wang

Student ID Number: 30845122

Email Address: pw2a19@soton.ac.uk

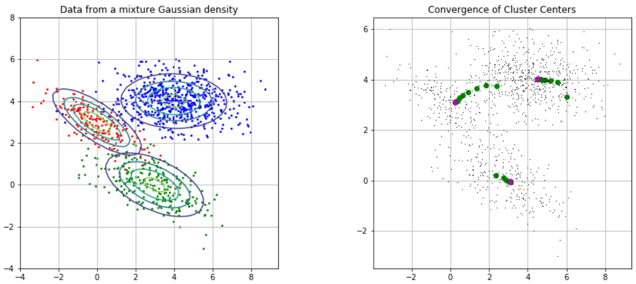
## 1.1 Lab 1-5

I have completed all the five labs and uploaded reports. And for where ”Incomplete” has been noted in the feedback provided (Labs 1-3), I have re-visited the tasks and completed them.

## 1.2 K-Means Clustering

### 1.2.1 K-means Algorithm Implementation

Sample data from a mixture Gaussian density and implement K means clustering algorithm. Its contours on the probability density I have used and and cluster centres from Kmeans clustering are shown in Figure 1.



(a)

(b)

Figure 1: Data from a mixture Gaussian density (a), and cluster centres from Kmeans clustering (b). Initial guess of cluster centres and their different estimates during iterations are marked in green and the converged answer in magenta.

As can be seen from Figure 1, the classification of the mixed Gaussian density is obvious, so the classification effect achieved by the K-means is accurate, and the clustering centers of the k-means are close to the centers of the three Gaussian density contours.

### 1.2.2 Comparison with Sklearn

The k-means algorithm in the sklearn library is used to cluster the mixed Gaussian density. The clustering result is shown in Figure 2.

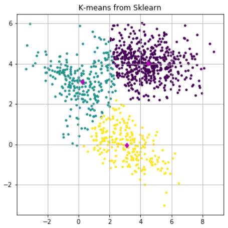


Figure 2: Clustering Results Using Sklearn

It can be seen that its clustering center is basically the same as the clustering center of the algorithm I implemented, which proves that the k-means algorithm I implemented has a good effect.

### 1.2.3 Initial Clusters Center and the Choice of K

The effect of the initial cluster center and k value selection on the clustering results is analyzed through a failed clustering case. The clustering results are shown in Figure 3.

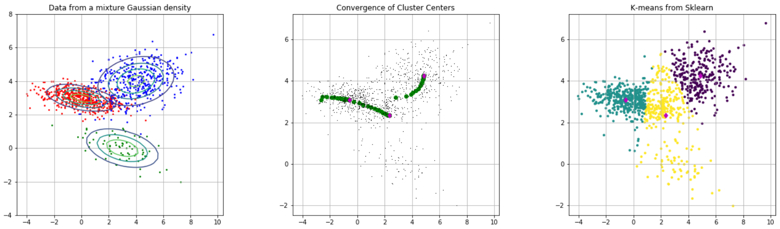


Figure 3: A Failed Clustering Case

In this case, the random Gaussian model appears with two larger clusters and one smaller cluster. Due to the random selection of the initial centroid, the problem of local optimal solution is caused, that is, the smaller clusters are ignored, and the clustering is performed in two larger clusters. Due to the instability of the K-Means algorithm, the initial centroid selection is different and the results are different. So to solve the local optimal method, one can run the algorithm multiple times and choose the group with the smallest SSE value as the final solution. This method solves the problem of randomly selecting the initial centroid by multiple runs and trials.

For the choice of k values, we can observe the results of Mean Distortions at different k values. For the case in 1.2.1, we draw a line chart of the mean distortion with the k value to determine the choice of the k value. The results are shown in Figure 4.

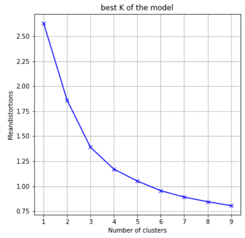


Figure 4: the Mean Distortion with the K Value

We can see that the larger the number of clusters, the smaller the mean distortion. But obviously the K value is not as large as possible. The larger the K value, the lower the information amount of the clustering result. In this data set, we can see from the above figure that, before K = 3, the mean distortion decline is more obvious, and after K = 3, the mean distortion decline slows down. It seems that K = 3 is a better choice.

### 1.2.4 Clustering Iris Dataset using K-means

I selected the iris dataset in the UCI database for k-means clustering. First process the iris dataset: delete the Species feature in column 5. The boxcox transform is then used to non-linearly scale the data to Increase its correlation. Next, delete the outliers on lines 14, 32, 59, and finally use PCA to reduce the data to 2-dimensional data.

Afterwards, we draw a graph of the mean distortion versus k value to select the k value. As shown in Figure 5.

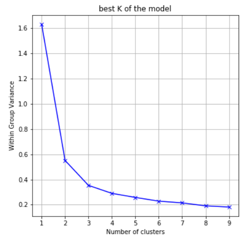


Figure 5: the Mean Distortion with the K Value

So I choose k = 3 as the number of clusters(This dataset contains three categories: Iris-setosa, Iris-versicolor, and Iris-virginica). The clustering results are shown in Figure 6.

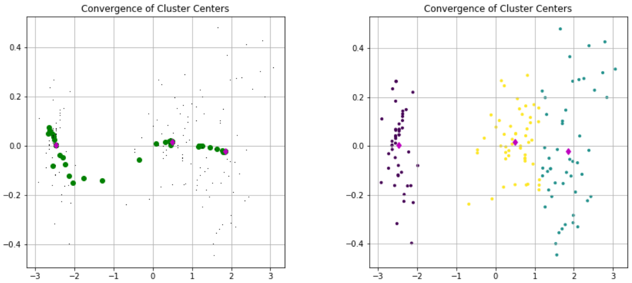


Figure 6: Clustering Iris Dataset using K-means

I use the silhouette coefficient to evaluate the quality of the clustering. The average silhouette coefficient of kmeans clustering on the iris dataset is 0.6498, which indicates that the clustering effect is good.