Assignment 3: HW8-9

Zhengyuan Zhu zhengyuan.zhu@mavs.uta.edu
1001778274

1. HW8

Explain why the K-means objective function decreases in each of the two steps in K-mean algorithm: (a) re-assign every data points to their nearest cluster centroids. (b) Given the grouping (or clustering), re-computer the cluster centroids.

1.1 Background

K-means objective function is to minimize the average Squared Euclidean distance of data points from their cluster centers where a cluster center is defines as the mean of centroid μ of the documents in a cluster w:

$$\hat{\mu}(w) = \frac{1}{|w|} \sum_{\hat{x}] \in w} \hat{x} \tag{1}$$

The objective function measures how well the centroid represent the members of their cluster, one avaiable function is **residual sum of squares**(RSS)

$$RSS_k = \sum_{\hat{x}] \in w_k} |\hat{x} - \hat{\mu}(w_k)|^2 \tag{2}$$

$$RSS = \sum_{k=1}^{K} RSS_k \tag{3}$$

1.2 solution

Through reassigning documents to the cluster with the closest centroid, the objective function decreases since **the distance that vector contributes to objective function decrease**.

1.3 solution

And the objective decrease in the recompute the cluster centroid step because **the new centroid is the vector for which objective reaches its minimum**.

2. HW9

(A) Generate Three Gaussian distributions, each with 100 data points in 2 dimensions, with centers at (3,3), (-3,3), and (0,-3) and standard deviation $\sigma = 2$. Draw them in a

Figure. Set K=3, do K-means clustering. Show the clustering results in the same Figure and compute the converged K-mean loss. Repeat this 5 times. Submit the 5 figures and losses, each represent the result of each K -means clustering. (B) Everything are same as (A), but with $\sigma = 4$. Submit the 5 figures and losses.

2.1 Code template

In the code template, I use "numpy" to generate three gaussian distribution and concatenate them as a dataset. The use "sklearn" to initiate a Kmeans class to fit the dataset. The centers and losses are built-in parameters in the Kmeans object. The figures and losses are shown below.

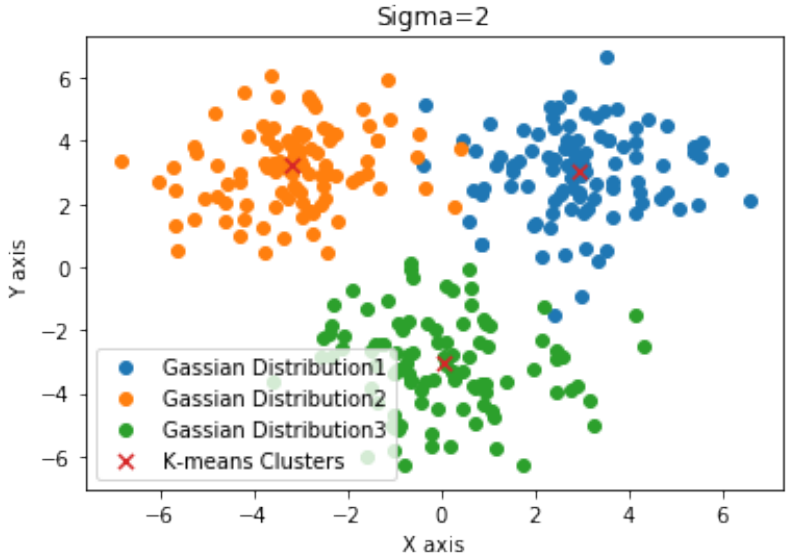

Figure 1: 1st running Kmeans when $\sigma = 2$, *accumulate loss* = 1119.87

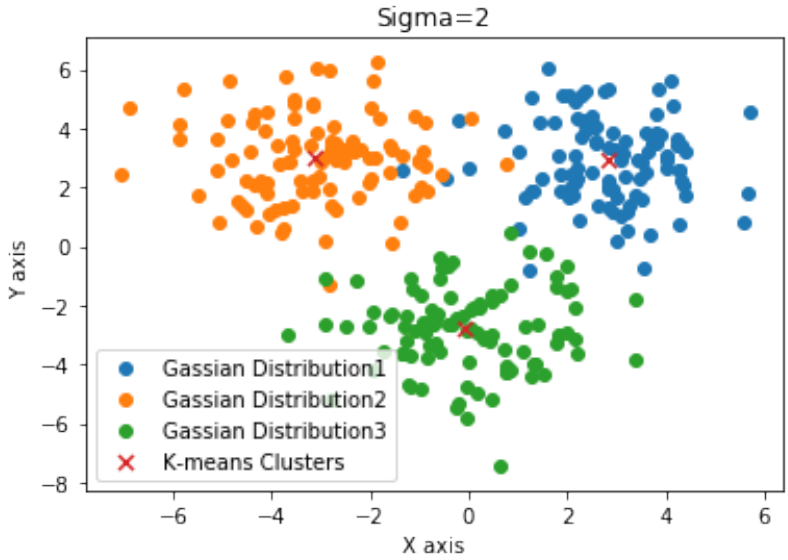

Figure 2: 2nd running Kmeans when $\sigma = 2$, *accumulate loss* = 1098.09

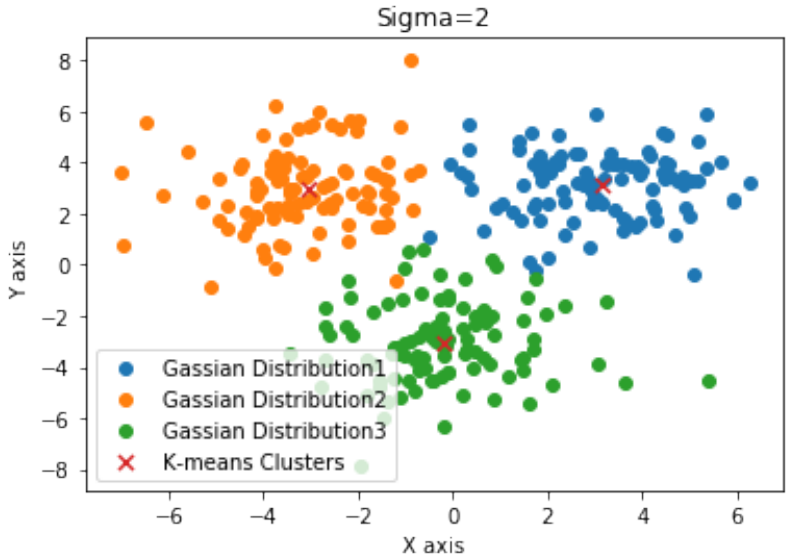

Figure 3: 3rd running Kmeans when $\sigma = 2$, *accumulate loss* = 1250.47

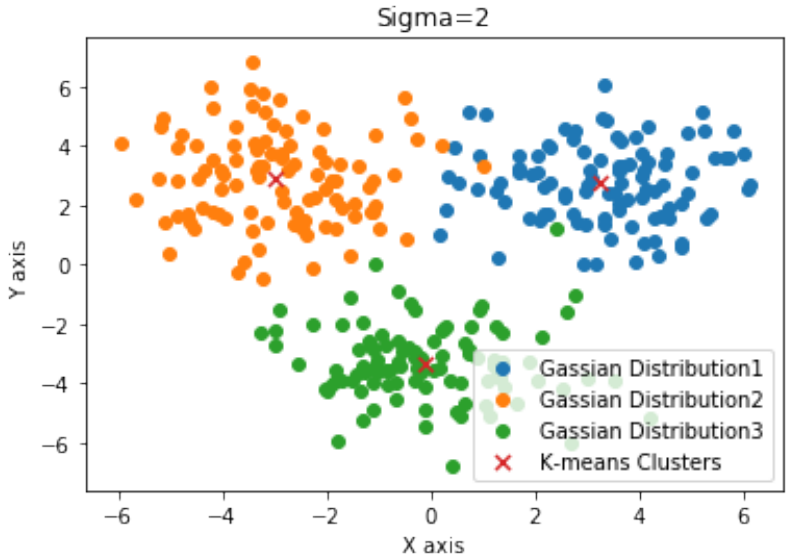

Figure 4: 4th running Kmeans when $\sigma = 2$, *accumulate loss* = 1154.10

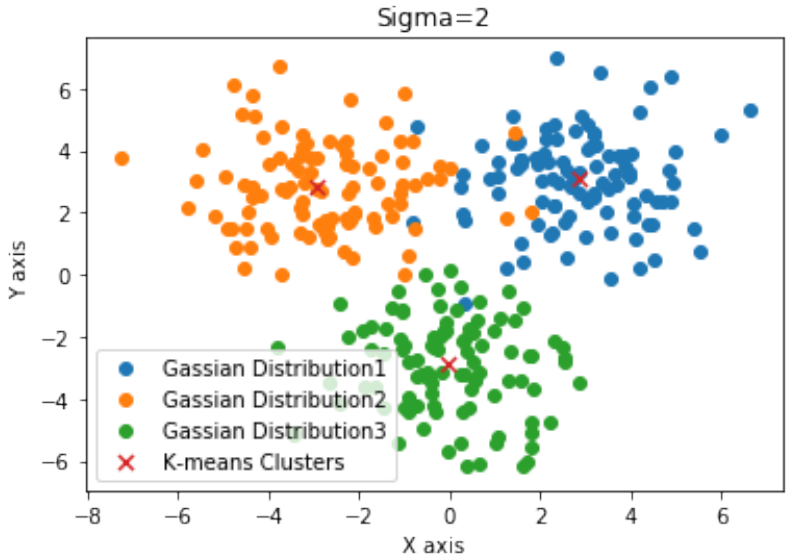

Figure 5: 5th running Kmeans when $\sigma = 2$, *accumulate loss* = 1201.70

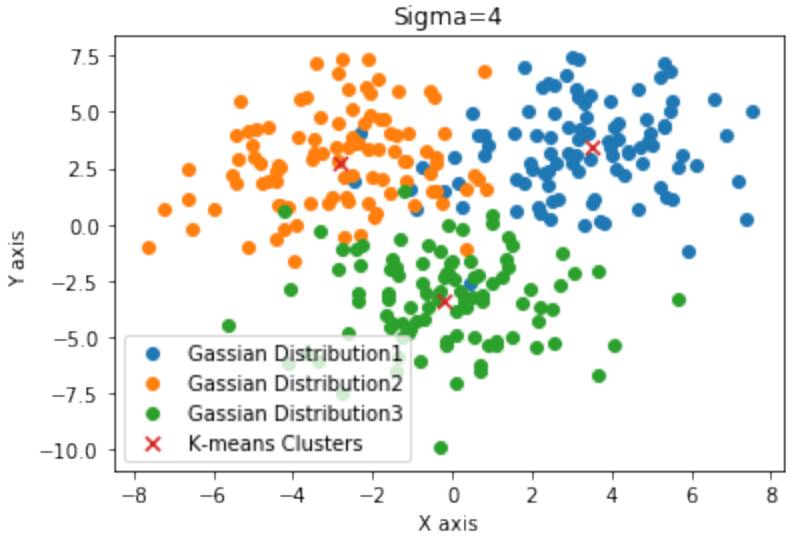

Figure 6: 1st running Kmeans when $\sigma = 4$, *accumulate loss* = 1939.61

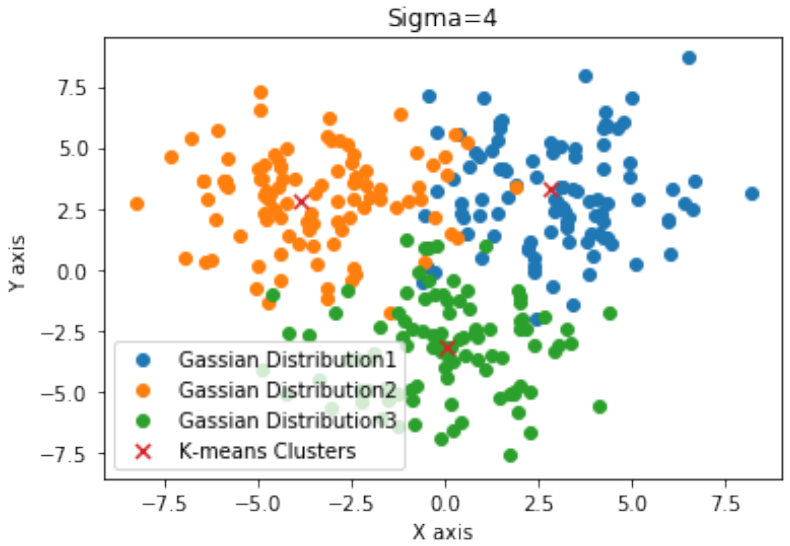

Figure 7: 2nd running Kmeans when $\sigma = 4$, *accumulate loss* = 2058.36

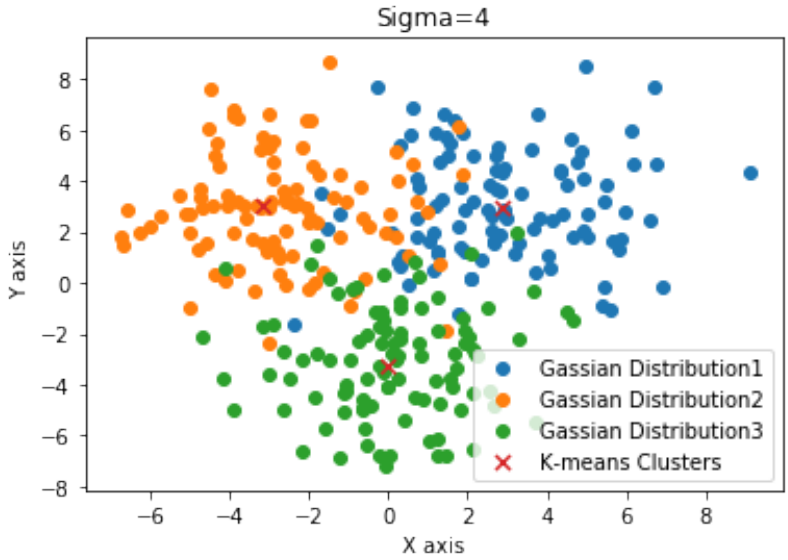

Figure 8: 3rd running Kmeans when $\sigma = 4$, *accumulate loss* = 2056.30

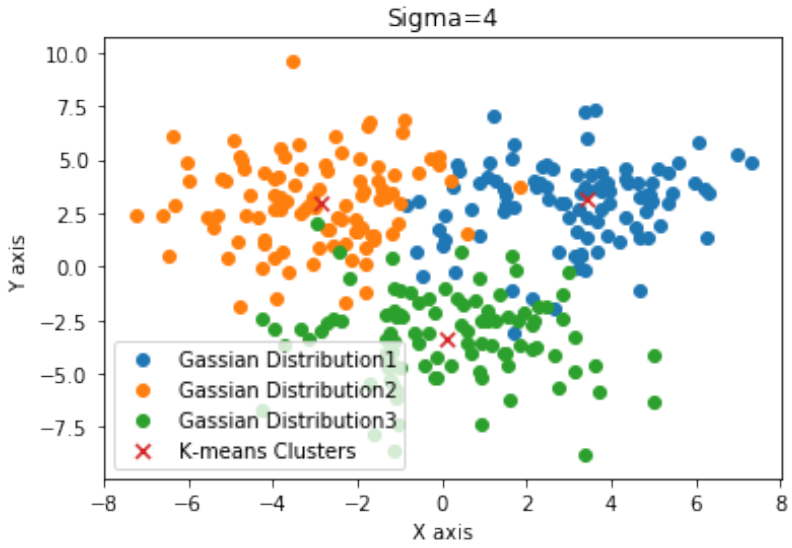

Figure 9: 4th running Kmeans when $\sigma = 4$, *accumulate loss* = 1973.79

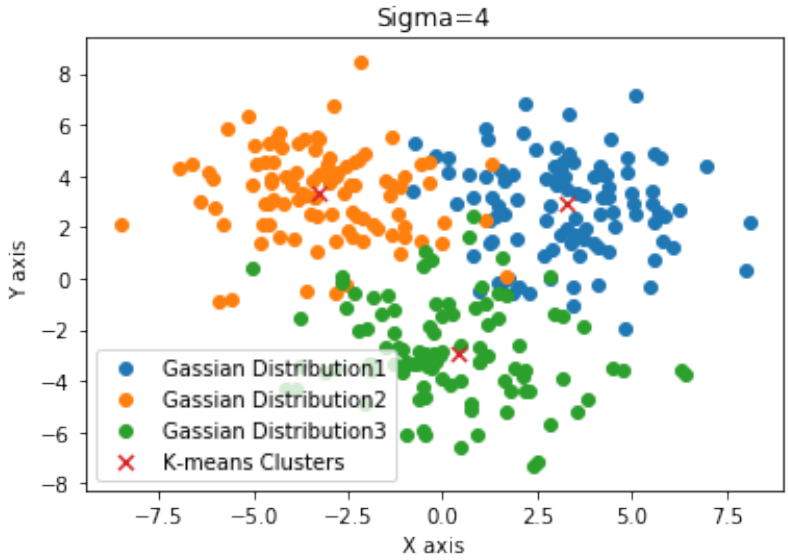

Figure 10: 5th running Kmeans when $\sigma = 4$, *accumulate loss* = 1916.41

References

- https://en.wikipedia.org/wiki/Multivariate_normal_distribution
- https://statweb.stanford.edu/~candes/teaching/acm118/Handouts/covariance.pdf
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html