COMP6245 : Lab 6

Thanakorn Panyapiang(tp2n19@soton.ac.uk)

1 Lab 1 - 5

I have completed all the five assignments, submitted reports and have taken on board any feedback provided.

2 K-Means Clustering

The sample data below is drawn from 3 Gaussian densities where all distributions have approximately the same amount of data.

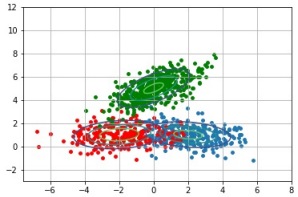


Figure 1: Data from a mixture of Gaussian density

Running 2 versions K-Means clustering(self-implementation and sklearn) on the dataset in Figure 1 gives the result as shown in Figure 2. Based on the prior knowledge about the data, the cluster centroids produced by both versions are consistent with the centers all 3 densities.

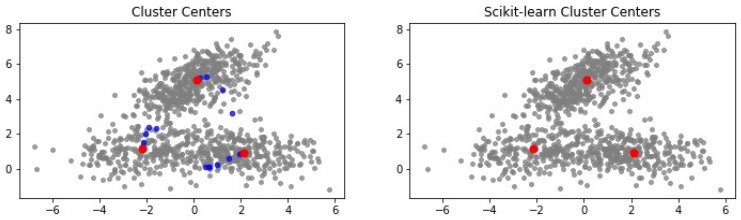


Figure 2

However, when running both versions of K-Means algorithm on the imbalance dataset displayed in Figure 3, both versions run into the same issue. From the result in Figure 4. , it can be noticed that K-Means locates two centroids on the majority class and one of the minor classes is not considered as a cluster.

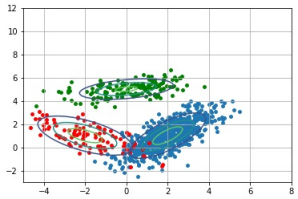


Figure 3: Imbalance dataset where one density has significantly more data than others

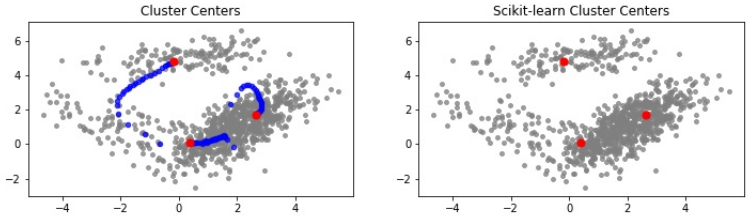


Figure 4: Results of K-Means clustering on the data in Figure 3

3 K-Means Parameters Impact

This section of the report will analyze the impact of two parameters of K-Means algorithm : K and *the initial centroids*.

3.1 Choice of K

The result of running K-Means clustering(self-implemented version) on the dataset in Figure 1. with various choices of K is shown in the figure below. What can be observed from the result is that as the value of K increases the size of cluster and the inter-cluster distance become smaller. However, too high value of K could also makes the cluster lack of generalization. For instance, on K = 7, the algorithm divides one density on the top into two groups and two densities below into four clusters.

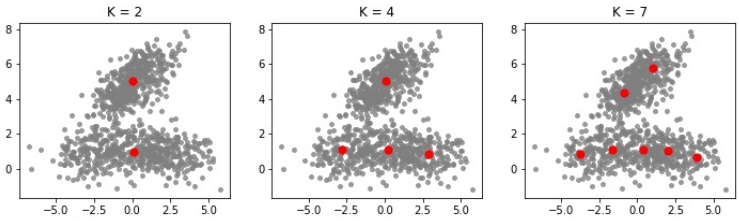


Figure 5: Clustering results for different values of K

To choose a proper K, one of the well-known techniques is to use the *Elbow Curve* which is the graph where X-axis represents the value of K and Y-axis represents average distance to centroid as displayed in Figure 6. From the elbow curve, K = 3 is the elbow of the curve which indicates the optimal value of K.

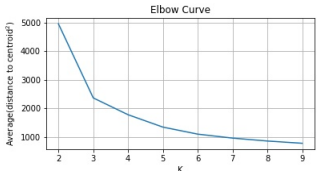
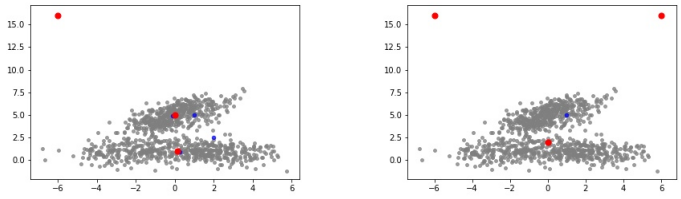


Figure 6

3.2 Initial Centroids

Initial centroids are another factor which has a huge impact on K-Means. The poor selection of centroids could leads to the algorithm returns incorrect results as illustrated in the figure below. In Figure 7a and 7b, there is one initial centroid which is the furthest and the closest to all datapoints respectively.



(a) (b)

Figure 7

The cause of this issue is the Expectation-Maximizatioin algorithm, which is the key of K-Means, stucks at the local minimum. In Figure 7a, no data is assigned to the top centroid in the first iteration. Therefore, when K-Means re-average the centroid to minimize the total inter-cluster distance and maximize the likelihood, the top centroid does not move and the algorithm ends up with

one empty cluster even though the K is correct. The similar situation also occurs on the Figure 7b where all samples are assigned to the bottom centroid.

One of the popular techniques for choosing initial centroids is to uniformly pick K points in the dataset as an initial centroid. Although this could reduce the chance, this method does not guarantee that the algorithm will not stuck at the suboptimal solution. The more sophisticated approach for this is *k-means++*[1] which is based on the same princleple but using weighted probability distribution instead of uniform sampling.

4 Testing with the UCI dataset

In this section, we will verify our K-Means implementation on the Iris dataset from UCI repository. The dataset contains 3 classes where each sample has 4 features and distributes as follow.

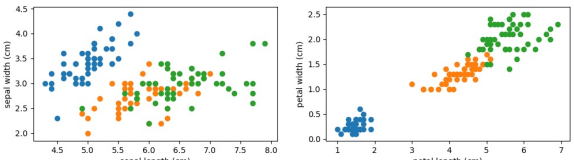
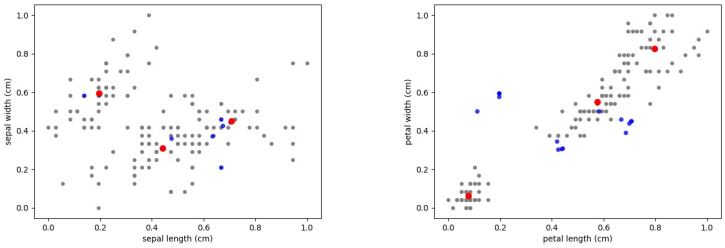


Figure 8

Using K-Means(with K = 3) to cluster the dataset gives the cluster centroids displayed in Figure 9. The result is matched with the dataset description.



(a) (b)

Figure 9

References

[1] D. Arthur and S. Vassilvitskii, *k-means++: The Advantages of Careful Seeding*. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, 322(10):891921, 2007.