2021

Assignment 3

HAIJIA ZHU

Contents

1. K-means Clustering

The K-means clustering is implemented in Matlab environment, two different initialization method is developed.

The stop criterion is set to be the number of iterations, and 10 as the default setting. However, to ensure the (local) minimum is found, the number of iterations is set to 50 in this experiment.

1.1. On datasets1

Figure 1 presents the cost of the cluster at different k for both methods. When the number of clusters is smaller than the “optimal” number, random initialization tends to have a larger cost compare to k-means++, as they approach the optimal number of clusters, the cost converges. Here in this dataset, the “optimal” number of clusters is selected using the elbow method, that is to pick the elbow of the curve as the number of clusters to use in this experiment. Here we pick 10 and 8 for random initialization and k-means++ method, respectively. The k-means method converges significantly faster which indicates a better clustering performance. The clustering result is presented in Figure 2 and Figure 3, and the result is almost identical if we consider the boxed clusters in Figure 2 as a group (consider the clusters in black circle and green circle as a whole).

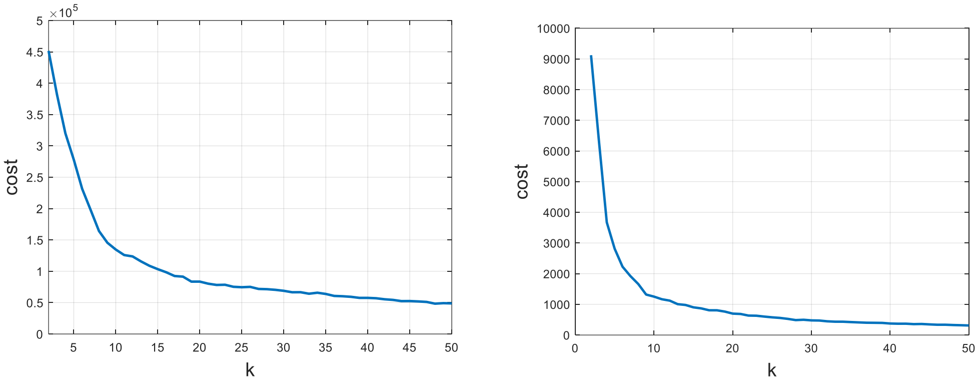

Figure 1 K-means, with random initialization (left) and k-means++ (right)

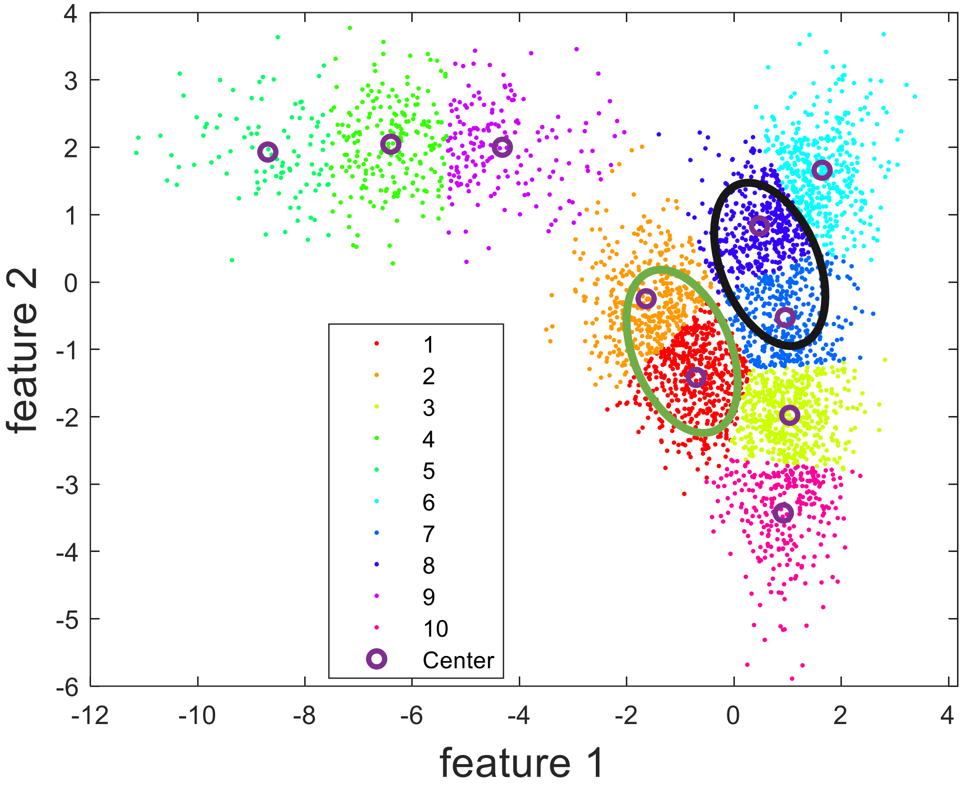

Figure 2 K-means, with random initialization and k=10

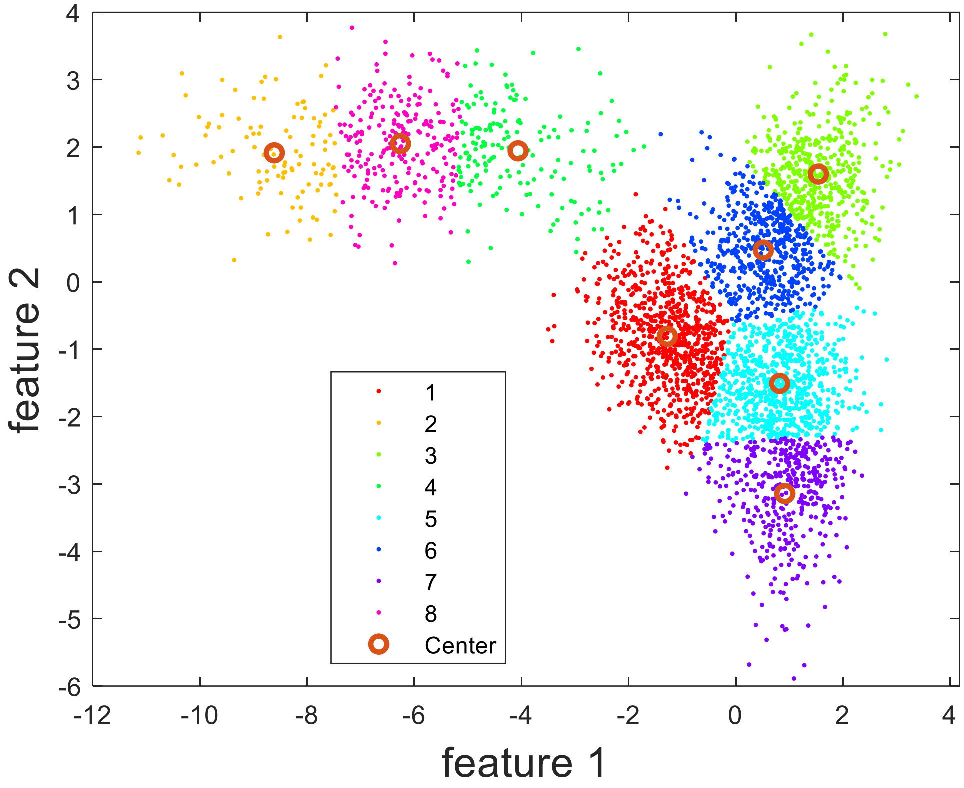

Figure 3 K-means++ and k=9

1.2. On datasets2

A similar result is obtained when performing k-means on datasets 2. The random initialization performs poorly when the number of clusters is small. The cost reduces as the k increase, but k-means++ has a smaller slope (cost reduce faster), as shown in the pictures below:

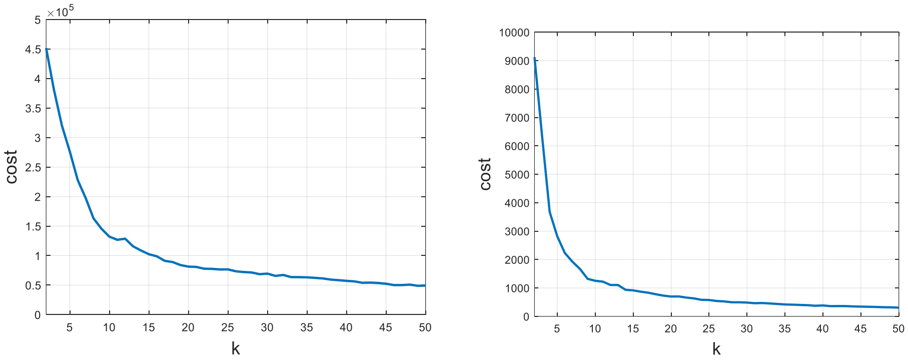

Figure 4 K-means, with random initialization (left) and k-means++ (right) on dataset 2

The clustering result differs, as shown in Figure 5 and Figure 6, although they have the same k. One reason is that the k from the random initialization is picked at the upper part of the elbow of the curve while the k-means++ is chosen from the lower part of the elbow of the curve (which has a lower cost). The cost of k-means++ is smaller than the random initialization method.

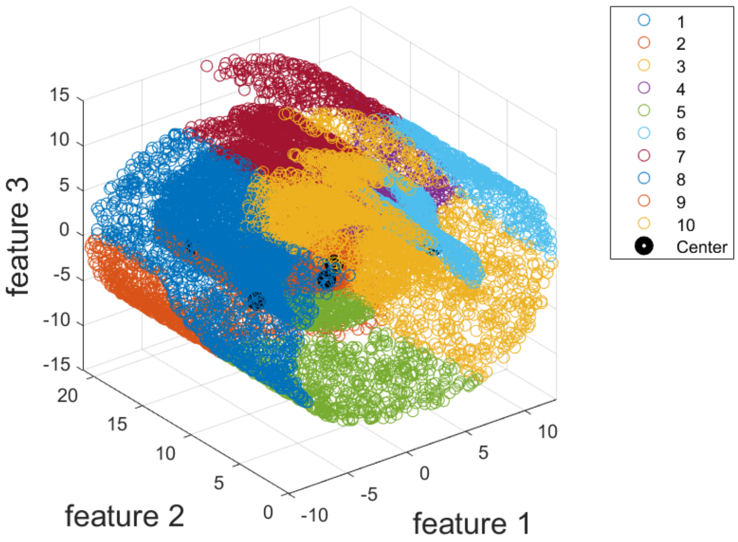

Figure 5 K-means, with random initialization and k=10

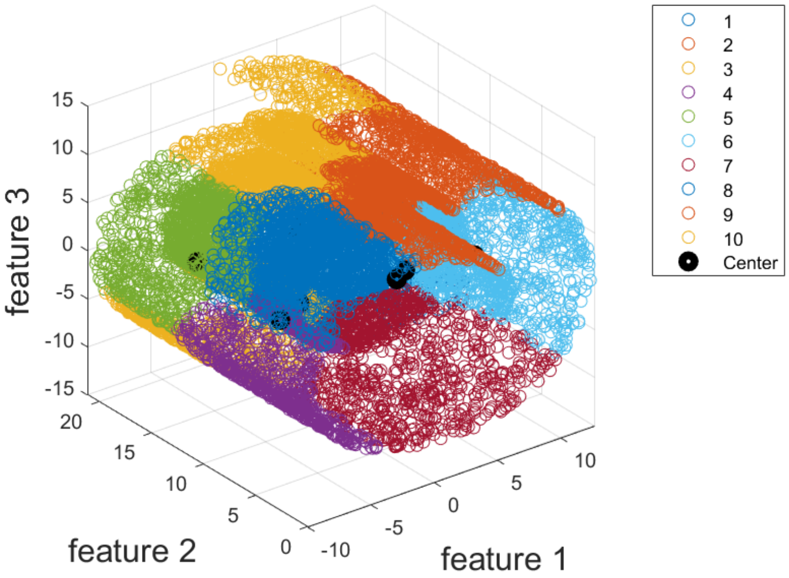

Figure 6 K-means++, and k=10

2. Hierarchical Clustering

A link that is approximately the same height indicates that there are no distinct divisions between the objects joined at this level of the hierarchy because the distance between the objects being joined is approximately the same as the distances between the objects they contain. However, it is hard to determine the cutoff (inconsistent values) parameters without an understanding of the data and the purpose of the cluster. Instead, we will define the max cluster number in this experiment.

2.1. On dataset1

2.1.1. Average

The black line in the picture below represents a potential optimal cluster number. By cutting along this line, the datasets are divided into 6 groups. The reason for selecting this line as the threshold is that the height (which represents the distance between the clusters) start to decrease (such as compare to divided into two clusters), and the resulting cluster has an even distribution (an uneven split can be found in next section).

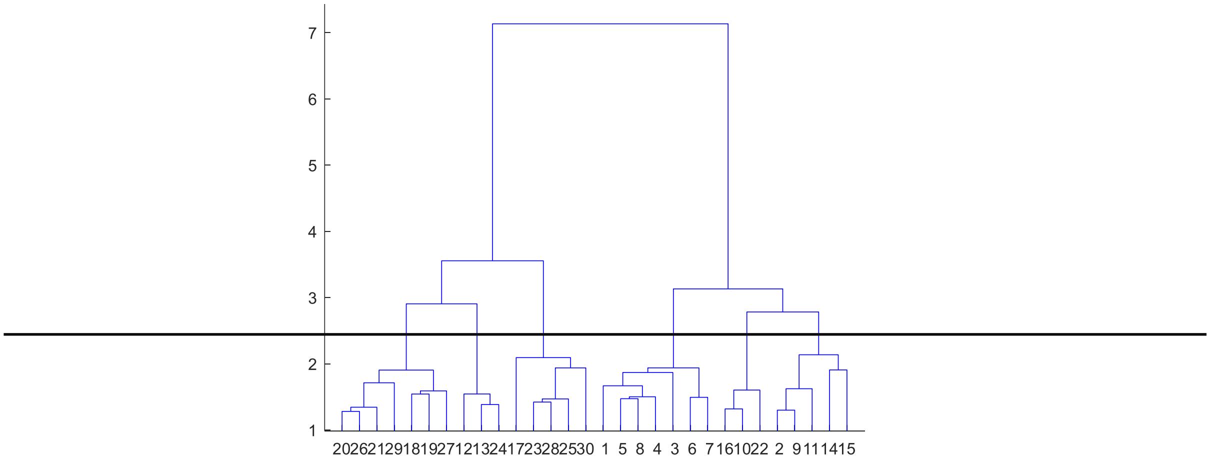

Figure 7

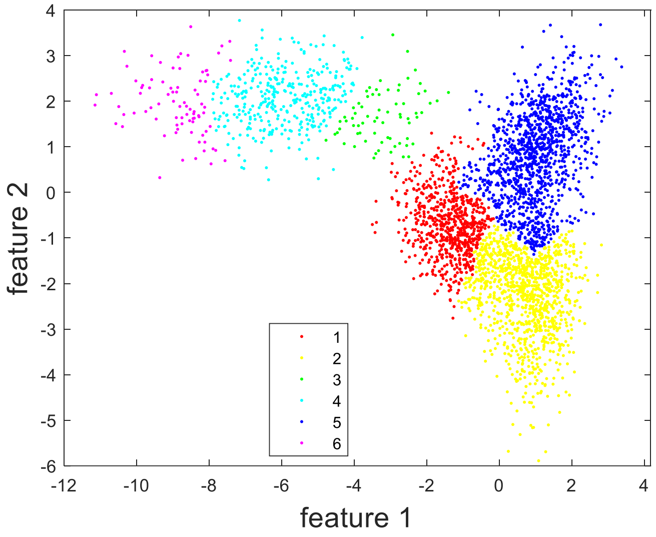

Figure 8

2.1.2. Single

When using the shortest distance between the cluster to group the data point, the dendrogram plot suggests that the distance between each cluster is small, and the largest height is only 0.75, and mainly gathered on the left-hand side. This suggests that most of the point is relatively consistent and it will be difficult to cluster the dataset evenly. And if the cluster number is assigned to 2 (or 3, or even more), the black box in Figure 9 suggests that this node distinct from other nodes and will be considered a cluster. Further experiment shows that this method struggles to divide the point into the different cluster and tend to divide most of the point into one group, as shown in Figure 10, Figure 11, Figure 13. Over single linkage is not suitable for this dataset. Even if the cluster number increases to 500, we can find that most of the point is group into on clusters, as shown in Figure 13. If one must pick a number, 100 is an optional number as it can split the dataset into several different groups.

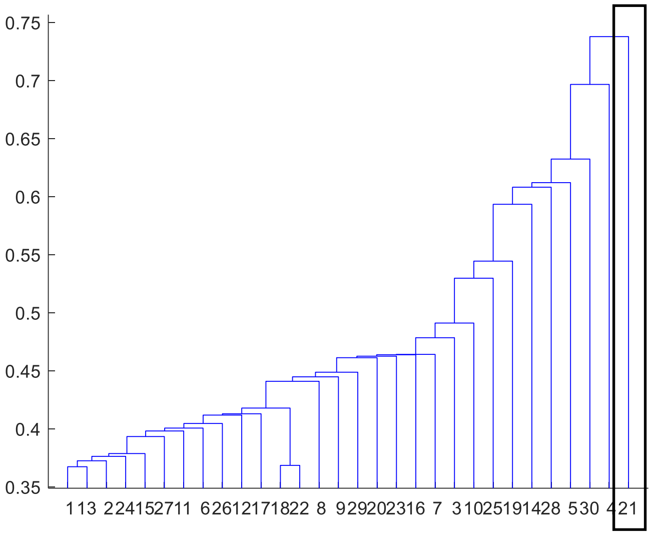

Figure 9 Tree

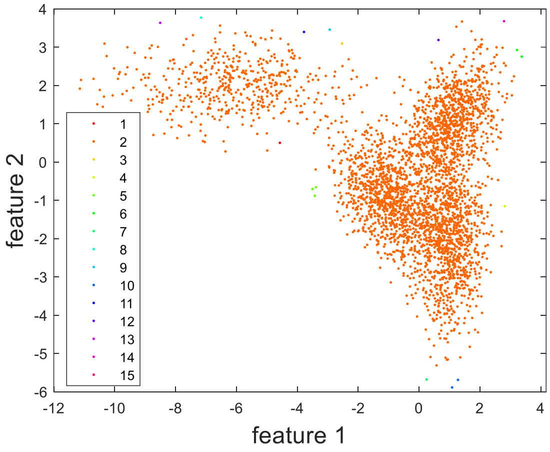

Figure 10   Single, max cluster number = 15

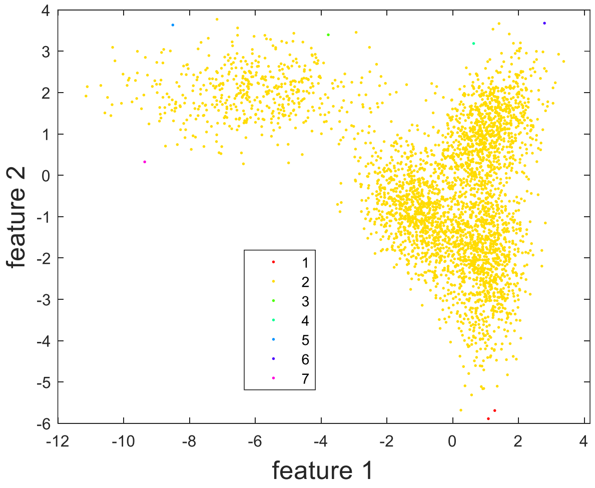

Figure 11 Single, max cluster number = 7

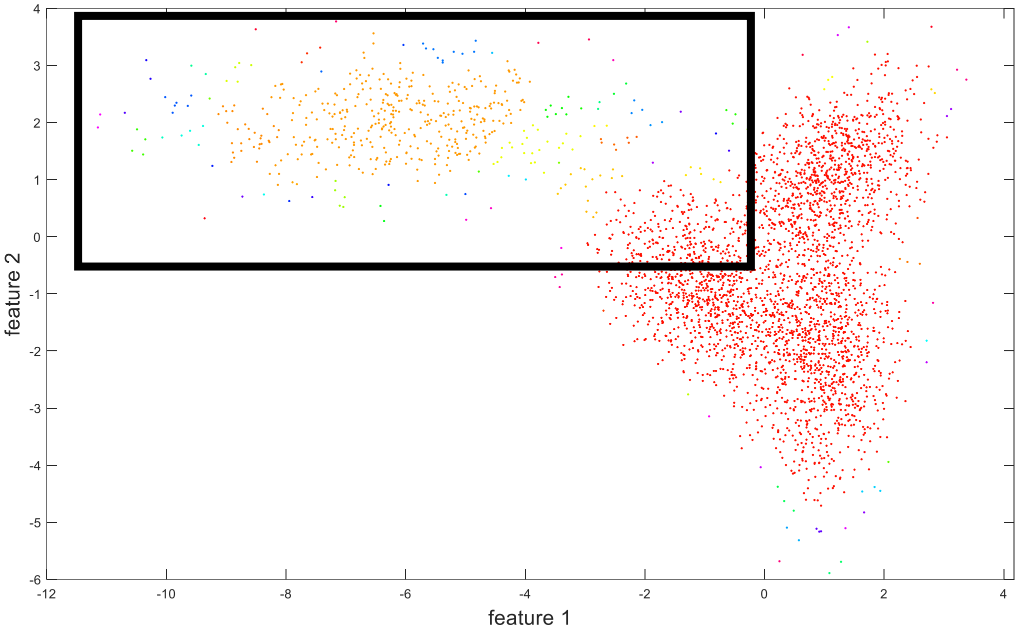

Figure 12 Single, max cluster number = 100

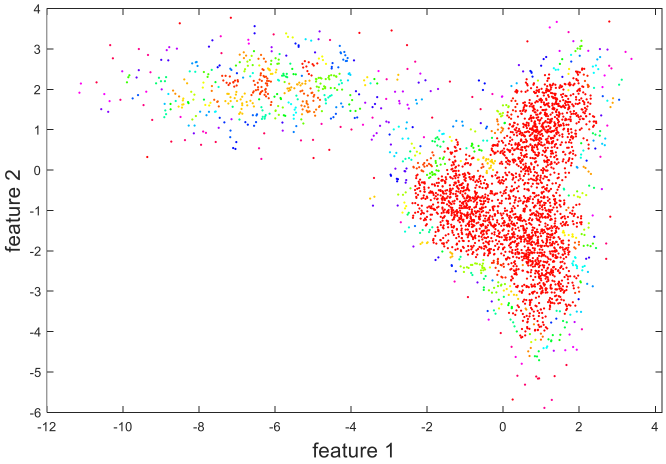

Figure 13 Single, 500 cluster

2.2. On dataset2

2.2.1. Average

For dataset2, several thresholds can be chosen but 13 is selected eventually. The reason for choosing this threshold is similar to the reason of chosen the threshold in On dataset1.

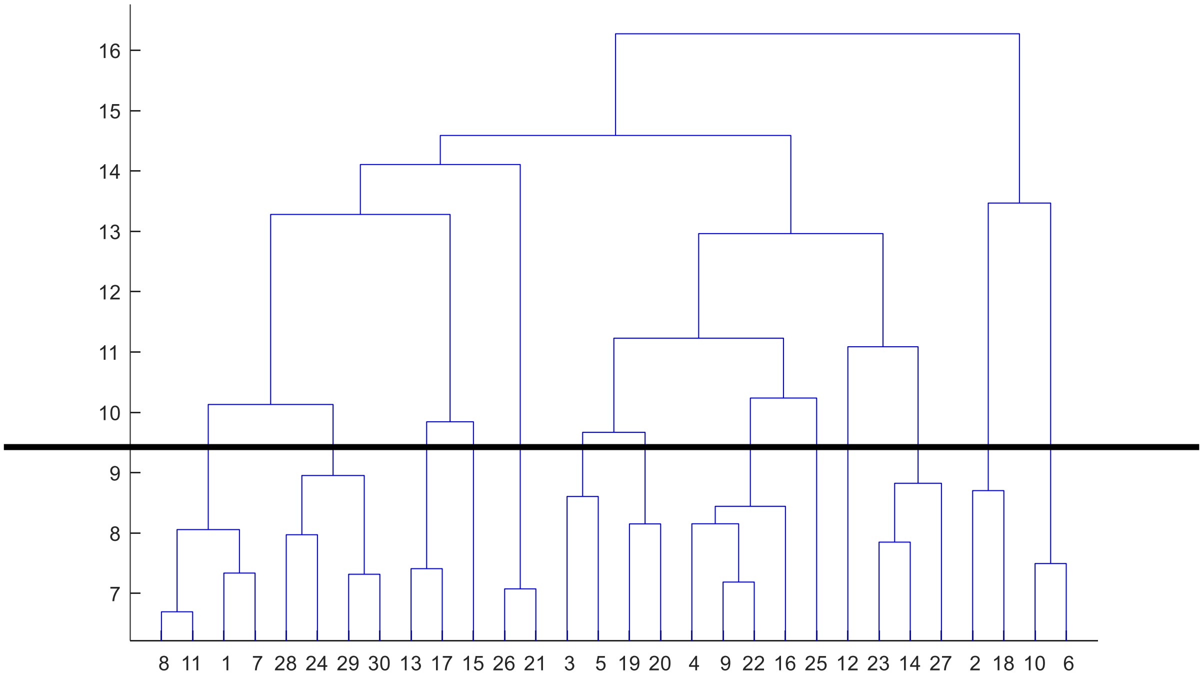

Figure 14 Hierarchical Binary Cluster Tree

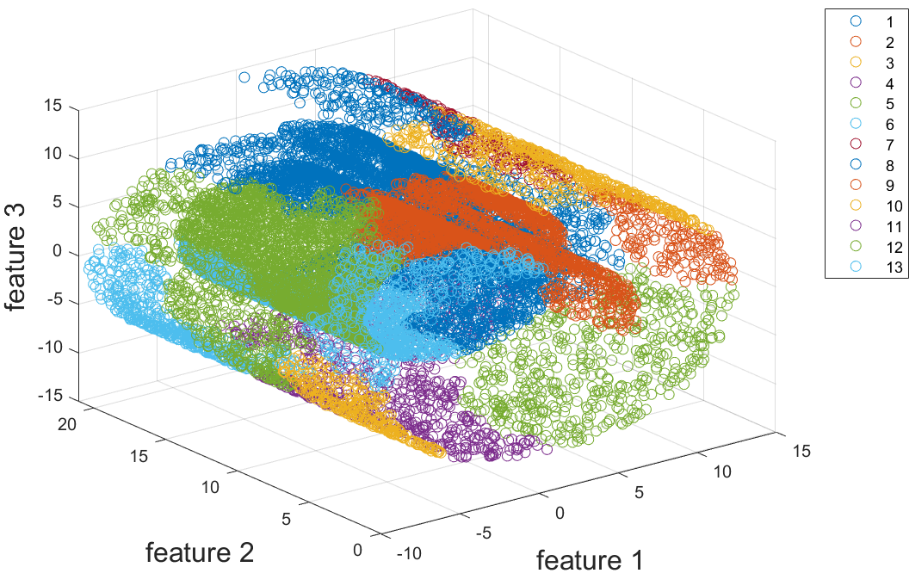

Figure 15   Single, 13 cluster

2.2.2. Single

The number of clusters is chosen as 2 for the hierarchical clustering using single, as it provides a good separation, acceptable result, and successful distinct inner circle with the outer circle, as shown in Figure 18. Depends on the purpose and the type of data, it is possible that this clustering is better than others. Intuitively, one would consider the inner circle as a group and the outer circle as another group.

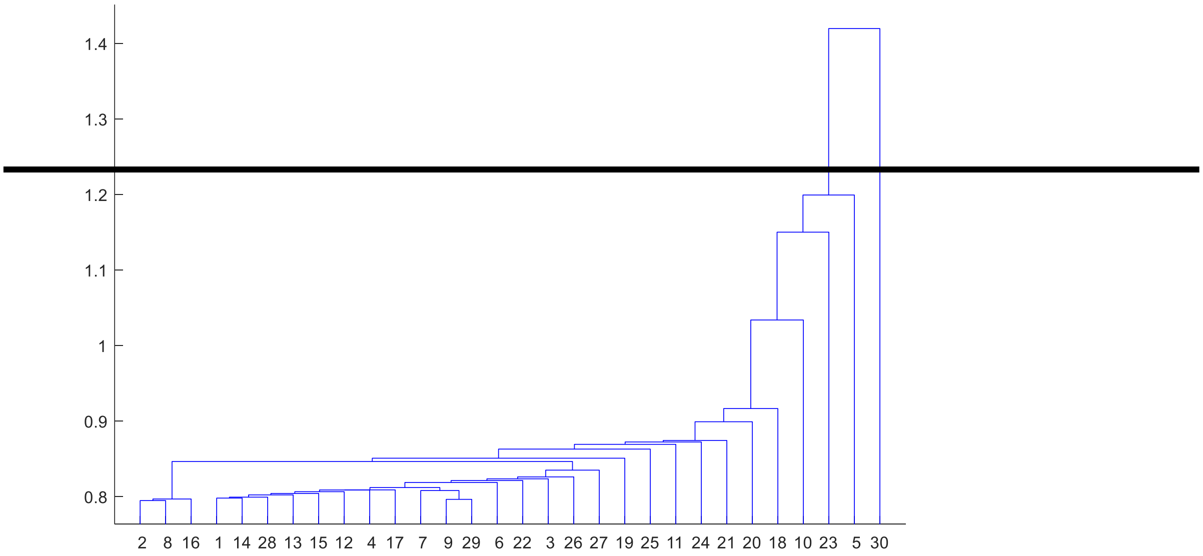

Figure 16 Hierarchical Binary Cluster Tree

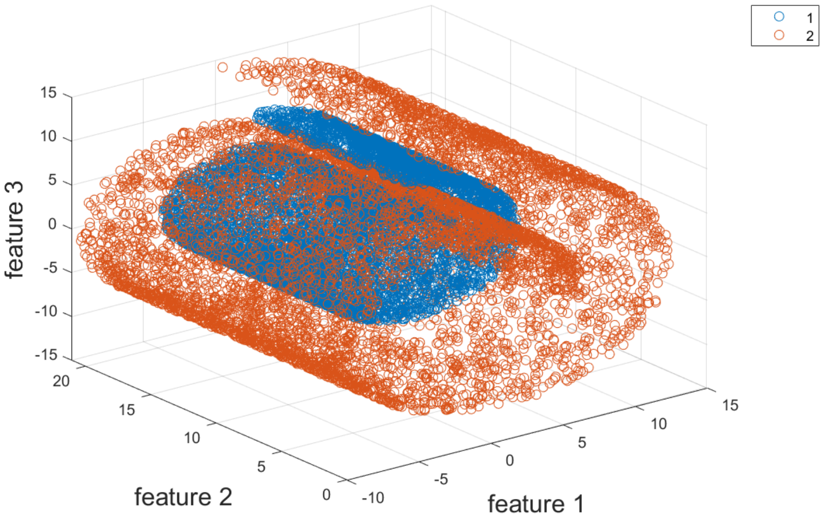

Figure 17   Single, 2 clusters

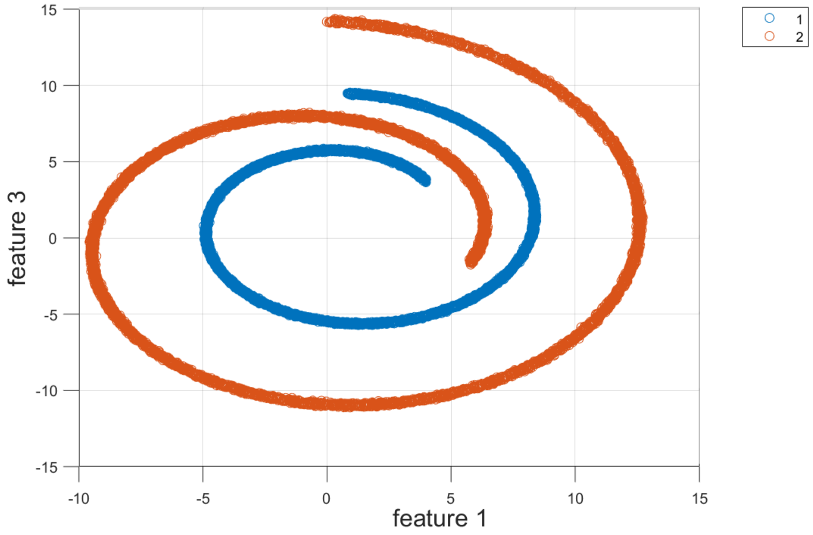

Figure 18   Single, 2 cluster, side view