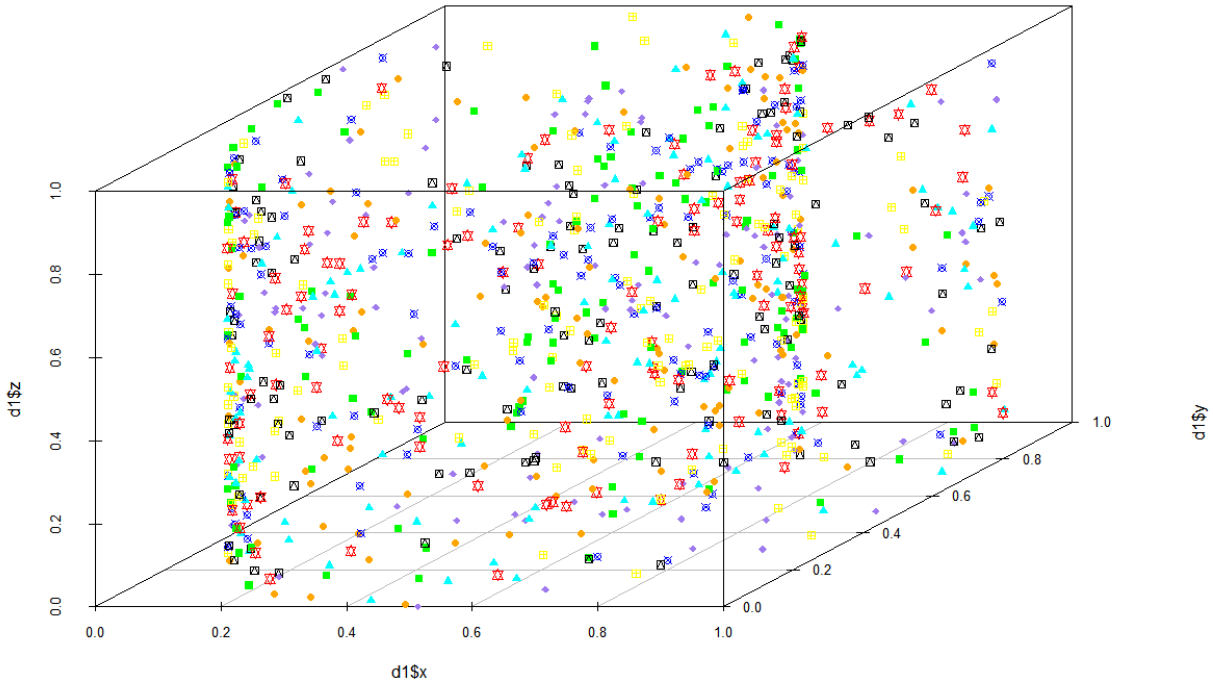
CIS 6930 Project II Report

Xi Yang 16216573

In this project, we have two tasks to perform. The first one is to try different clustering methods on a data set with ground truth to understanding the performance of each methods. The second task is based on the first task using an appropriate method selected from task one to cluster a data set with 1 million individual data points.

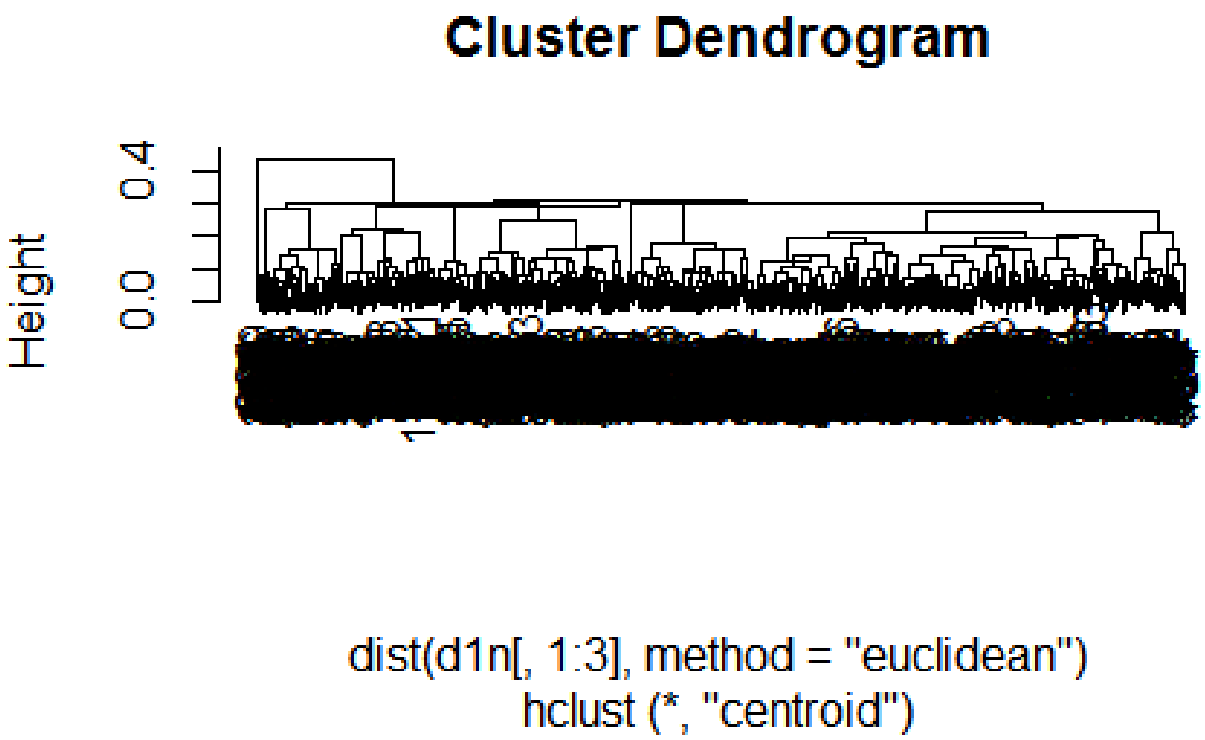
Detail of the first task:

First plotting the data set, a 3D plot depicted data is shown as follow:



From the plot, we can see the data are not clearly separated. Therefore, the low accuracy can be expected. Before applying clustering method on the data, a data scaling (normalization) is executed to normalize the data.

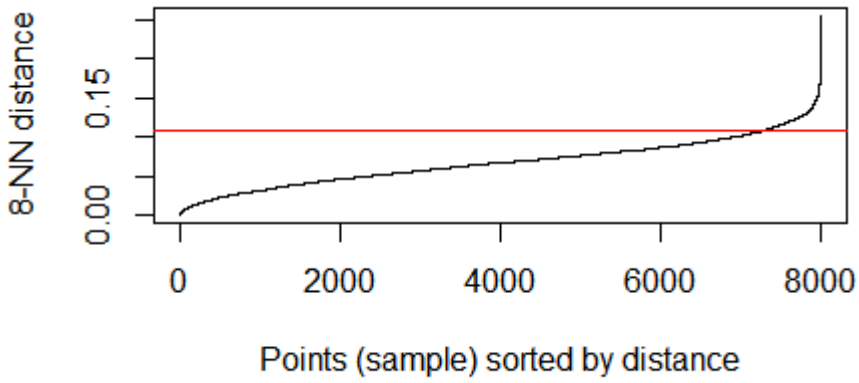
The first method we used is hierarchy method with divisive fashion based on distance matrix. After clustering, a dendrogram like structure is generated as follow:



At the bottom, every data points are in one cluster, then the closeness will be calculated between each two of them, and merged the two closed points into one cluster until only one cluster left. Complete linkage and mean linkage clustering are the ones used most often to determine the closeness of two clusters. Applying this method on the data set and comparing the clustered result with the ground truth, an accuracy of 0.126 is obtained.

The second method is Kmeans which is a very common and fast method for clustering. The K means method aims to partition the points into k groups such that the sum of squares from points to the assigned cluster centres is minimized. To utilize this method, initially a set of random centers are used, so each time the clustering result might vary. Therefore, 5 times repeat experiments are executed and the average accuracy is reported as 0.129. In the experiment, 25 start points are used as initial centers and the calculation algorithm is set to “MacQueen” instead of “Lloyd” which seems yield better results.

The third method is density based method specifically named dbscan. The dbscan gives a set of points in some space. it groups together points that are closely packed together (points with many nearby neighbors), marking as outlier points that lie alone in low-density regions (whose nearest neighbors are too far away). To perform the experiment, we first need to determine the two import parameters: eps(size of the epsilon neighborhood) and minpts(number of minimum points in the eps region). A standard way to determine eps is using knn distance histogram and selecting the point where dramatical change happened as eps. On the other hand, there is no general way of choosing minPts and to obtain the cluster group numbers close to ground truth, 8 was chosen. Running knn distance histogram, a plot was obtained as follow and from the plot, we can see the changing point can be set to 0.108.



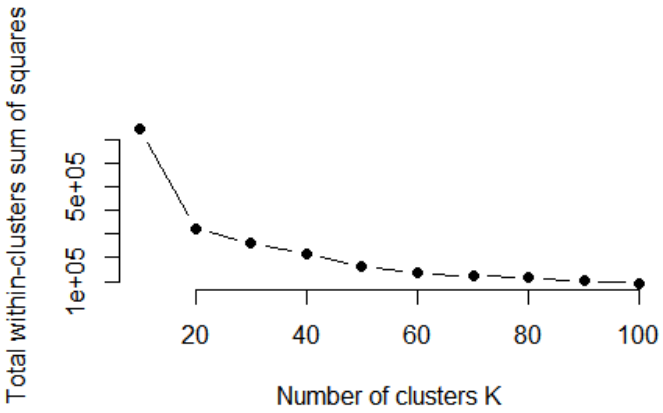
Clustering with the dbscan method and comparing the result with the ground truth, the accuracy obtained is 0.123.

The last method should be a graph based method. The most common method should be clique method. However, due to the data set, the clique method took too long to yield result. Instead, SNN clustering method is used which is based on Jarvis-Patrick algorithm and involve a construction of a SSN graph followed by weighting, filtering and clustering. To run this experiment, we also need to provide parameters as K(Neighborhood size for nearest neighbor sparsification to create the shared NN graph) , eps and minPts. In order to obtain cluster groups closing to 8, the parameters are determined by screening. After the clustering, the accuracy is calculated as 0.117.

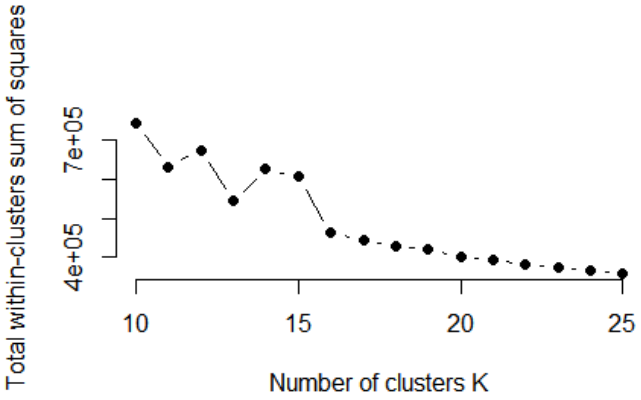
Detail of the second task:

First step is to determine which method will be used with respecting to the fact that we have about 1 million data points. The hierarchy and SNN methods are not feasible because both method need to initialize a matrix based on data so in this case the matrix will be too large (3700GB) for the memory. Comparing kmeans with dbscan, the kmeans runs significantly faster and the parameters can be easily obtained. Therefore, the kmeans is used for the task 2.

The next question is to determine how many clusters should be assigned to the data set. The elbow method is used. The first elbow screen used the input as (10, 20, 30, 40, 50, 60, 70, 80, 90, 100) and the plot obtained show as



An interpretation illustrates the group numbers should be between 10 and 25. Then a second elbow experiment ran with input from 10 to 25. From the plot, a conclusion of cluster group number can be determined as two possible number: 15 or 16.



Running K means method with k=16 on data set, we obtain an evaluation parameters table as

```
totss   tot.withinss    betweenss
4024728       590563    3434165
```

The totss is the total sum of squares and betweenss is the between groups deviance. The good clustering fit usurally have a bss/tss ratio close to one means most data points are separated into different groups. In the case of k=16, the averaged ratio of bss and tss is 85.3% which indicate a good fit. And when k=15, the averaged ratio of bss and tss is 84.8%.