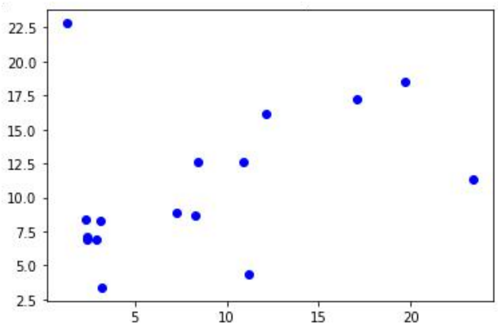
Bingbing Li
18210705

Data Mining COMP40370
Practical 6

Date 03/11/2019

Q1: Using all the attributes, performs the k-means algorithm for three clusters. Discuss the obtained clustering results in your report.



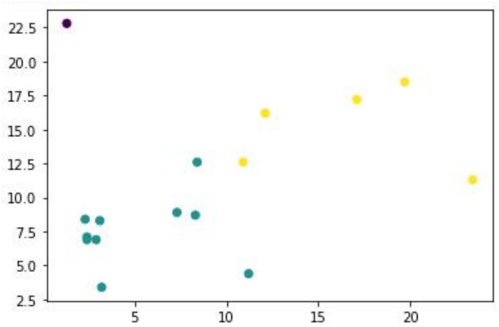
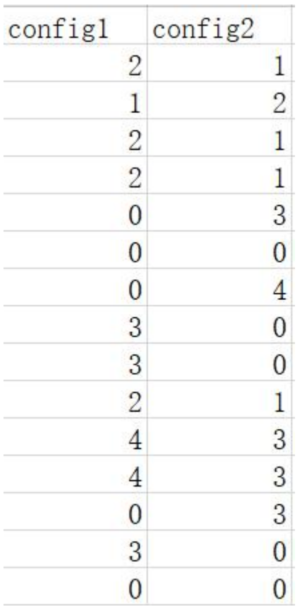


Figure 1 shows the original data loaded into two dimensions using plt.scatter() method. Figure is the result after using K-means algorithm to all the attributes. Because the parameters n_clusters in K-means was set to the value of three, the data set was divided into 3 clusters. The points which are have the same colour belong to one cluster: the “green”, “yellow” and “purple” demonstrate three different clusters. The main area of these three clusters is X(0:10) Y(2.5:12.5), X(10:25) Y(10:20) and a point which is far away from the former two clusters.

Q2:

(1) Are the clustering results obtained with the first configuration different from the results obtained with the second configuration?

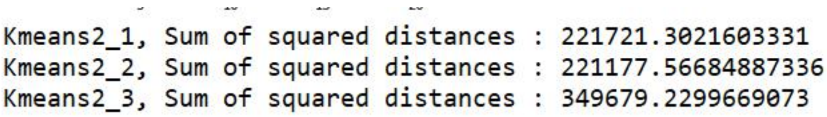


config1	config2
2	1
1	2
2	1
2	1
0	3
0	0
0	4
3	0
3	0
2	1
4	3
4	3
0	3
3	0
0	0

We can have a look at the generated columns in “output/question 2.csv”.
The config1 was the cluster labels generated using the method : Run the k-means algorithm using 5 clusters as target, 5 maximum runs,and 100 maximum optimization steps. Keep the random state to 0.
The config2 was the cluster labels generated using the method : Run the k-means algorithm using 5 clusters as target, 100 maximum runs,and 100 maximum optimization steps. Keep the random state to 0.
We can see from the config1 and the config2 that after taking the name of the cluster was different (such as 1 in config1,2 in config2 represent the same cluster), there is still some points was set into different clusters when using different parameters. For instance, the row 5 and row 7 are both cluster0 in config1 but was set into cluster 3 and 4 in config2.
The reasons of the difference is that the difference of n_init which is number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia. Therefore ,the higher the value is ,the more likely the k-means algorithm can generate better result. So the second method and config2 should be better.

(2) Which clustering solution is the better?

Because the original data set have several dimensions, so it would not be feasible to use the plt to plot the points in different colors to show the result after clustering. Therefore, I print the sum of squared distances of samples to their closest cluster center using the parameter .inertia_.

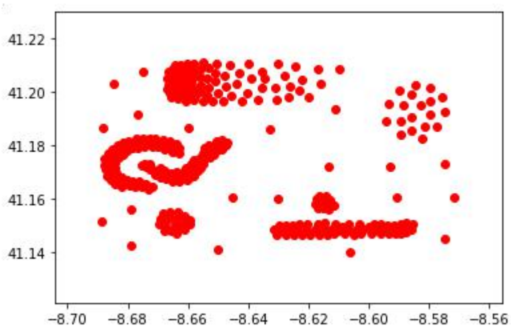


```
Kmeans2_1, Sum of squared distances : 221721.3021603331
Kmeans2_2, Sum of squared distances : 221177.56684887336
Kmeans2_3, Sum of squared distances : 349679.2299669073
```

As the result shows the Keams2_2 which is:
 run the k-means algorithm using 5 clusters as target, 100 maximum runs,and 100 maximum optimization steps. Keep the random state to 0
has the least sum of squared distances. So the best solution is the second method. Moreover, from the experiment we can see that when the number of the clusters, the time of the maximum runs and the time of maximum optimization steps are relatively bigger, the cluster performance is better.

Q3:

(1) Discuss the cluster results in your report.



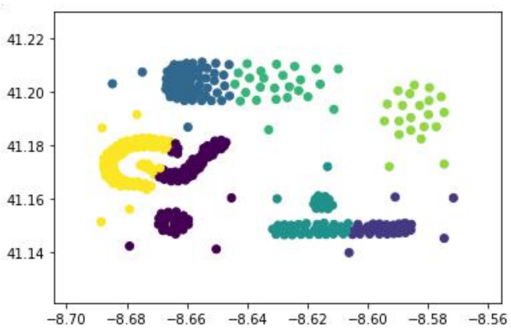
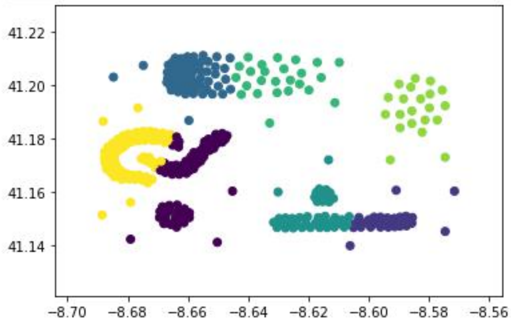
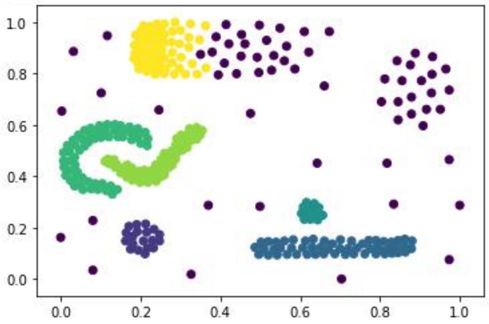
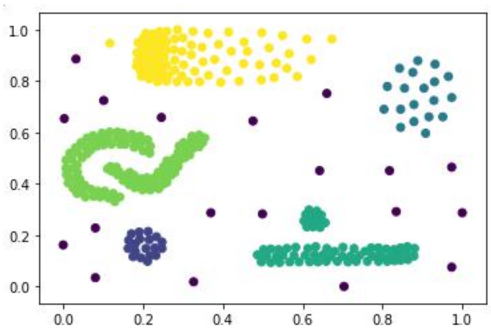


Figure 1 shows the original data loaded into two dimensions using plt.scatter() method. Figure is the result after using K-means algorithm to all the attributes.Because the parameters n_clusters in K-means was set to the value of seven, the data set was divided into 7 clusters. But we can see from the figure two that the cluster performs badly to those non convex shaped clusters and outlier points in the data.

(2) Discuss the different clustering solutions in your report. Which solution is the best? What is the reason behind the differences in the results?







The figure1 shows the clustering result using k-means algorithm. The figure2 and figures3 show the clustering result using DBSCAN algorithm.

Comparing the figure 1 with figure 2 and figure 3, we can see that k-means performs badly with those non convex shaped clusters and outlier points in the data.

In terms of the figure 2 and figure 3, we can see that the DBSCAN can perform better with the non convex shaped clusters and outlier points.

But there is still difference between these two cluster result which is that the method two the data set which has the same shape was set into different clusters, while in the figure 3, it cluster the data set correctly.

The reason to this difference is the different value of the eps parameter which is the maximum distance between two samples for one to be considered as in the neighborhood of the other. The eps in method 2 is 0.04 and in method 3 is 0.08. So in conclusion, the higher the eps value is , the data sets of the same shape are more likely to remain in the same cluster. In conclusion, the third solution is the best one.