MDM-2021 Assignment 2

Jeyhun Yagublu (SN:1011331)

October 24, 2021

1 Task 3

In this task, you should study two internal clustering validation indices, Silhouette index (SI) and Davies-Bouldin index (DB), and one external index, Normalized Mutual Information (NMI), the version by Strehl and Ghosh, 2003 (see lecture 5 slides). Load two data sets, "balls.txt" and "spirals.txt". Both are two-dimensional data, where the third feature ("class") contains the ground-truth labels. Remember to discard the label before running the clustering algorithms.

1.1 A)

Task: (Warm-up) Cluster "balls.txt" with i) K-means and ii) spectral clustering using a Gaussian kernel and a Laplacian matrix of your choice. You can try different values of the kernel parameter, to see if it has any effect. Note: if you are using a software package, try to figure out which Laplacian matrix it uses. The distance measure is Euclidean.
Test values K = 2, . . . , 5 and determine the optimal number of clusters for both methods using all three indices (SI, DB, NMI). Report the results as a table. Which method and K are best for the data?
Answer: I have scatterplotted the data according to given class:

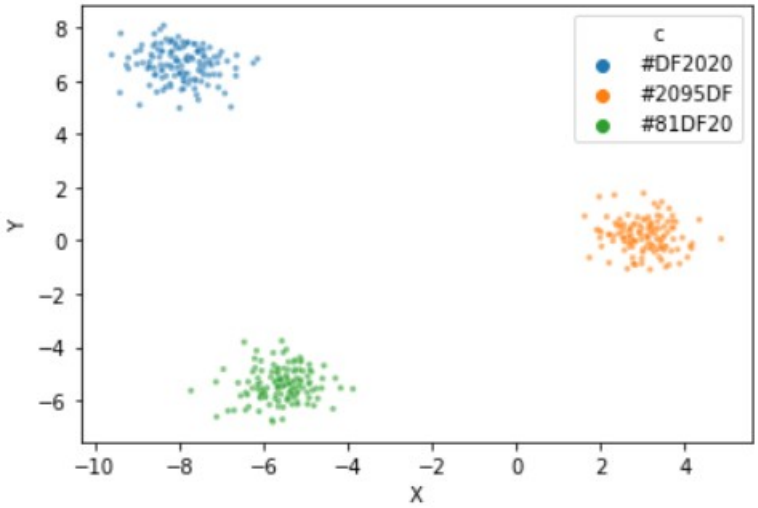

Figure 1: scatter plot of balls.txt

Table 1: Result of clustering

	SI	DB	NMI
K-means	0.9015 for K=3	0.1359 for K=3	1.0 for K=3
Spectral	0.9015 for K=3	0.1359 for K=3	1.0 for K=3

From the plots and data we can see that both Kmeans and Spectral Clustering gives same clusters for this data and best K for this data would be 3.

1.2 B)

Task: Repeat a) for "spirals.txt".

Answer: Results of clustering of spirals.txt :

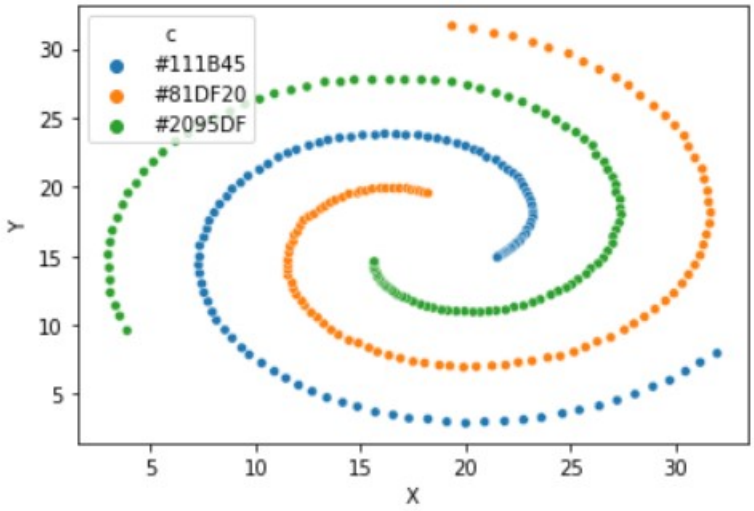

Figure 2: scatter plot of Spirals.txt for given classification

Table 2: Result of clustering

	SI	DB	NMI
K-means	0.3600 for K=3	0.8779 for K=3	0.0066 for K=5
Spectral	0.0253 for K=2	5.4585 for K=5	1.0 for K=3

As we can clearly see from NMI index Kmeans method is creates clusters which are very different from the given classification on the other hand Spectral clustering creates exactly same cluster as given classification when K=3.

1.3 C)

Task: Explain and analyze your observations. Which index captured the performance of the algorithm most accurately? Why some indices failed to reflect good performance? Can you use internal indices to determine optimal K for spectral clustering? It is recommended to look at the definitions of indices to better understand their objectives.

Answer: The most successful index in these cases were NMI. It is obvious that since NMI compares the clusters to the given classification it was the most successful to differentiate between good clustering and bad clustering. the other indices fail because they try to manage the clusters in a way that they intracluster distance is minimum and intercluster distances are maximum meaning that cluster try to create dense population of points in space which are far from the other dense points. For data which has non convex clusters these indices fail miserably as we have seen from the spiral data. I would say that indices are good depending on data type rather than clustering algorithm since we have seen that Spectral clustering gave very good results also for balls data and indices were succesful in identifying the correct K.