CS-E4650 - Assignment 2

Arttu Häkkinen - 596077

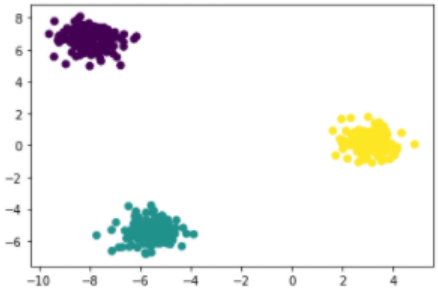
October 23, 2021

Task 3

For K-means clustering I used package [4] and for spectral clustering package [7]. Spectral clustering uses normalized Laplacian matrix. For clustering validation methods I used packages Silhouette index (SI) [2], Davies-Bouldin index (DB) [5] and Normalized Mutual Information (NMI) [6]. I also checked from the source code of NMI [6] that it uses the version by Strehl and Ghosh (lecture 5).

a)

I downloaded the data from "balls.txt" and plotted it coloring the data points based on their label. Three clear clusters are visible:

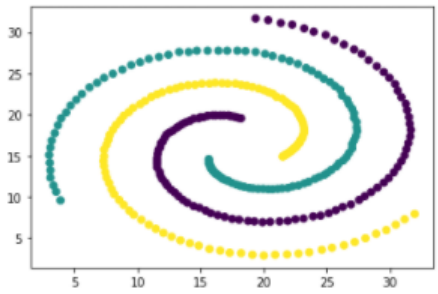


The best K for both of the clustering methods is obviously $K = 3$ and bth of the methods cluster the data perfectly with $K = 3$. The results table below is exactly similar for both of the clustering methods.

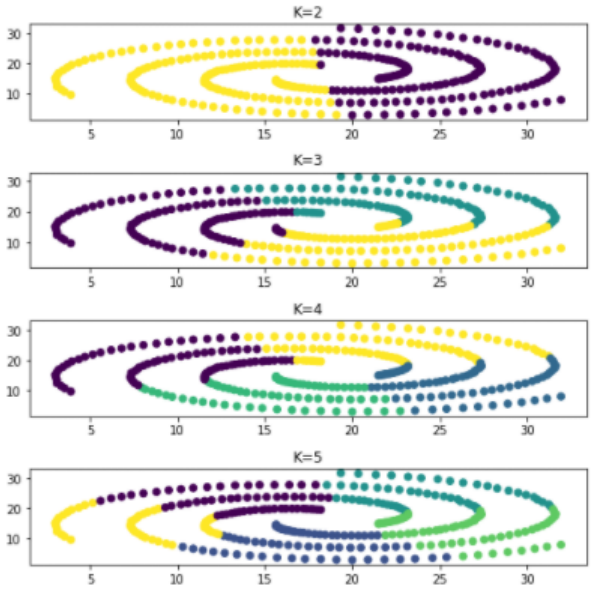
	SI	DB	NMI
K=2	0.67	0.52	0.73
K=3	0.90	0.14	1.00
K=4	0.71	0.65	0.90
K=5	0.52	0.95	0.83

b)

Downloaded the data and plotted it coloring based on the true class of the data points.



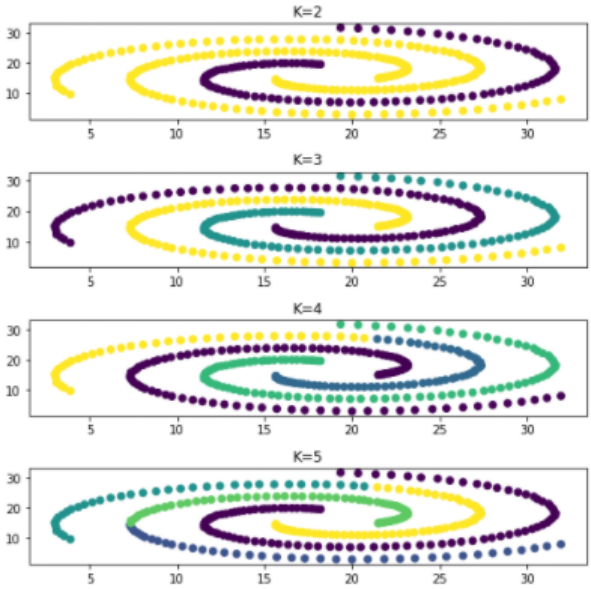
K-means clustering did not do well on this data set. The clusterings with different values of K are plotted below. None of them seems good.



Results of K-means clustering validation indexes on "spirals.txt" on table below:

	SI	DB	NMI
K=2	0.35	1.17	0.00
K=3	0.36	0.88	0.00
K=4	0.35	0.88	0.00
K=5	0.35	0.90	0.01

Spectral clustering on the other hand, managed to cluster this data set correctly with $K = 3$ as can be seen from the plots below.



Results of spectral clustering validation indexes on "spirals.txt" on table below:

	SI	DB	NMI
K=2	0.03	6.31	0.73
K=3	0.00	5.88	1.00
K=4	-0.01	6.91	0.91
K=5	0.02	5.46	0.83

So it can be concluded that from these two, the best method for the "spirals.txt" data set was spectral clustering with $K = 3$.

c)

The index that captured the performance of the algorithm most accurately was NMI. It scales the clustering results between 0 (no mutual information) and 1 (perfect correlation). Since $K = 3$ yielded best clustering for the data set, the NMI score also told us that the clustering was perfect even without plotting and making sure visually.

Some indices (mainly the internal ones) failed to reflect good performance because they only use the internal information of the data set. This means that

they do not have the information of the real labels and since the "spirals.txt" data set's real classes are not linearly separable, it can be hard to describe the performance solely based on the features and predicted labels. In other words, the unsupervised performance evaluation is a much harder task compared to external indices' supervised performance evaluation where the real labels are known.

Internal indices can't be used to determine optimal K for spectral clustering because SI suggests that the best clustering is $K = 2$ and DB suggests that it is $K = 5$, when the correct answer is $K = 3$ - as we knew already.

References

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html`

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html`

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html`

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html`