Daniel Illner
888013562
19.12.2019

Advanced Practical Course in Machine Learning - Solution Exercise 03

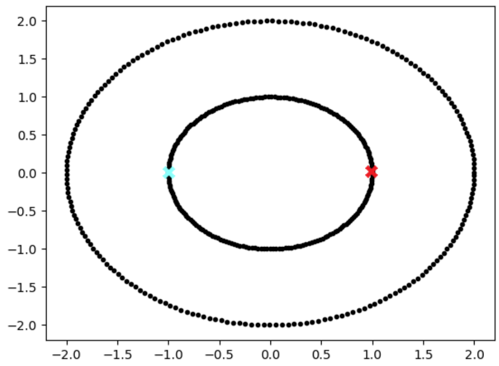
Clustering

Theoretical Questions

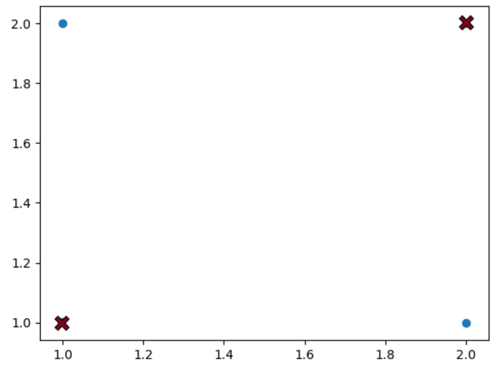
1.1)

1.2)

Results for 1 & 2:



(a) **1**: $k = 2$ and crosses as initial centroids.



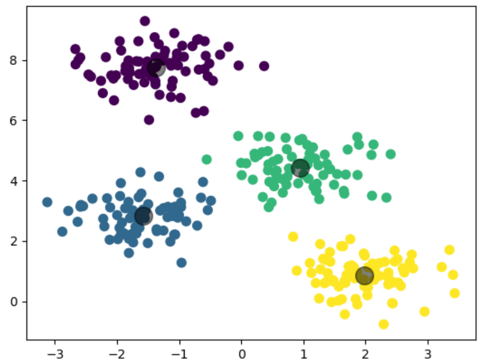
(b) **1**: $k = 2$ and crosses as initial centroids.

1.3)

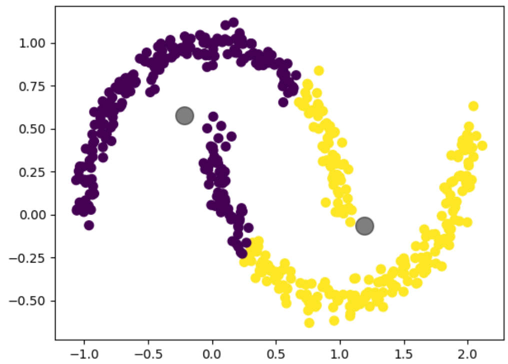
Practical Exercise

2.1.1) Implement a **K-means++** function and test it on synthetic data

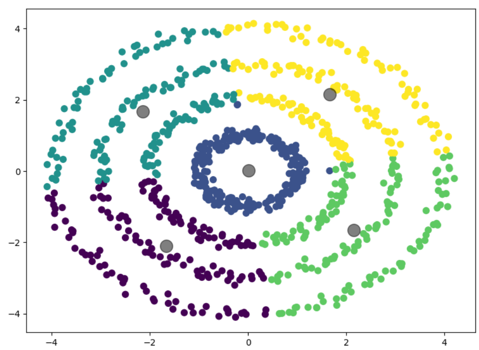
These are the results with the certain parameters:



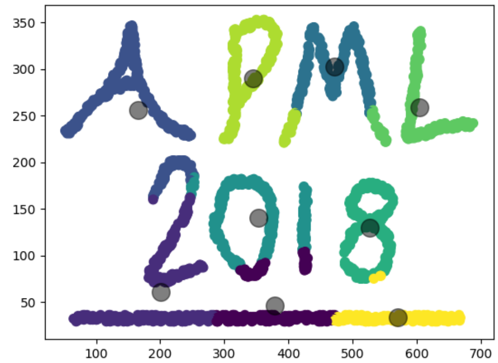
(a) K-means++ for random Blobs: $k = 4$



(b) K-means++ for two moons with noise: $k = 2$



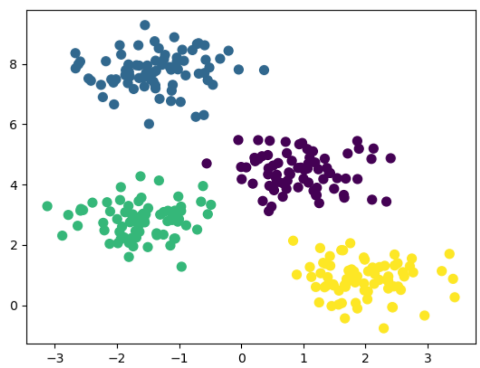
(a) K-means++ for synthetic circles: $k = 4$



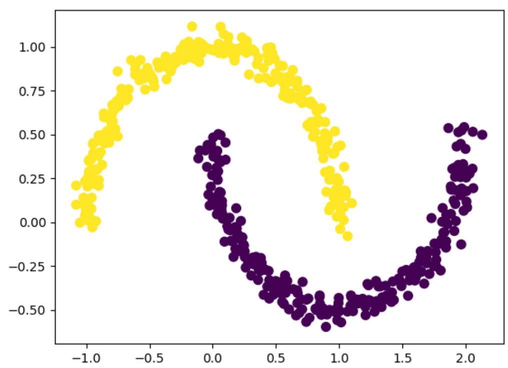
(b) K-means++ for apml: $k = 9$

2.1.2) Write a **spectral clustering** function and test it on synthetic data

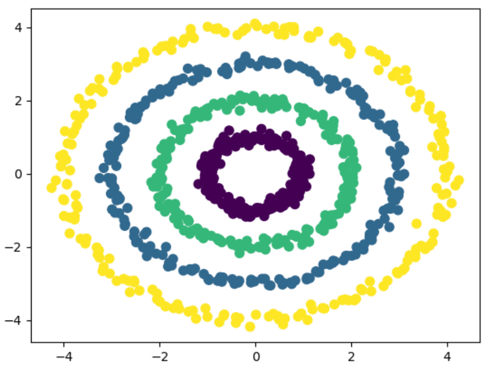
These are the results with the certain parameters:



(a) Spectral clustering for Blobs: $k = 4$ & $\sigma = 0.08$

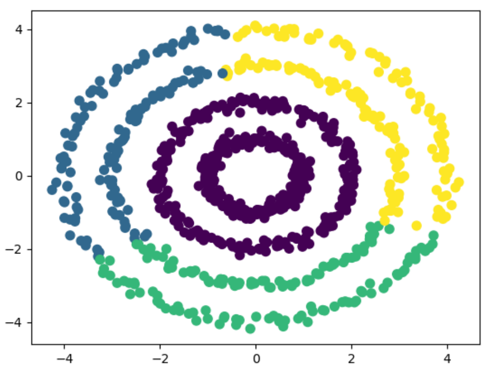


(b) Spectral clustering for two moons: $k = 2$ & $\sigma = 0.04$

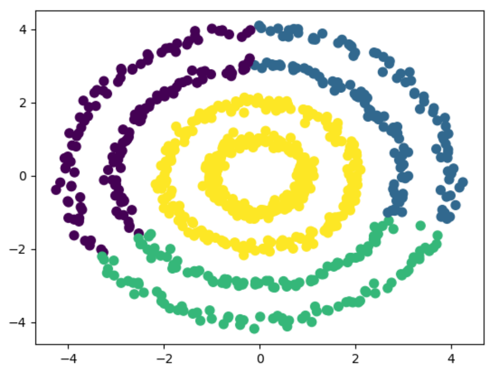


(a) Spectral clustering for synthetic circles: $k = 4$ & $\sigma = 0.006$

As you can see by comparing the figures, the *spectral clustering* solves the problems that *kmeans++* has with complex data structure. Choosing the right parameter is sufficient. E.G if you compare the result of *spectral clustering* for different σ values. For the apml data the function *scipy.sparse.linalg.eigs* made some issues on the RAM of my computer, so I was not able to run it properly. But the answer should be sufficient nevertheless.



(a) SC for random Blobs: $k = 4$ & $\sigma = 0.26$

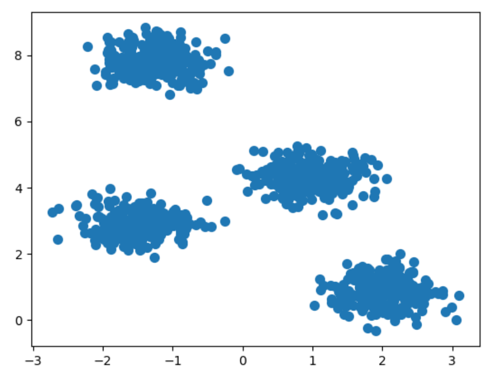


(b) SC for two moons with noise: $k = 4$ & $\sigma = 0.34$

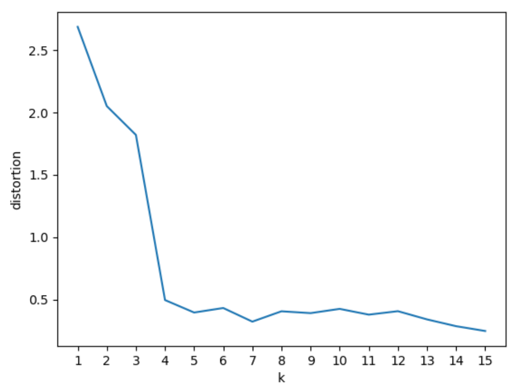
For finding values for σ, I was iterating over a list of different values and choose a σ that fitted my purpose.

2.1.3) Demonstration of the **elbow-method** on synthetic data (see function *def elbow_evaluation()* in the script).

These are the results on the dataset (left figure, 1000 datapoints & standart deviation: 0.4). You see the elbow graph on the right.



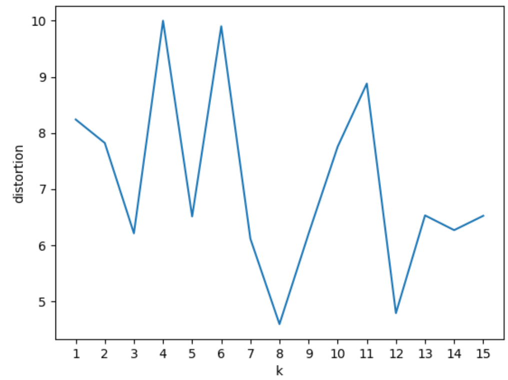
(a) Dataset with 1000 datapoints & $std = 0.4$



(b) Elbow Method result graph: $k = 4$

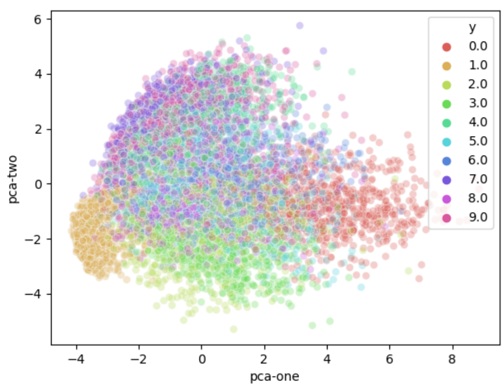
2.1.4) Applying the K-means and spectral clustering to the microarray data set.

Trying to figure out the optimal k in this case. For $kmeans++$ the elbow method gives me no result at all. The elbow graph is useless in this case.

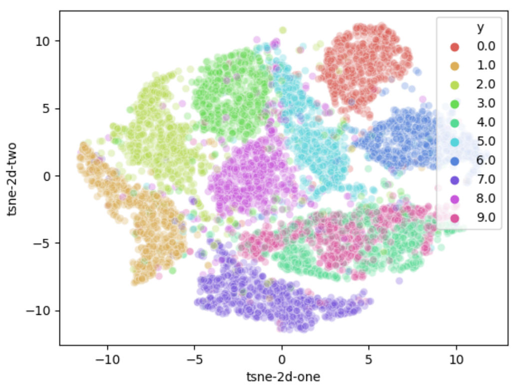


(a) Elbow Graph of k-means++ on microarray data

2.1.5) Compare **t-SNE** to **PCA** by using MNIST data set



(a) PCA on MNIST (8000 datapoints)

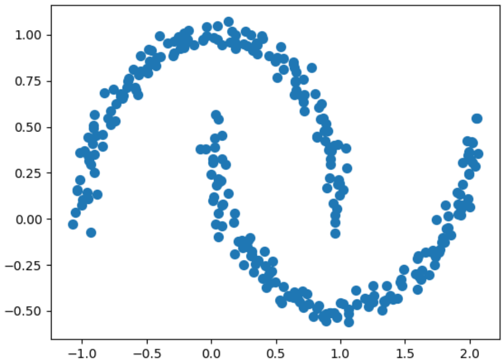


(b) t-SNE on MNIST (8000 datapoints)

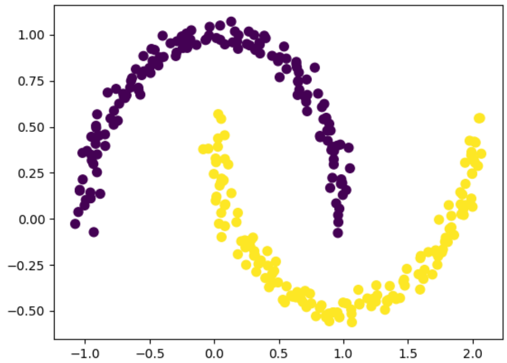
As you can see, the t-SNE does a quite good jobs clustering the 10 different numbers in the dataset. Compared to PCA, PCA does not manage to define the different clusters in a proper way. For the visualization the top two columns defined by the largest two eigenvalues where used. Check the function *def methods_comparison(n=8000)* for comparison and reproducing the results.

2.4.1) Plotting the Similarity Graph

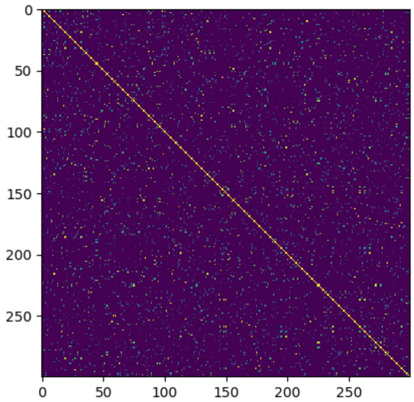
I plotted the similarity graph $\mathcal{W}$ for the following data (300 datapoints and $k = 2$). These are the results. I used the Gaussian Kernel for computing $\mathcal{W}$. You clearly see how $\mathcal{W}$ changed from random or unsorted to a clear structure, which directly results in a perfect clustering result.



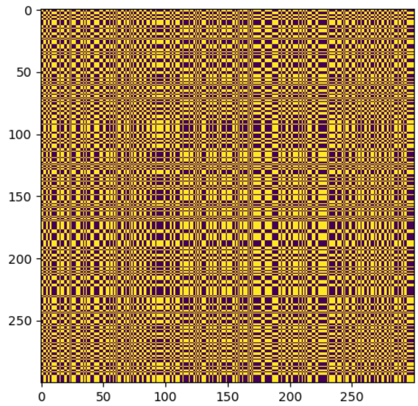
(a) Used dataset with 300 datapoints



(b) Result of the spectral clustering: $k = 2$ & $\sigma = 0.08$



(a) Unsorted Similarity Graph



(b) Sorted Similarity Graph