Machine Learning HW5

0556562 陳鴻君

(All programs are implemented in Python 2.7 in Linux.)

(Videos is made by “ffmpeg”.

The usage is “ffmpeg –y –framerate 3 –i $sourgeImages $outputpath”)

1. K-means clustering

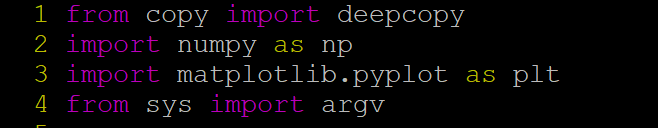
Here’s import part.

“deepcopy” is used to copy array elements in detail.

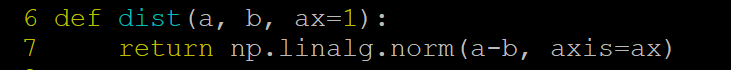
“numpy” is a powerful matrix operation library.

“matplotlib” is used to visualize data.

“argv” is input argument list.

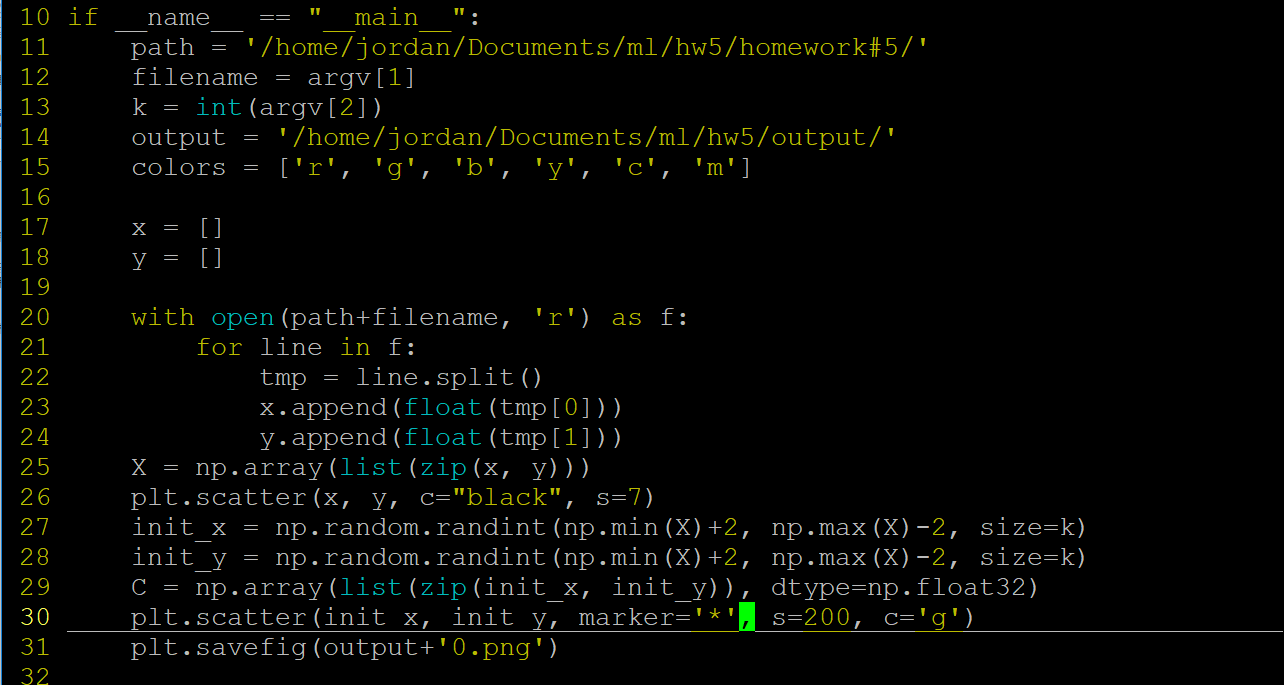


A function named “dist” is used to calculate norm from all elements in matrix a to all elements in matrix b.

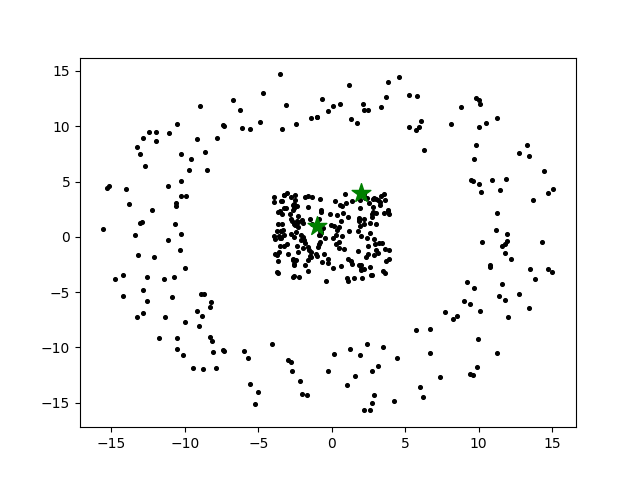


In the first part of main function, save initial in following steps:

* Read data from test file, and k from arguments.
* Initial centroid matrix C, plot and save them in a figure.

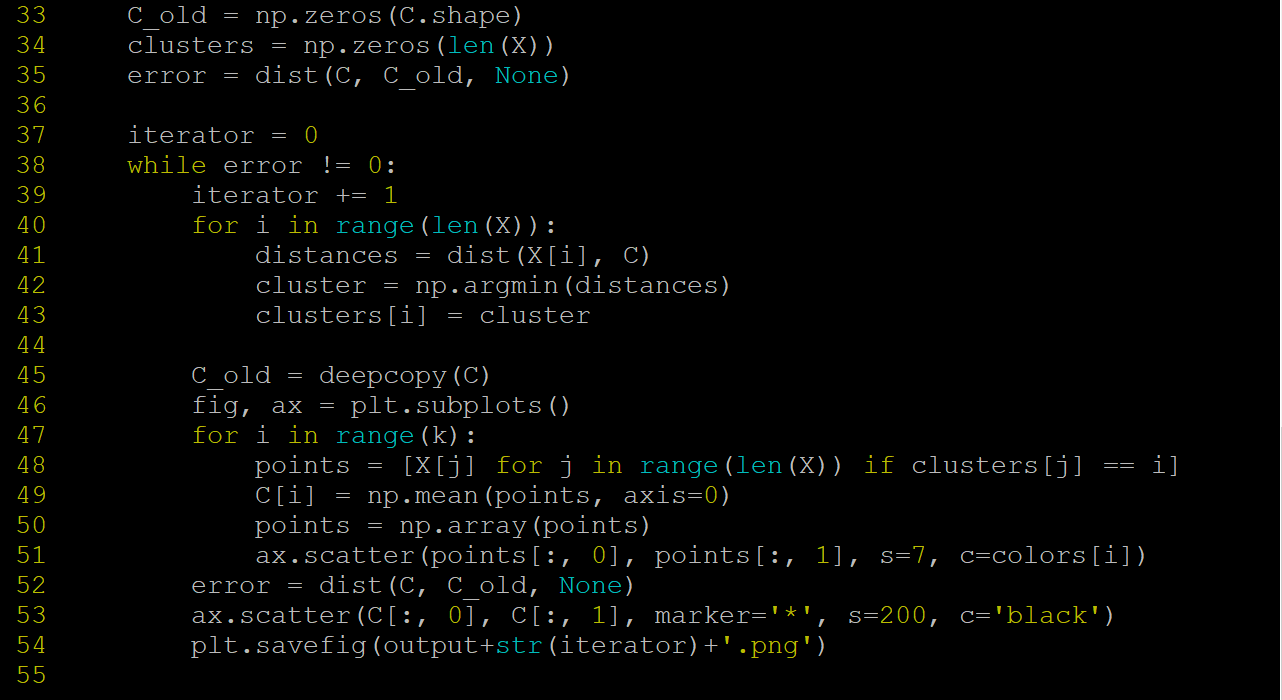


For example:

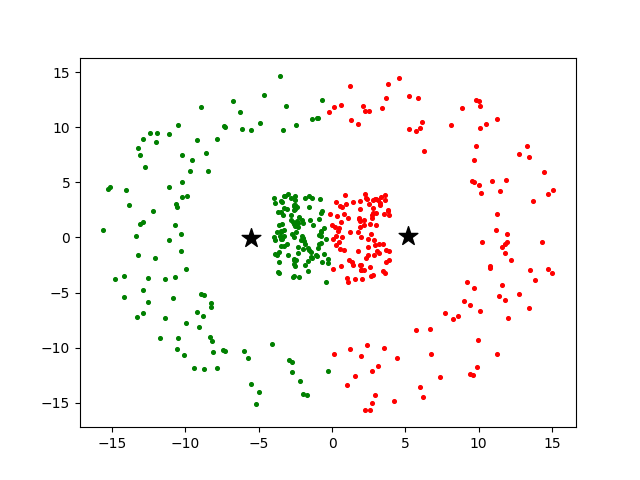


In the second part of main function, keep clustering data into k clusters until centroid fixed.

* Define “error” is the converge condition, which is the difference from this turn’s centroid to previous one.
* The data are clustered into k clusters by their distance to centroids. They will be clustered into the shortest one.
* Update centroids, by calculating the mean of all data belong to this cluster.
* Save clustering result figure in each iterator.



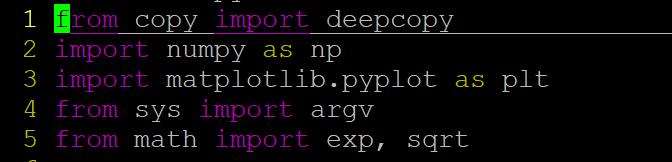
The result of test 2:



1. Kernel K-means clustering with RBF kernel

The import is similar to K-means.

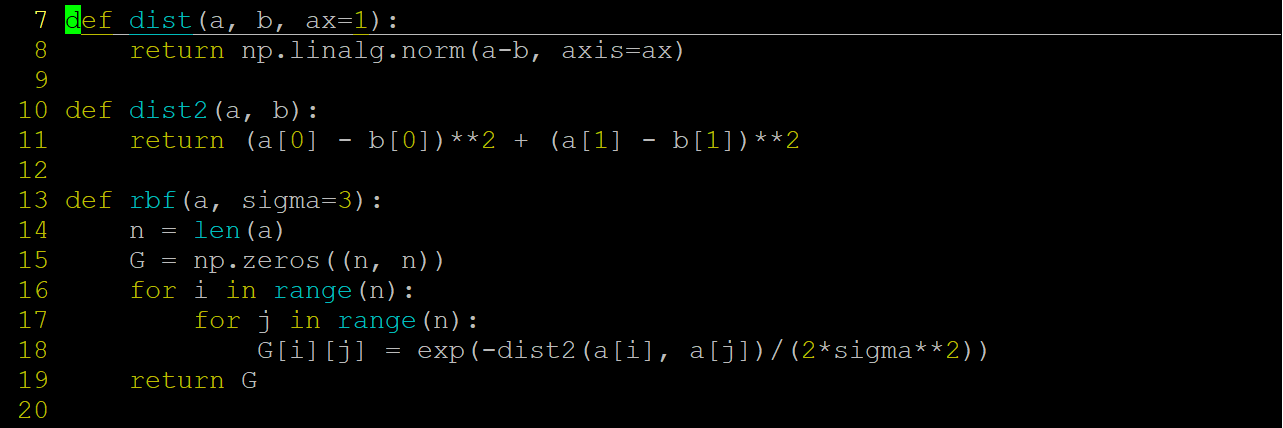
“exp” is the exponent of e. “sqrt” is square root.



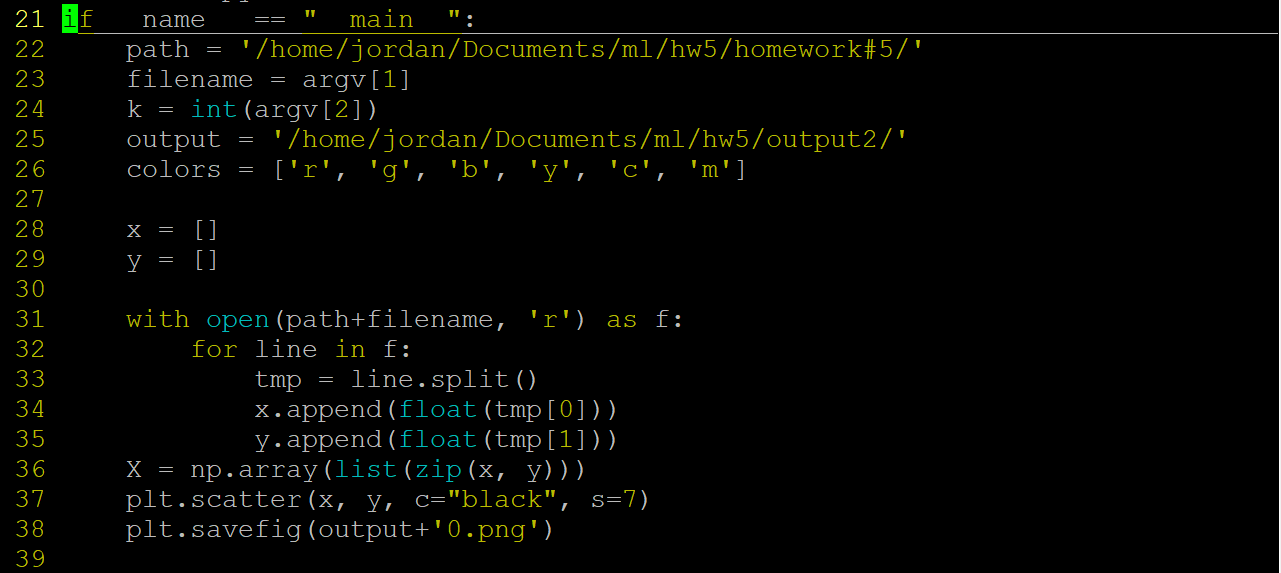
“dist” function is the same of k-means.

“dist2” function is used to calculate norm between two data points.

“rbf” function is used to calculate gram matrix G. Sigma default equals to 3.

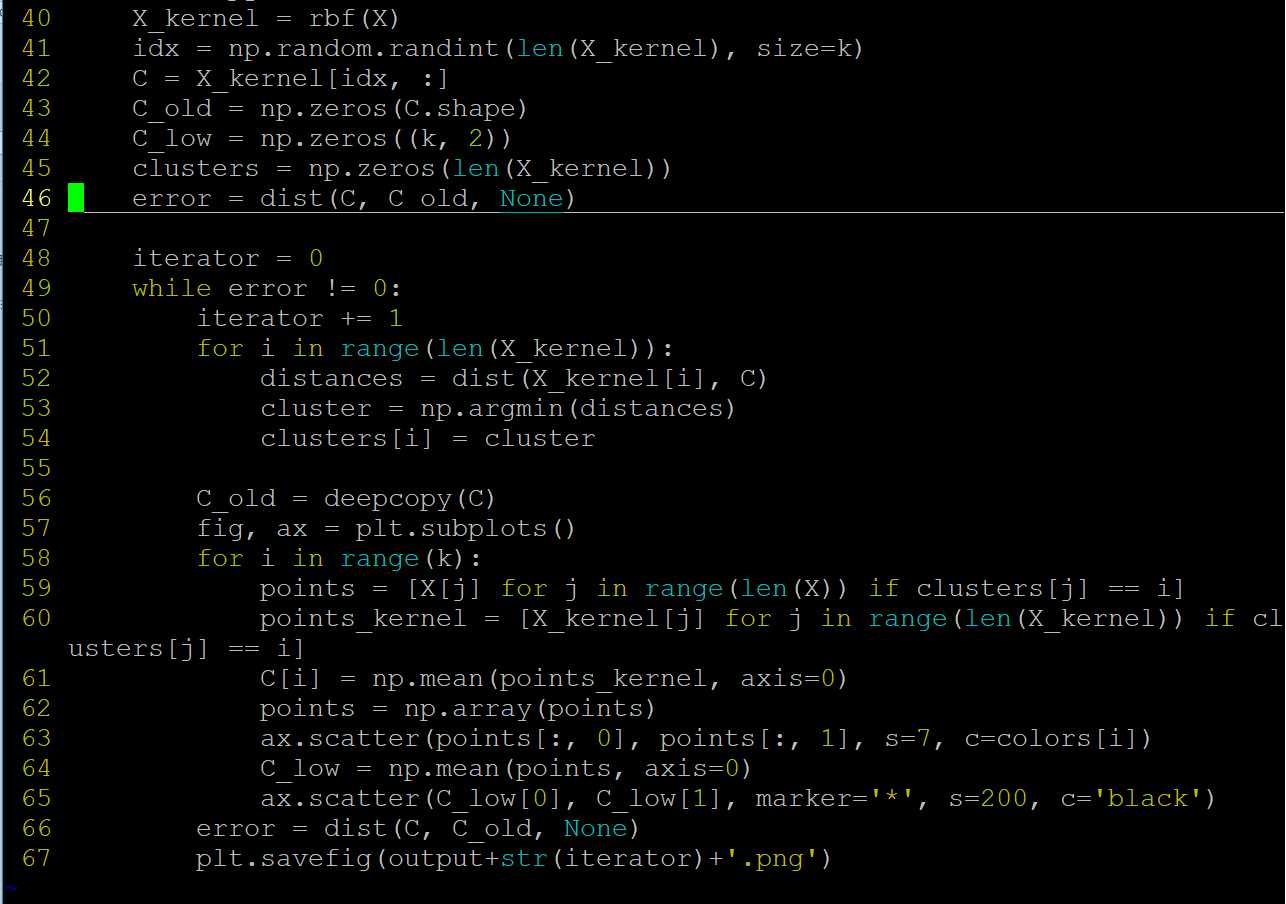


The first part of main function is same as k-means. Load and plot data.

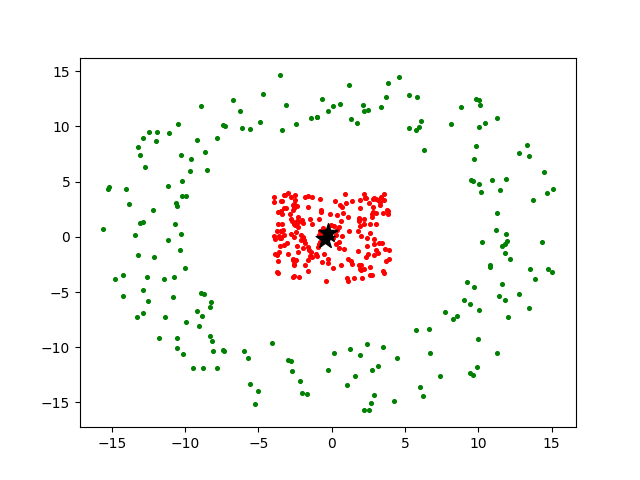


In the second part of main function, plotting data points into high dimensional space, and doing k-means with high dimension data.

* Project data into high dimensional space by RBF kernel, and randomly select k rows as initial centroids C. (C\_old is last round’s C, and it is used to calculate error. C\_low is the mean of corresponding original data points in each cluster.)
* For clustering phase, meaningful, it’s using high dimensional data to do k-means clustering. Therefore, the progress is the same.
* Plotting corresponding original data points and their centroids.



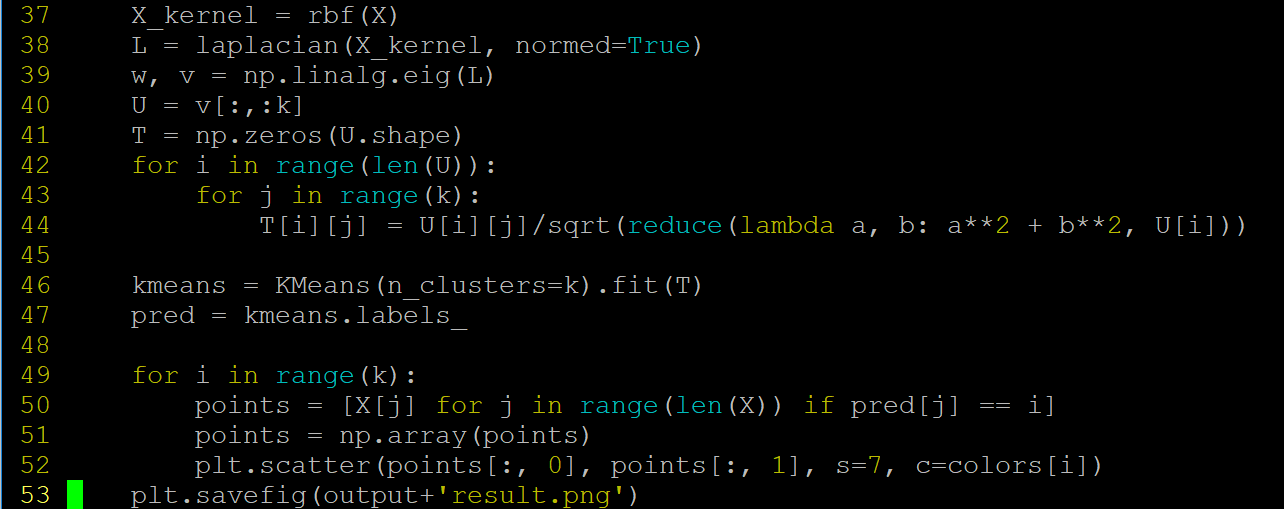
The result of test 2:



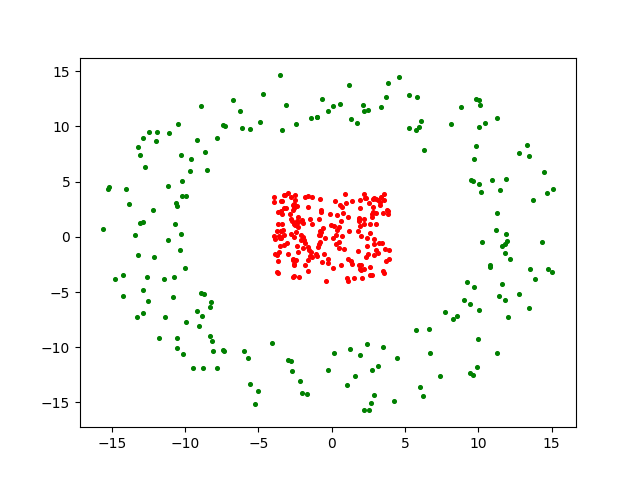
1. Spectral clustering

Import, “dist” function, and “rbf” function is skipped because they are almost the same with K-means and kernel K-means.

* Project data into high dimensional space with RBF kernel.
* Calculate normalized Laplacian matrix L.
* Eigenvalue matrix of L is called v, and select first k columns as U.
* T is the row normalized matrix of U.
* Do k-means with matrix T



The result of test 2:

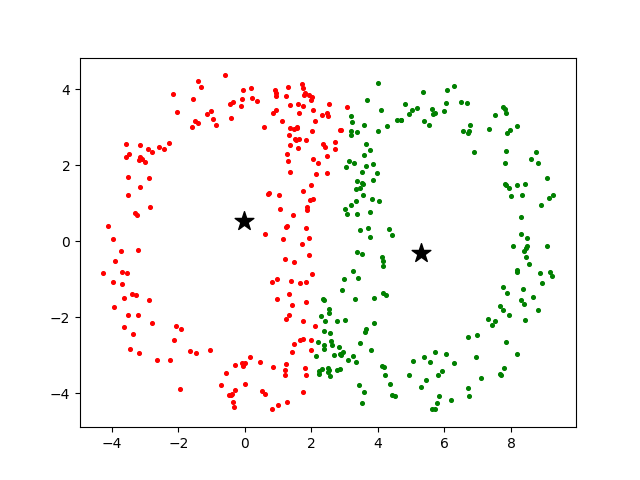


1. More clusters result

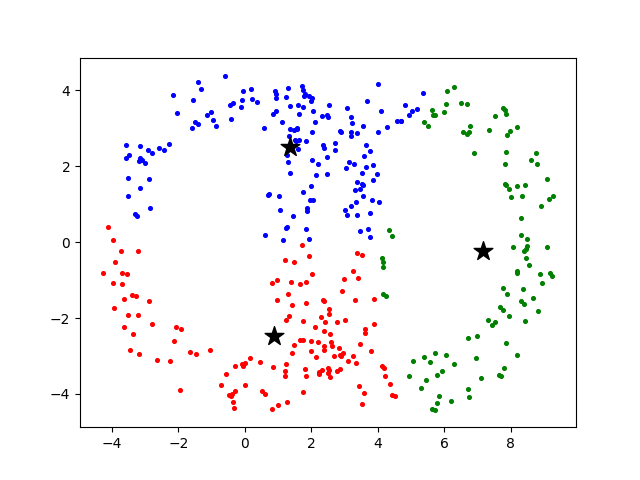
K-means:

Data = test\_1

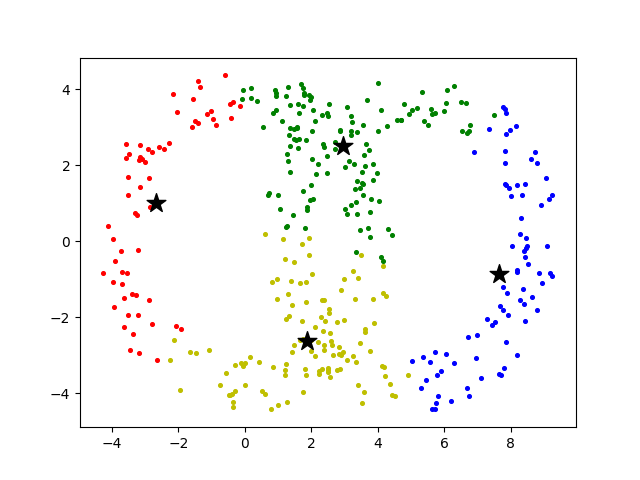
Cluster = 2



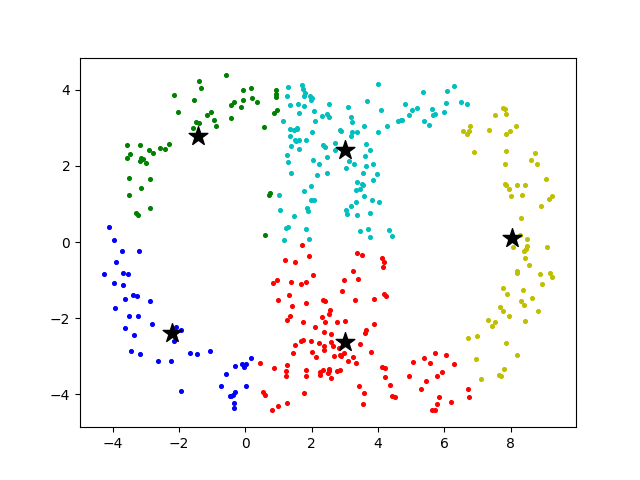
Cluster = 3



Cluster = 4

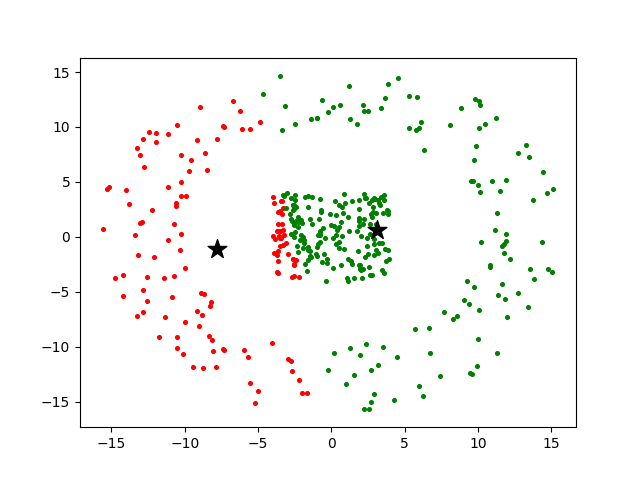


Cluster = 5

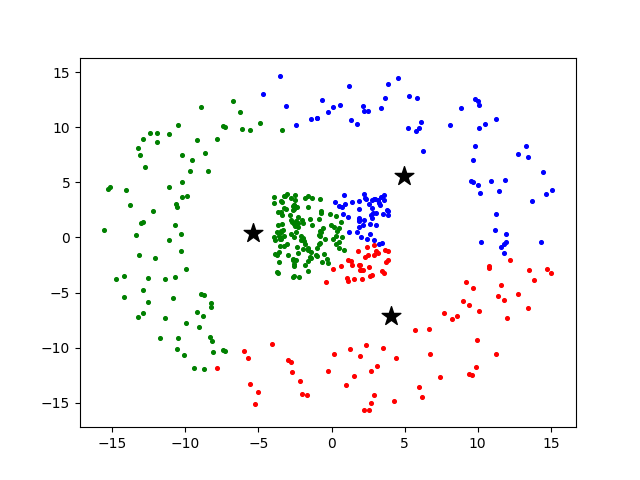


Data = test\_2

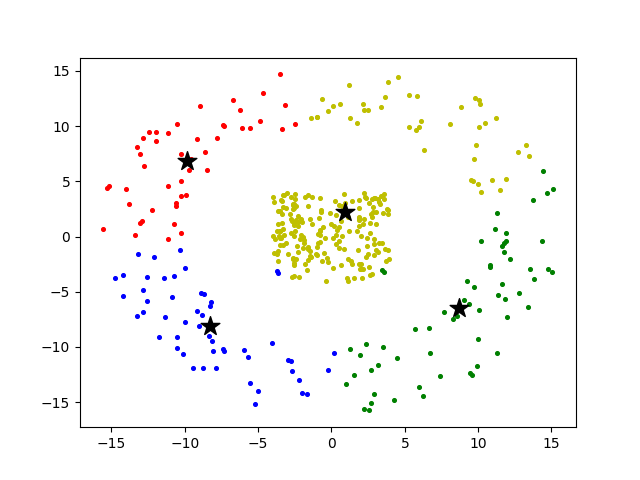
Cluster = 2



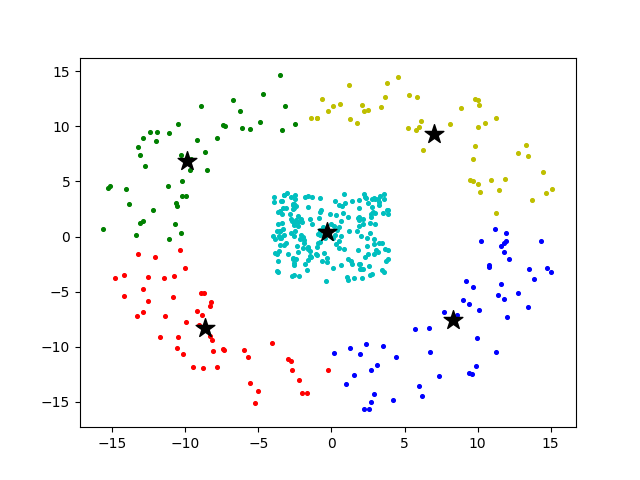
Cluster = 3



Cluster = 4



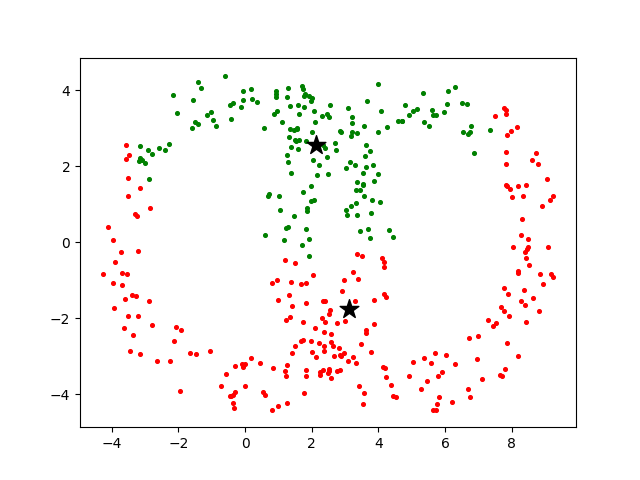
Cluster = 5



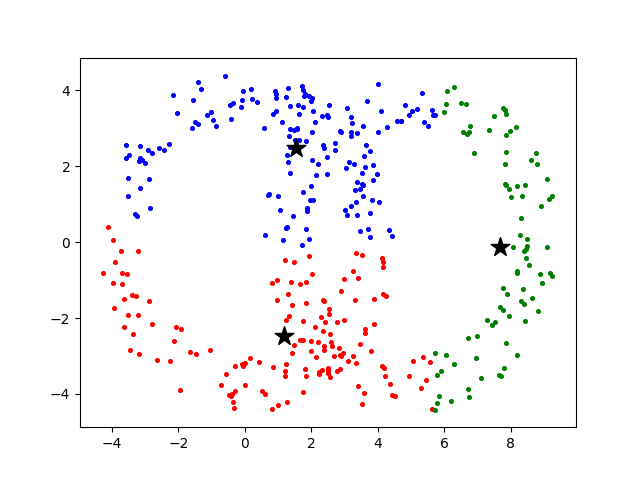
Kernel K-means:

Data = test\_1

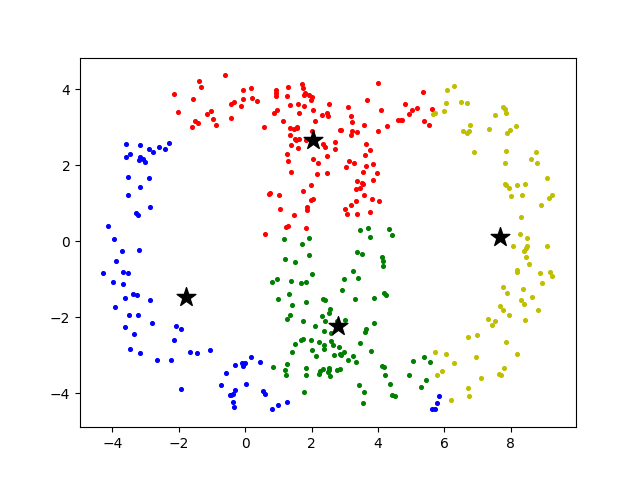
Cluster = 2



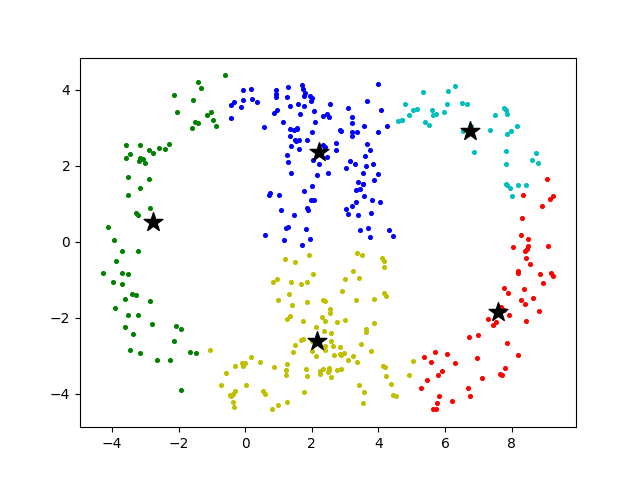
Cluster = 3



Cluster = 4

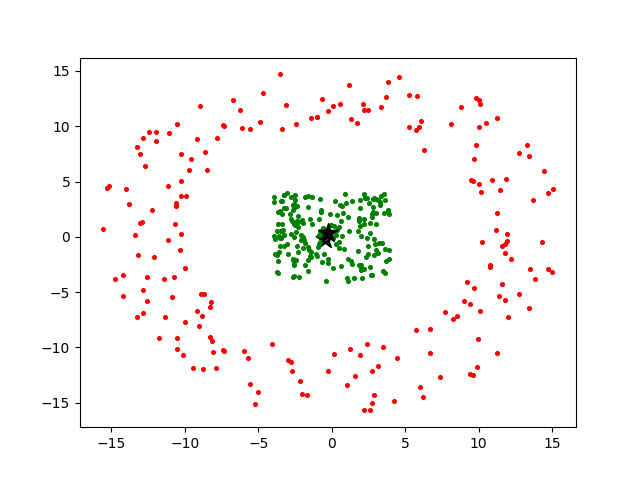


Cluster = 5

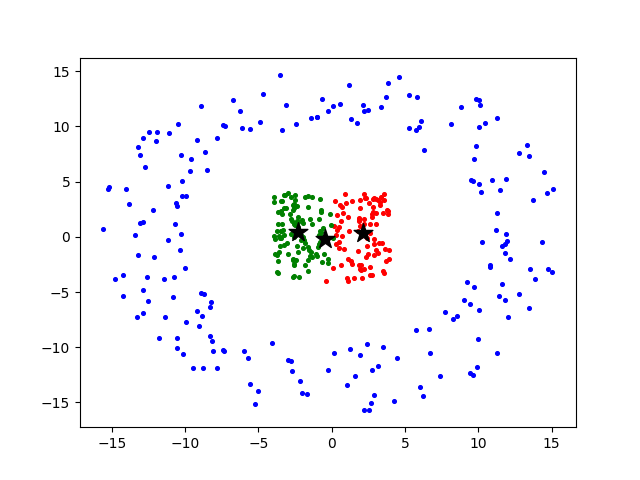


Data = test\_2

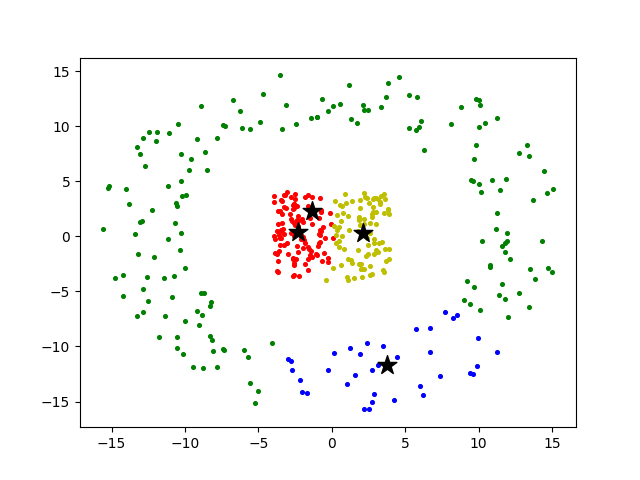
Cluster = 2



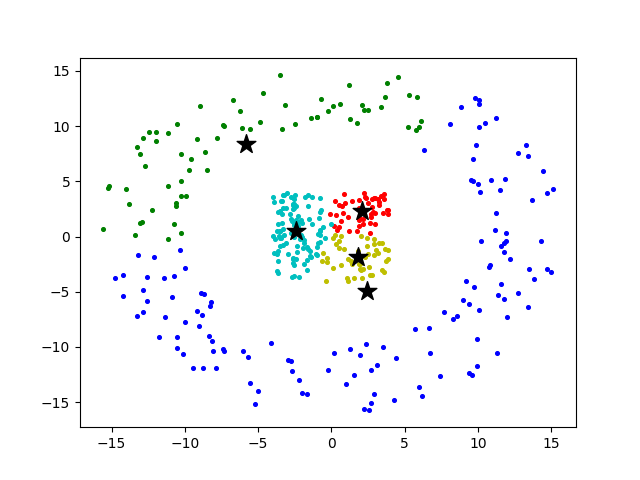
Cluster = 3



Cluster = 4



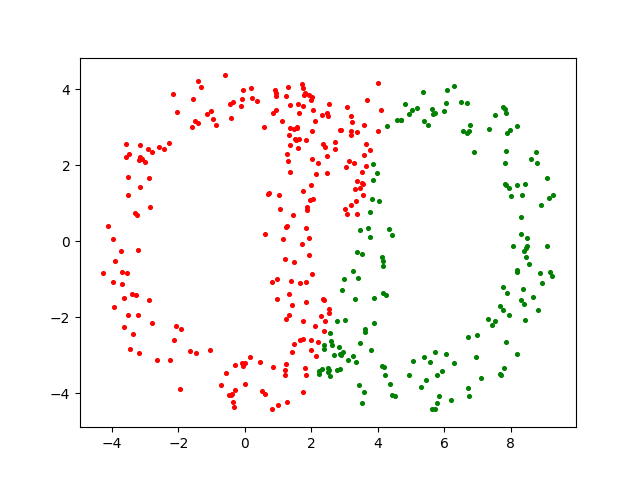
Cluster =5



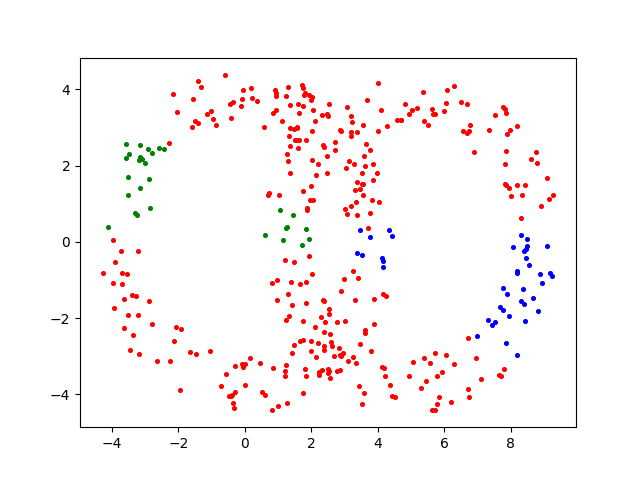
Spectral clustering:

Data = test\_1

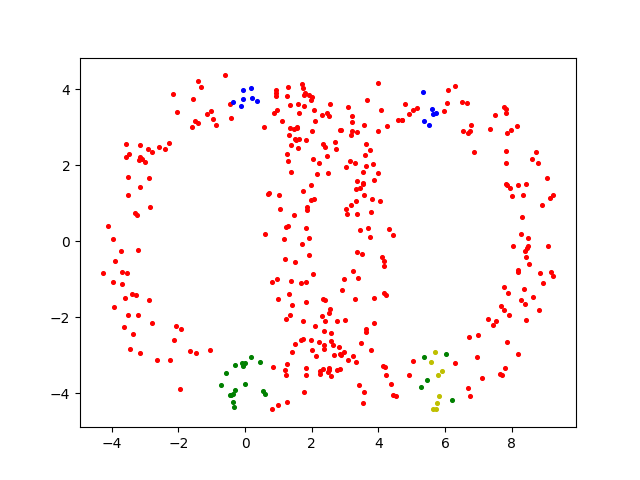
Cluster = 2



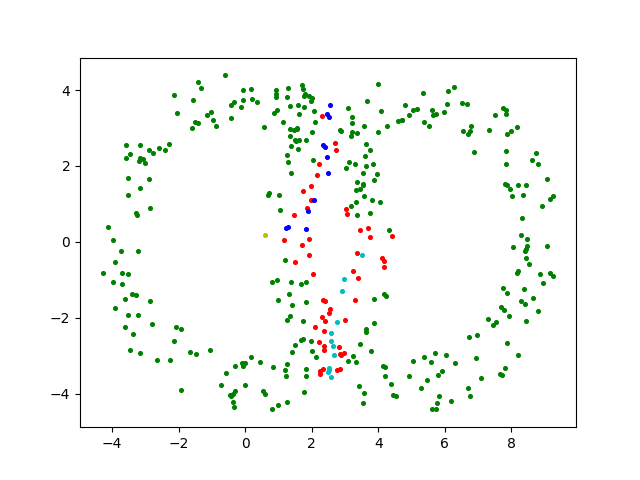
Cluster = 3



Cluster = 4

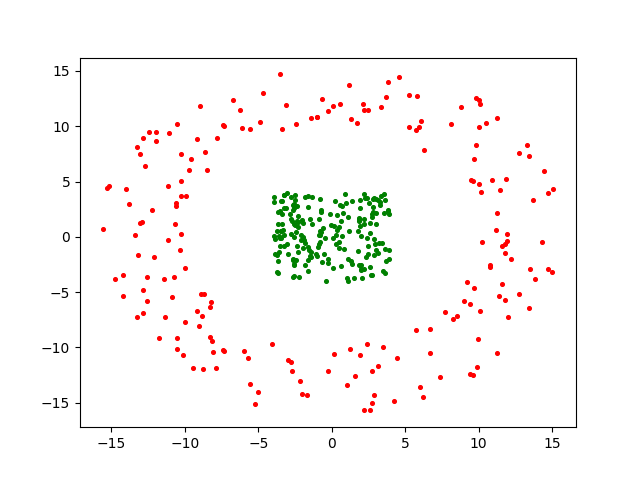


Cluster = 5

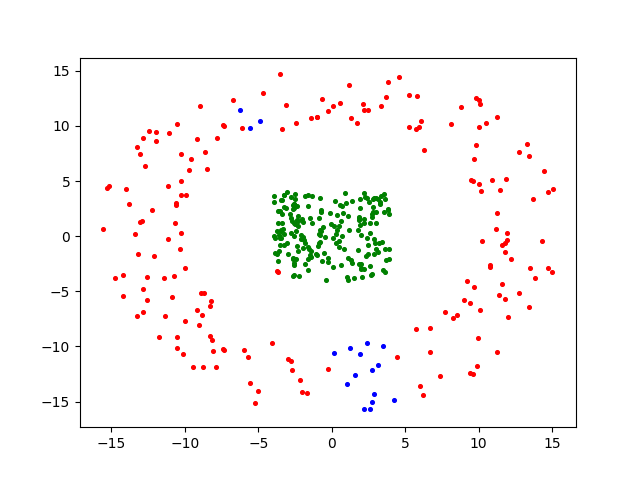


Data = test\_2

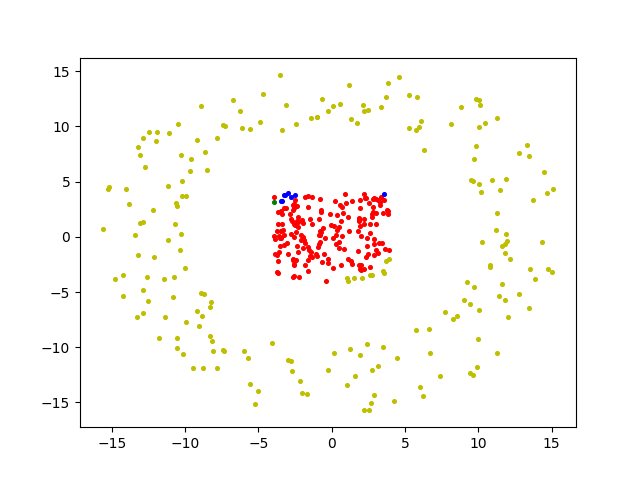
Cluster = 2



Cluster = 3



Cluster = 4



Cluster = 5

