Aristidis Chronarakis 389

06/06/18

L7- Data Mining

Homework #2

Exercise 1

The main purpose of this exercise is to measure the generalization performance for the Bagging and Random Forest classification methods for different number of classifiers that take part on the ensemble. Below is the table with the results

	Bagging (accuracy from cross validation)	**Random Forest (out-of-bag error)**
25	0.6790	0.3100
50	0.6860	0.2820
75	0.6980	0.2690
100	0.7180	0.2320

The conclusion from the results is that we are getting better performance on generalization as we increase the number of classifiers that participate on the ensemble. This is something we expected, because the ensemble techniques are used for this purpose, to give us better accuracy (generalization performance) by taking advantage of the average value approach that they use.

Compared to the results from the first set of exercises, we get better accuracy results (better generalization performance) if we use the ensemble approach, due to the above explanation.

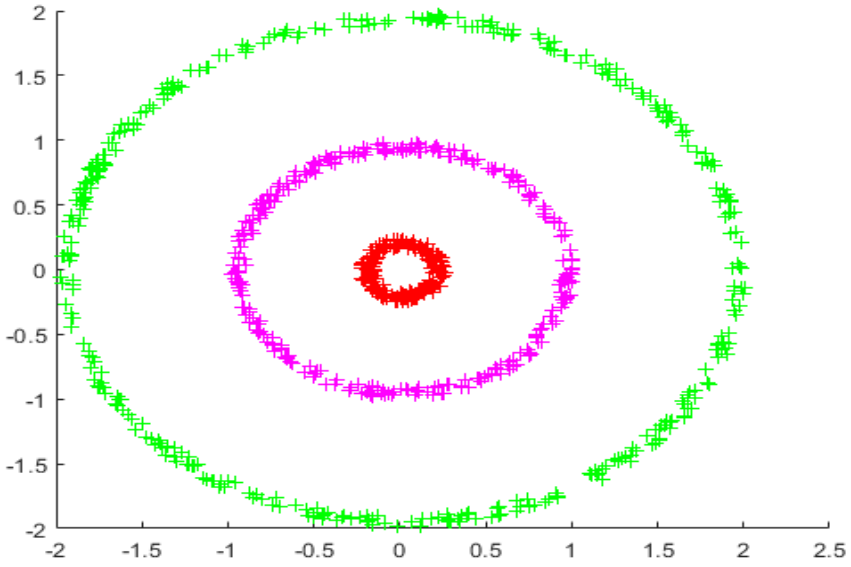
Below is the Matlab code that is used for the implementation of the exercise

```
%load data
load('AchronarakisData.mat');
rng default
NumTrees = 25;

%BAGGING
B = fitcensemble(X,classes,'Method','Bag','NumLearningCycles',NumTrees,'Kfold',10)
accuracy_cv = 1 - kfoldLoss(B)

%RANDOM FOREST (Min_Leaf=5)
RF =
TreeBagger(NumTrees,X,classes,'Method','Classification','NumPredictorsToSample','al
l','MinLeafSize',5 ,'OOBPrediction','on');
out_of_bag_error = oobError(RF)
```

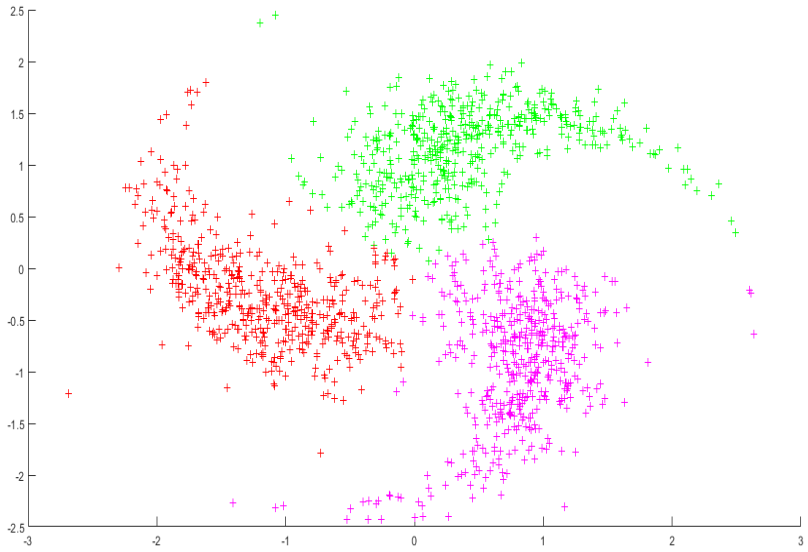
Exercise 2

The first part of this exercise asking the clustering of the 7 given datasets using the 6 different methods (k-means, agglomerative with single and average link and spectral with sigma=0.1, 0.5, and 1).

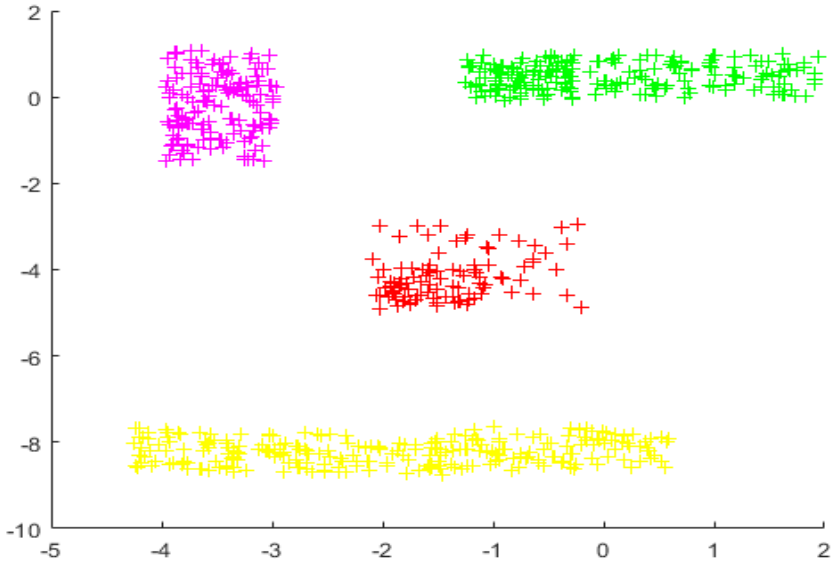
- ***Dataset*** : 3rings.mat
  Best clustering method : spectral with sigma = 0.1



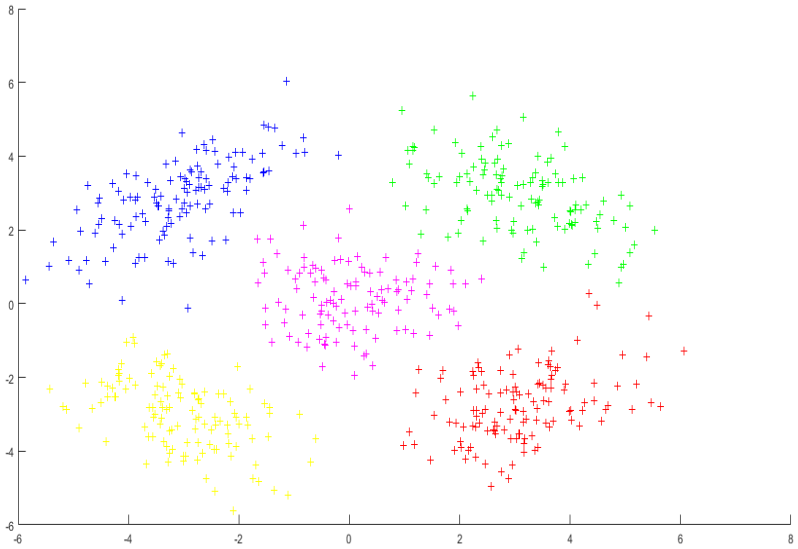
- ***Dataset*** : 3wings.mat
  Best clustering method : spectral with sigma = 0.5



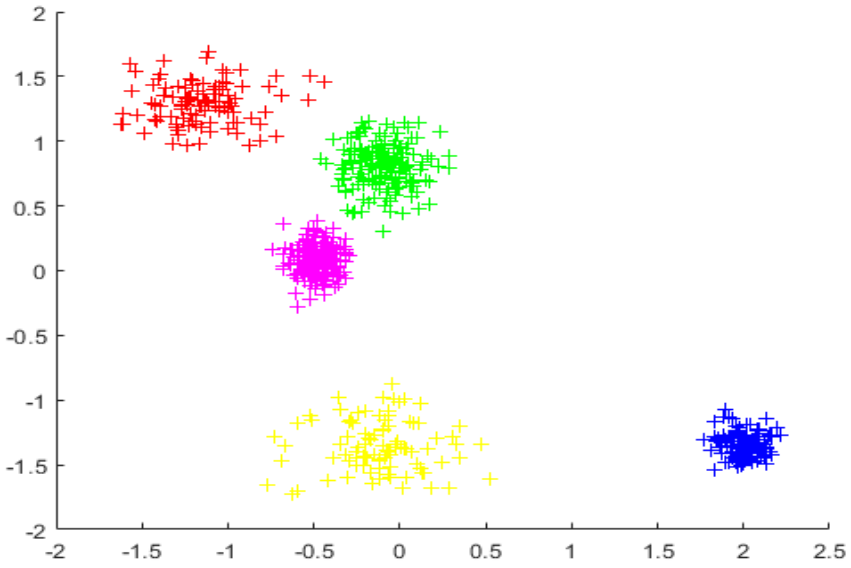
- ***Dataset*** : 4rectangles.mat
  Best clustering method : k-means, agglomerative with single link, agglomerative with average link and spectral with sigma = 0.1 give the same clustering



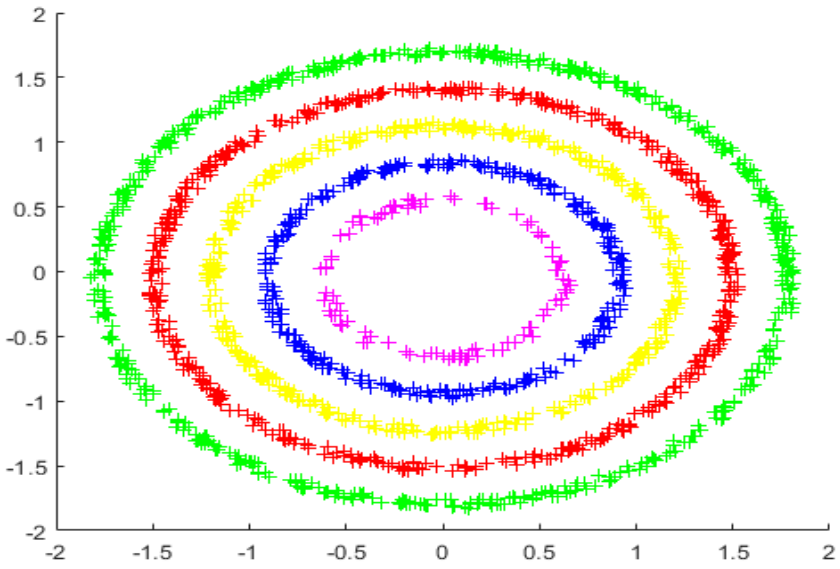
- ***Dataset*** : 5clusters.mat
  Best clustering method : k-means, agglomerative with average link, and spectral with sigma = 0.5 and sigma = 1 give good clustering results and I select as the best the agglomerative method with average link



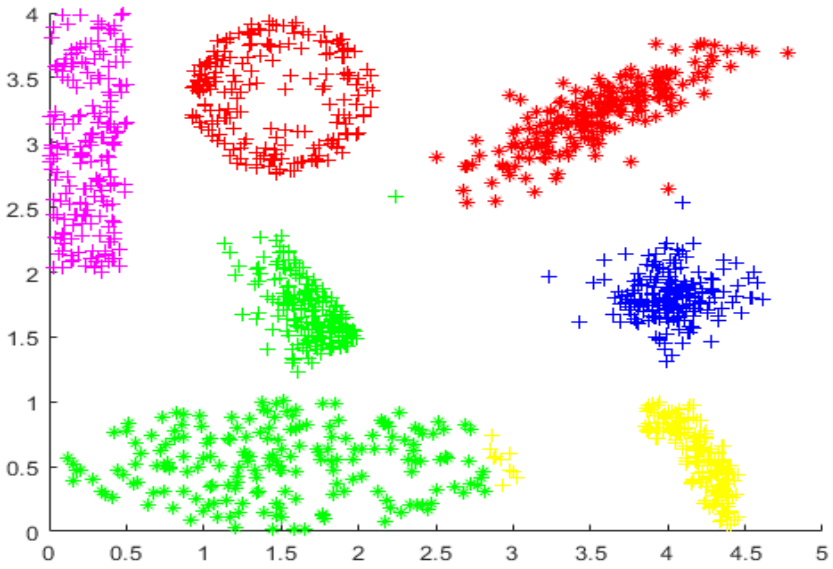
- ***Dataset*** : 5Gaussians.mat
  Best clustering method : agglomerative with average link and spectral with sigma = 1 give the same best results



- ***Dataset*** : 5rings.mat
  Best clustering method : agglomerative with single link and spectral with sigma = 0.1 give the same best results



- ***Dataset*** : 7clusters.mat
  Best clustering method : for this dataset there is no method which gives the right one clustering. The best results between the 6 methods, are given using k-means, spectral with sigma = 0.5 and and spectral with sigma = 1. I select as the best one, the k-means method.



The second part of this exercise is asking the finding of the sigma value in the range [0.1 , 0.4] which gives the best results on spectral clustering using the dataset gauss_rings which has 5 clusters.

The sigma value which gives the best clustering for the 5 clusters is 0.1. As the sigma value increases we get worse clustering. The worse clustering is reflected by the fact that the two rings are clustered on the same cluster. By increasing the sigma value the diameter which used on spectral clustering increases and as a result the two rings are clustered together.

The conclusions drawn from the results of the clustering are :

- For the data sets with the rings, the best clustering results are achieved with the agglomerative with single link and with spectral with suitable sigma value (0.1). All other clustering methods like the widely used k-means have poor clustering results for datasets with rings.

- For the remaining data sets almost all methods of classical clustering as well as methods based on spectral clustering can be used. More specifically k-means clustering method can be used to give good results on all remaining data sets except the Gaussian. Between agglomerative methods, the method with average link gives better clustering results on all remaining data sets except on data set with rectangles in which both the agglomerative methods achieve the same good results. Spectral clustering method can be used to give good clustering results if the sigma value of the method has been tuned on the suitable value for that data set.

The spectral clustering algorithm does not give the same solution every time. This is due to the fact that, on the last stage when the clustering of the lines of the matrix formed by the eigenvectors (spectral analysis of the similarity matrix) take place, the method which is used for the clustering is the k-means, a method whose results are affected by the initial placement of the k centers of the clusters.

The third part of this exercise is asking for an evaluation of the optimal number of clusters on datasets (not included datasets with rings) using k-means and silhouette as evaluation's criterion.

Dataset	Real number of clusters given from evaluation
3wings.mat	3
4rectangles.mat	5
5clusters.mat	5
5Gaussians.mat	5
7clusters.mat	8

The silhouette criterion measures how well data lie within each cluster. For the datasets 4rectangles and 7clusters the evaluation from the above method differs from the number of clusters that we know that is the real number. More specifically, it gives as optimal number of clusters a number which is bigger than the real one by one, but the evaluation's values for these two numbers of clusters are very close. Thus, when we have close quality evaluation values, we can select the number of clusters which we consider as optimal.

Below is the Matlab code that is used for the implementation of the exercise

```
%load data
load('3rings.mat');
NoC=3;

%K-MEANS
idkm = kmeans(X,NoC);
plot_max10_clusters(X,idkm)

%AGGLOMERATIVE (SINGLE LINK)
idags = clusterdata(X,'Linkage','single','Maxclust',NoC);
plot_max10_clusters(X,idags)

%AGGLOMERATIVE (AVERAGE LINK)
idaga = clusterdata(X,'Linkage','average','Maxclust',NoC);
plot_max10_clusters(X,idaga)

%SPECTRAL CLUSTERING WITH RBF KERNEL (SIGMA = 0.1)
spectral(X,NoC,0.1);

%SPECTRAL CLUSTERING WITH RBF KERNEL (SIGMA = 0.5)
spectral(X,NoC,0.5);

%SPECTRAL CLUSTERING WITH RBF KERNEL (SIGMA = 1)
spectral(X,NoC,1);

%ESTIMATE THE NUMBER OF CLUSTERS
evaNoC = evalclusters(X,'kmeans','silhouette','KList',[1:10]);
```