ML ASSIGNMENT 4

(All Errors are reported on a scale of 100)

Problem 1: PCA and Feature Selection

SVMs and PCA

1) Refer to the file **PCA.py**

   Top 6 eigen values of the data covariance matrix are [8.94757052, 9.68540952, 10.71620451, 16.41583154, 36.76148171, 63.56349684]

2) Refer to the file **SVMwithPCA.py**

Errors of the learned Classifier on the Validation set for each pair of K/C value

K/C	C = 1	C = 10	C = 100	C = 1000
K = 1	40.384615	40.384615	40.384615	40.384615
K = 2	40.384615	40.384615	40.384615	40.384615
K = 3	38.461538	38.461538	38.461538	38.461538
K = 4	40.384615	46.153846	46.153846	46.153846
K = 5	23.07692	26.923076	19.230769	19.2307692
K = 6	23.076923	26.9230769	26.923076	28.846153

The best pair of k and c value is 5 and 100

3) Refer to the file **SVMwithPCA.py**

Error of the best k/c pair on the test data is 23.076923

Error on the best classifier without feature selection is 15.384615

The observations that can be deduced from above values are as follows

1. Although only the top 5 features(K = 5) are taken, the accuracy achieved is more which is 77% that is slightly less than the accuracy achieved by the best classifier of SVM without feature selection.

2. If some more features are taken into account, then the accuracy achieved will be nearly equal to that of achieved by classifier without feature selection.

4)
The heuristic that can be used to select the value of K before evaluating the performance on Validation Set is

- Taking the first k eigenvectors that capture at least 85% of the total variance.(which mean selecting only the best K features that covers whole data)

PCA for Feature Selection

1. Naive bayes without feature selection (Refer to **naiveBayes.py**)
   Accuracy on the test set of the Gaussian naive bayes classifier is 69.23076923076923.

2. Naive bayes with feature selection (Refer to **PCA_FeatureSel.py**)

- π defines a probability distribution as the sum of the probabilties in that distribution is equal to one. Each value in the list corresponds to the probability of selecting that feature and it corresponds to the summation of squares of eigen vector value corresponding to that feature.

- Average test error over 100 iterations

Error on test data for value of k = 1 and s = 1 is 44.019230769230774
Error on test data for value of k = 1 and s = 2 is 43.24999999999999
Error on test data for value of k = 1 and s = 3 is 42.634615384615365
Error on test data for value of k = 1 and s = 4 is 43.115384615384585
Error on test data for value of k = 1 and s = 5 is 42.36538461538463
Error on test data for value of k = 1 and s = 6 is 42.423076923076906
Error on test data for value of k = 1 and s = 7 is 41.96153846153847
Error on test data for value of k = 1 and s = 8 is 42.86538461538461
Error on test data for value of k = 1 and s = 9 is 41.88461538461539
Error on test data for value of k = 1 and s = 10 is 42.82692307692306
Error on test data for value of k = 1 and s = 11 is 42.21153846153846
Error on test data for value of k = 1 and s = 12 is 42.365384615384585
Error on test data for value of k = 1 and s = 13 is 42.519230769230774
Error on test data for value of k = 1 and s = 14 is 42.46153846153844
Error on test data for value of k = 1 and s = 15 is 42.01923076923074
Error on test data for value of k = 1 and s = 16 is 42.788461538461526
Error on test data for value of k = 1 and s = 17 is 41.51923076923076
Error on test data for value of k = 1 and s = 18 is 43.32692307692306
Error on test data for value of k = 1 and s = 19 is 42.57692307692306
Error on test data for value of k = 1 and s = 20 is 42.46153846153846

Error on test data for value of k = 2 and s = 1 is 44.11538461538462
Error on test data for value of k = 2 and s = 2 is 42.53846153846153
Error on test data for value of k = 2 and s = 3 is 42.403846153846146
Error on test data for value of k = 2 and s = 4 is 42.26923076923075
Error on test data for value of k = 2 and s = 5 is 41.73076923076922
Error on test data for value of k = 2 and s = 6 is 41.17307692307693
Error on test data for value of k = 2 and s = 7 is 41.51923076923076
Error on test data for value of k = 2 and s = 8 is 41.519230769230745
Error on test data for value of k = 2 and s = 9 is 39.61538461538462

Error on test data for value of k = 2 and s = 10 is 40.269230769230774
Error on test data for value of k = 2 and s = 11 is 40.192307692307665
Error on test data for value of k = 2 and s = 12 is 40.17307692307691
Error on test data for value of k = 2 and s = 13 is 40.84615384615382
Error on test data for value of k = 2 and s = 14 is 40.269230769230745
Error on test data for value of k = 2 and s = 15 is 40.19230769230768
Error on test data for value of k = 2 and s = 16 is 40.11538461538459
Error on test data for value of k = 2 and s = 17 is 40.134615384615365
Error on test data for value of k = 2 and s = 18 is 40.730769230769226
Error on test data for value of k = 2 and s = 19 is 40.961538461538446
Error on test data for value of k = 2 and s = 20 is 40.61538461538461

Error on test data for value of k = 3 and s = 1 is 44.519230769230774
Error on test data for value of k = 3 and s = 2 is 42.2307692307692
Error on test data for value of k = 3 and s = 3 is 41.55769230769229
Error on test data for value of k = 3 and s = 4 is 41.730769230769255
Error on test data for value of k = 3 and s = 5 is 41.82692307692306
Error on test data for value of k = 3 and s = 6 is 41.84615384615384
Error on test data for value of k = 3 and s = 7 is 40.59615384615379
Error on test data for value of k = 3 and s = 8 is 41.903846153846125
Error on test data for value of k = 3 and s = 9 is 40.903846153846146
Error on test data for value of k = 3 and s = 10 is 39.98076923076919
Error on test data for value of k = 3 and s = 11 is 40.59615384615384
Error on test data for value of k = 3 and s = 12 is 39.96153846153843
Error on test data for value of k = 3 and s = 13 is 40.17307692307691
Error on test data for value of k = 3 and s = 14 is 40.692307692307686
Error on test data for value of k = 3 and s = 15 is 40.307692307692314
Error on test data for value of k = 3 and s = 16 is 40.23076923076924
Error on test data for value of k = 3 and s = 17 is 40.55769230769231
Error on test data for value of k = 3 and s = 18 is 41.30769230769227
Error on test data for value of k = 3 and s = 19 is 41.13461538461535
Error on test data for value of k = 3 and s = 20 is 40.50000000000001

Error on test data for value of k = 4 and s = 1 is 44.423076923076934
Error on test data for value of k = 4 and s = 2 is 42.63461538461538
Error on test data for value of k = 4 and s = 3 is 42.51923076923075
Error on test data for value of k = 4 and s = 4 is 41.94230769230768
Error on test data for value of k = 4 and s = 5 is 42.211538461538446
Error on test data for value of k = 4 and s = 6 is 42.307692307692314
Error on test data for value of k = 4 and s = 7 is 40.55769230769229
Error on test data for value of k = 4 and s = 8 is 41.09615384615382
Error on test data for value of k = 4 and s = 9 is 41.30769230769229
Error on test data for value of k = 4 and s = 10 is 42.05769230769228
Error on test data for value of k = 4 and s = 11 is 40.69230769230768
Error on test data for value of k = 4 and s = 12 is 41.67307692307691
Error on test data for value of k = 4 and s = 13 is 42.05769230769229
Error on test data for value of k = 4 and s = 14 is 40.86538461538461
Error on test data for value of k = 4 and s = 15 is 41.07692307692307
Error on test data for value of k = 4 and s = 16 is 40.26923076923074
Error on test data for value of k = 4 and s = 17 is 40.019230769230745
Error on test data for value of k = 4 and s = 18 is 40.807692307692285
Error on test data for value of k = 4 and s = 19 is 42.269230769230745
Error on test data for value of k = 4 and s = 20 is 42.46153846153843

Error on test data for value of k = 5 and s = 1 is 44.94230769230768
Error on test data for value of k = 5 and s = 2 is 40.13461538461538
Error on test data for value of k = 5 and s = 3 is 37.461538461538474
Error on test data for value of k = 5 and s = 4 is 38.50000000000002
Error on test data for value of k = 5 and s = 5 is 36.75000000000002
Error on test data for value of k = 5 and s = 6 is 35.903846153846175
Error on test data for value of k = 5 and s = 7 is 34.98076923076924
Error on test data for value of k = 5 and s = 8 is 33.55769230769232
Error on test data for value of k = 5 and s = 9 is 34.019230769230774

Error on test data for value of k = 5 and s = 10 is 32.78846153846156
Error on test data for value of k = 5 and s = 11 is 33.173076923076934
Error on test data for value of k = 5 and s = 12 is 31.884615384615373
Error on test data for value of k = 5 and s = 13 is 33.30769230769229
Error on test data for value of k = 5 and s = 14 is 32.09615384615387
Error on test data for value of k = 5 and s = 15 is 32.05769230769232
Error on test data for value of k = 5 and s = 16 is 31.634615384615387
Error on test data for value of k = 5 and s = 17 is 32.01923076923079
Error on test data for value of k = 5 and s = 18 is 31.28846153846156
Error on test data for value of k = 5 and s = 19 is 30.846153846153854
Error on test data for value of k = 5 and s = 20 is 31.249999999999986

Error on test data for value of k = 6 and s = 1 is 44.961538461538446
Error on test data for value of k = 6 and s = 2 is 41.84615384615382
Error on test data for value of k = 6 and s = 3 is 40.21153846153846
Error on test data for value of k = 6 and s = 4 is 37.46153846153848
Error on test data for value of k = 6 and s = 5 is 37.057692307692314
Error on test data for value of k = 6 and s = 6 is 36.038461538461526
Error on test data for value of k = 6 and s = 7 is 35.500000000000014
Error on test data for value of k = 6 and s = 8 is 34.86538461538464
Error on test data for value of k = 6 and s = 9 is 35.80769230769229
Error on test data for value of k = 6 and s = 10 is 34.423076923076934
Error on test data for value of k = 6 and s = 11 is 32.92307692307695
Error on test data for value of k = 6 and s = 12 is 34.230769230769226
Error on test data for value of k = 6 and s = 13 is 32.05769230769232
Error on test data for value of k = 6 and s = 14 is 32.15384615384613
Error on test data for value of k = 6 and s = 15 is 32.423076923076934
Error on test data for value of k = 6 and s = 16 is 32.48076923076924
Error on test data for value of k = 6 and s = 17 is 32.4807692307692
Error on test data for value of k = 6 and s = 18 is 31.884615384615415
Error on test data for value of k = 6 and s = 19 is 31.211538461538453
Error on test data for value of k = 6 and s = 20 is 31.846153846153854

Error on test data for value of k = 7 and s = 1 is 42.67307692307691
Error on test data for value of k = 7 and s = 2 is 41.461538461538495
Error on test data for value of k = 7 and s = 3 is 39.50000000000003
Error on test data for value of k = 7 and s = 4 is 36.74999999999998
Error on test data for value of k = 7 and s = 5 is 36.19230769230771
Error on test data for value of k = 7 and s = 6 is 35.34615384615387
Error on test data for value of k = 7 and s = 7 is 34.90384615384613
Error on test data for value of k = 7 and s = 8 is 34.980769230769226
Error on test data for value of k = 7 and s = 9 is 34.423076923076934
Error on test data for value of k = 7 and s = 10 is 32.19230769230769
Error on test data for value of k = 7 and s = 11 is 32.28846153846152
Error on test data for value of k = 7 and s = 12 is 32.07692307692311
Error on test data for value of k = 7 and s = 13 is 31.346153846153854
Error on test data for value of k = 7 and s = 14 is 32.46153846153844
Error on test data for value of k = 7 and s = 15 is 32.92307692307689
Error on test data for value of k = 7 and s = 16 is 31.749999999999986
Error on test data for value of k = 7 and s = 17 is 30.980769230769255
Error on test data for value of k = 7 and s = 18 is 30.826923076923094
Error on test data for value of k = 7 and s = 19 is 32.269230769230774
Error on test data for value of k = 7 and s = 20 is 31.346153846153854

Error on test data for value of k = 8 and s = 1 is 44.038461538461526
Error on test data for value of k = 8 and s = 2 is 40.44230769230769
Error on test data for value of k = 8 and s = 3 is 37.46153846153847
Error on test data for value of k = 8 and s = 4 is 36.46153846153846
Error on test data for value of k = 8 and s = 5 is 36.98076923076922
Error on test data for value of k = 8 and s = 6 is 35.76923076923079
Error on test data for value of k = 8 and s = 7 is 35.673076923076934
Error on test data for value of k = 8 and s = 8 is 33.826923076923066
Error on test data for value of k = 8 and s = 9 is 34.40384615384616

Error on test data for value of k = 8 and s = 10 is 33.88461538461536
Error on test data for value of k = 8 and s = 11 is 33.653846153846175
Error on test data for value of k = 8 and s = 12 is 32.903846153846175
Error on test data for value of k = 8 and s = 13 is 33.115384615384585
Error on test data for value of k = 8 and s = 14 is 31.961538461538453
Error on test data for value of k = 8 and s = 15 is 33.80769230769231
Error on test data for value of k = 8 and s = 16 is 31.80769230769232
Error on test data for value of k = 8 and s = 17 is 31.615384615384 64
Error on test data for value of k = 8 and s = 18 is 31.134615384615415
Error on test data for value of k = 8 and s = 19 is 31.019230769230802
Error on test data for value of k = 8 and s = 20 is 32.1346153846154

Error on test data for value of k = 9 and s = 1 is 43.78846153846157
Error on test data for value of k = 9 and s = 2 is 41.30769230769231
Error on test data for value of k = 9 and s = 3 is 39.73076923076924
Error on test data for value of k = 9 and s = 4 is 39.211538461538495
Error on test data for value of k = 9 and s = 5 is 35.98076923076924
Error on test data for value of k = 9 and s = 6 is 35.500000000000014
Error on test data for value of k = 9 and s = 7 is 34.51923076923079
Error on test data for value of k = 9 and s = 8 is 35.115384615384585
Error on test data for value of k = 9 and s = 9 is 32.80769230769229
Error on test data for value of k = 9 and s = 10 is 32.63461538461539
Error on test data for value of k = 9 and s = 11 is 33.30769230769231
Error on test data for value of k = 9 and s = 12 is 31.538461538461547
Error on test data for value of k = 9 and s = 13 is 31.211538461538467
Error on test data for value of k = 9 and s = 14 is 31.615384615384 64
Error on test data for value of k = 9 and s = 15 is 32.46153846153848
Error on test data for value of k = 9 and s = 16 is 30.769230769230788
Error on test data for value of k = 9 and s = 17 is 30.865384615384627
Error on test data for value of k = 9 and s = 18 is 30.07692307692308
Error on test data for value of k = 9 and s = 19 is 30.961538461538453
Error on test data for value of k = 9 and s = 20 is 32.36538461538464

Error on test data for value of k = 10 and s = 1 is 42.442307692307715
Error on test data for value of k = 10 and s = 2 is 41.94230769230771
Error on test data for value of k = 10 and s = 3 is 39.57692307692306
Error on test data for value of k = 10 and s = 4 is 38.346153846153854
Error on test data for value of k = 10 and s = 5 is 35.230769230769226
Error on test data for value of k = 10 and s = 6 is 34.88461538461539
Error on test data for value of k = 10 and s = 7 is 33.596153846153854
Error on test data for value of k = 10 and s = 8 is 34.92307692307692
Error on test data for value of k = 10 and s = 9 is 33.28846153846153
Error on test data for value of k = 10 and s = 10 is 33.903846153846175
Error on test data for value of k = 10 and s = 11 is 32.59615384615388
Error on test data for value of k = 10 and s = 12 is 32.05769230769229
Error on test data for value of k = 10 and s = 13 is 32.30769230769232
Error on test data for value of k = 10 and s = 14 is 32.13461538461539
Error on test data for value of k = 10 and s = 15 is 32.00000000000003
Error on test data for value of k = 10 and s = 16 is 31.53846153846156
Error on test data for value of k = 10 and s = 17 is 33.19230769230771
Error on test data for value of k = 10 and s = 18 is 31.65384615384619
Error on test data for value of k = 10 and s = 19 is 32.50000000000003
Error on test data for value of k = 10 and s = 20 is 30.82692307692308

- **Pros and Cons**
  Yes, this provides a reasonable alternative solution as the accuracy achieved is equal to that of achieved by naive bayes without feature selection and the computation required is less.

Pros

1. With Less Computation time that is selecting only the best features gives high accuracy.

Cons

1. If all the features are equally important, then whole data needs to be taken into account where sampling does not help.
2. For finding the best value of K and S, more computation time is needed needs to be done iterating over different values of K and S.

Problem 2: Spectral Clustering

The Basic Algorithm (Refer to **Spectral_Algo.py)**

(T) corresponds to transpose

A matrix L is defined as positive semi definite if v(T)Lv >= 0

Consider Laplacian Matrix L
As L = D-A
Where D is the Diagonal Matrix
And A is the similarity Matrix(Square of distance between 2 input vectors)

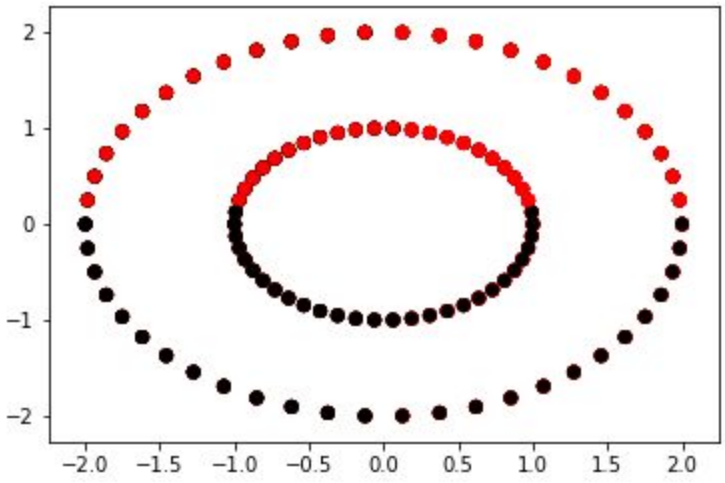
L = M(T)M

Lambda = v(T)Lv Lambda corresponds to eigen values
= v(T)M(T)Mv
= (Mv)(T)Mv
Lambda is the inner product of vector Mv with itself which is positive.
As Lamda is positive, Matrix L is positive semidefinite.

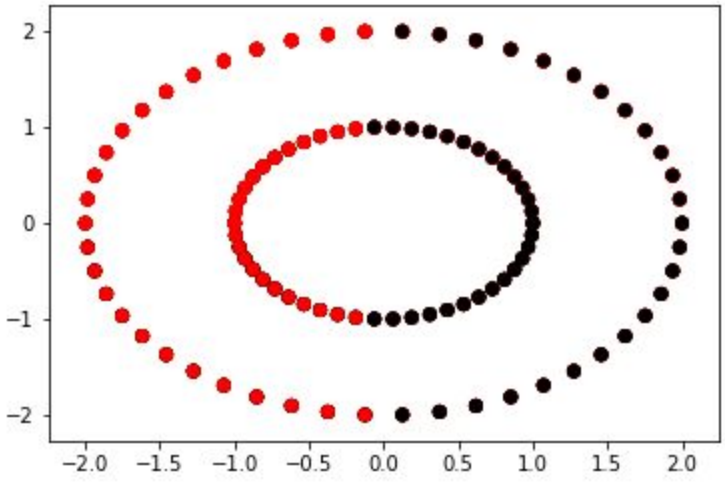
A Simple Comparison(Refer to **Spectral_Circle.py)**

2)Scatter Plots for different values of sigma and K = 2 of Spectral Clustering

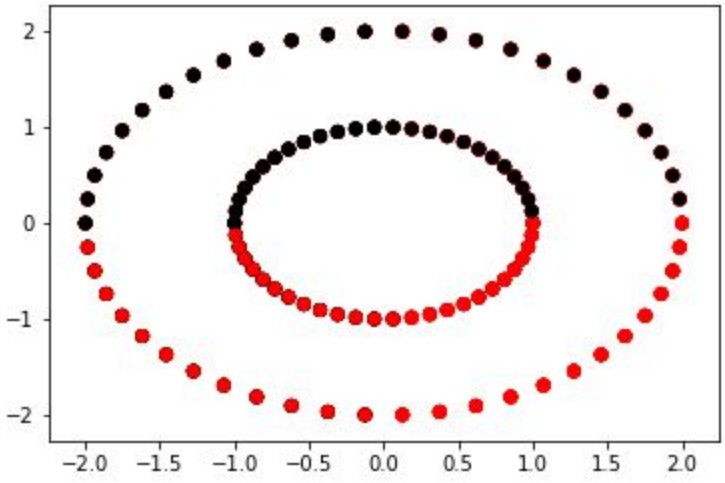
Sigma = 10000



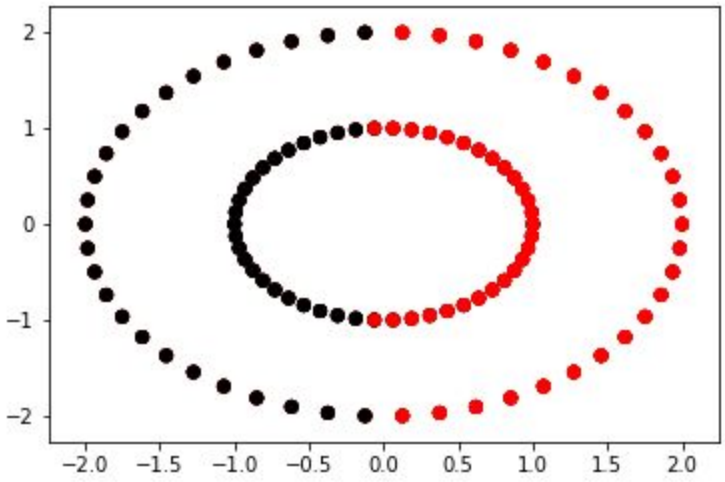
Sigma = 1000



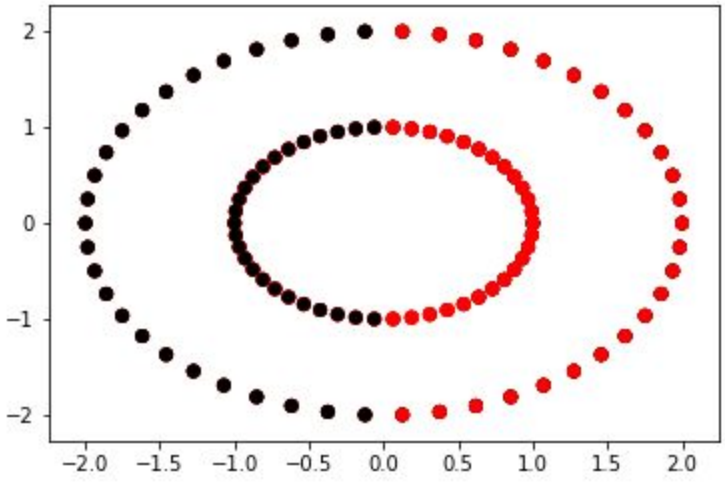
Sigma = 100



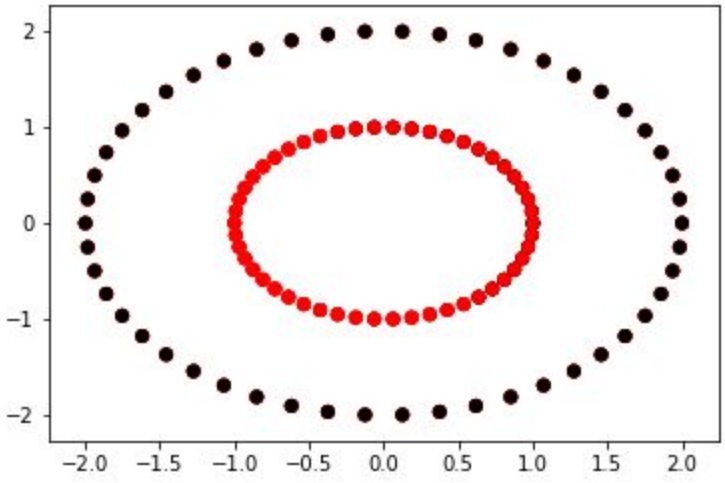
Sigma = 10



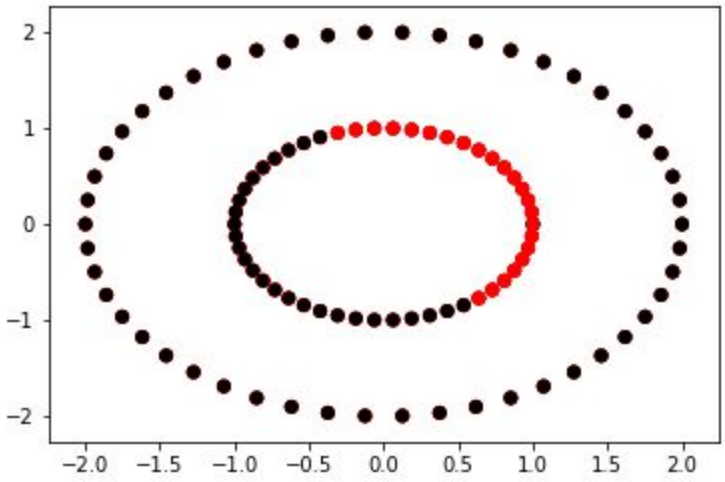
Sigma = 1



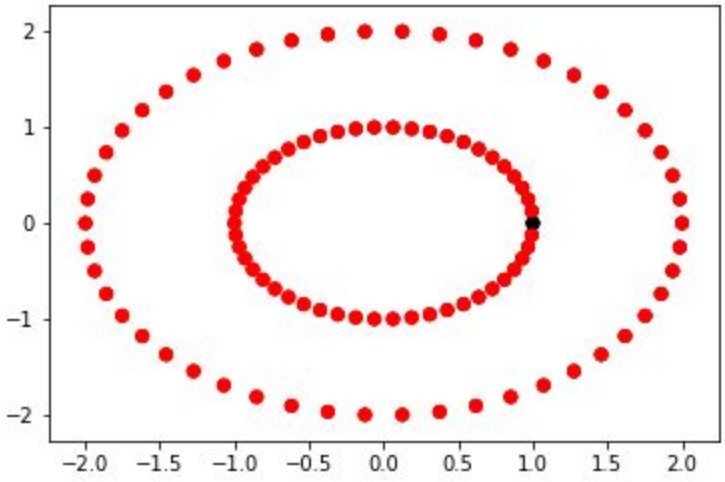
Sigma = 0.1



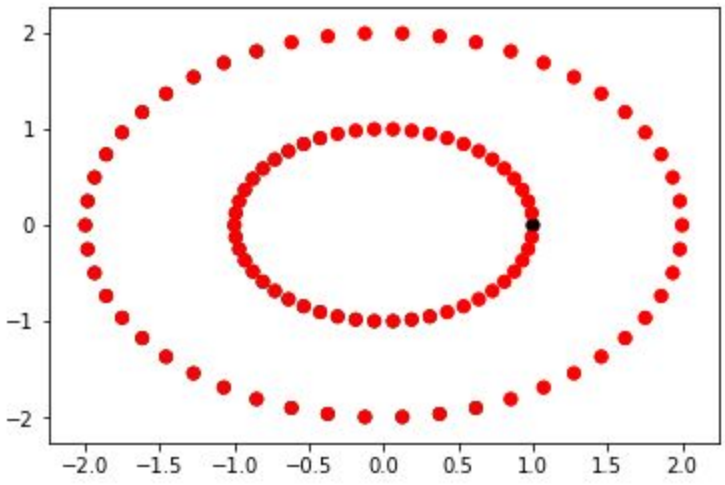
Sigma = 0.01



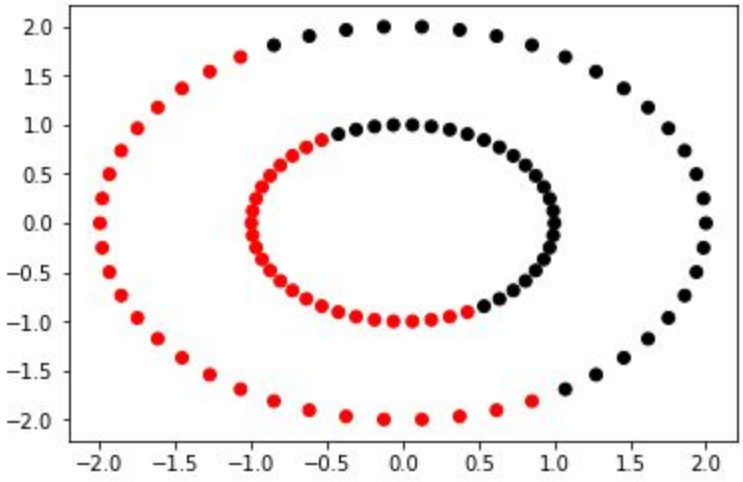
Sigma = 0.001



Sigma = 0.0001



Kmeans Scatter Plot



3) The choice of Sigma for which the Spectral Clustering Outperforms Kmeans is 0.1 which can be seen from the above scatter plots.

(Detailed Comparision between Spectral and Kmeans by using loss function with respect to cluster centers is in **Spectral_Circle.py)**

Loss_kmeans 158.69083991428775
Loss_spectral 5.535360064558555e-27 **For Sigma = 0.1**

There is no K-Means Solution because the data given is Concentric circles data which is unordered. So, Kmeans does not perform well on Unordered data as it tries to minimize the loss function w.r.to cluster centers. So, No choice of Cluster Centers and Clusters will outperform Spectral Clustering. On the other hand Spectral Clustering will project the data into new dimensional spaces and it applies K-means on the projected data.

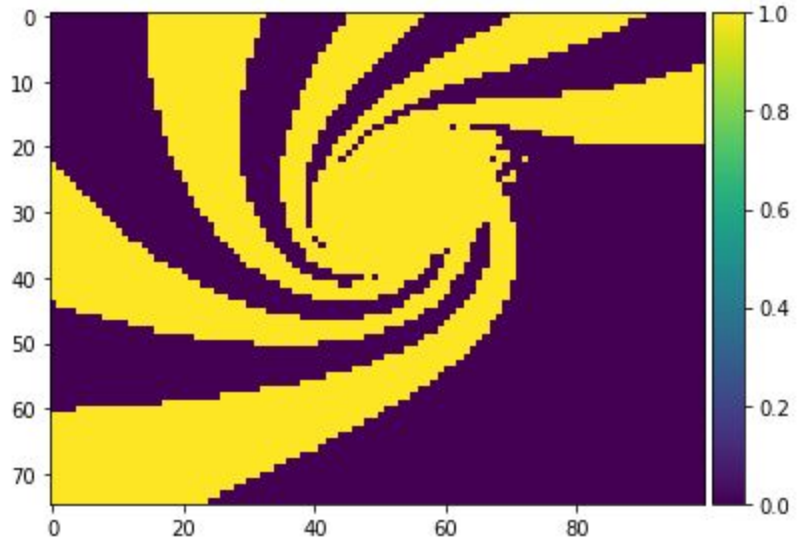
Partitioning Images (Refer to **partitionImg.py**)
Spectral Clustering For different Values of Sigma and K = 2

Sigma = 0.1

Sigma = 0.01

Sigma = 1

Kmeans



For sigma = 0.1 and K = 2, Spectral Clustering outperforms Kmeans.

Loss_kmeans 8552679.70449823
Loss_Spectral 51603522.79023312 for sigma = 0.1

Loss_kmeans 8552679.704498377
Loss_Spectral 83615225.92841788 for sigma = 0.01
Loss_kmeans 8552679.70449823
Loss_Spectral 82425681.27803117 for sigma = 1