Introduction to Data Mining: Assignment #4

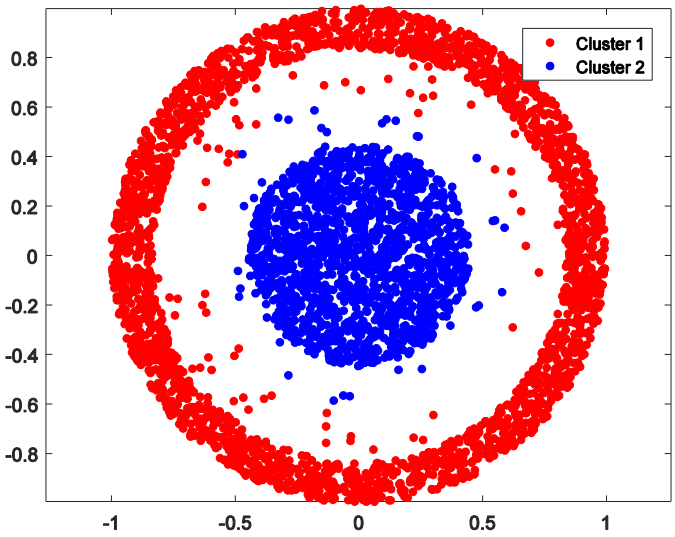
Junpei Zhou

3140102286

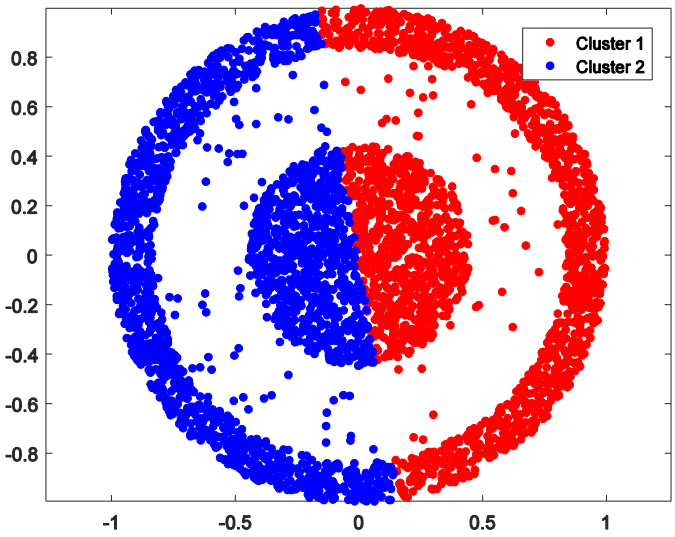
1. Spectral Clustering

a) Spectral Clustering on synthesis data

The result of the spectral clustering is:



The result of the knn clustering is:



We can see that for the situation where the data cannot be “seperated” into two convex shape, the knn performs bad. However, the spectral clustering can project it into another space where the data can be clustered easily.

b) Spectral Clustering on real-world data

For the spectral clustering, the accuracy is 0.800265 and the normalized mutual information is 0.617138

For the k-means clustering, the accuracy is 0.542509 and the normalized mutual information is 0.322413

2. Principal Component Analysis

a) Recover the rotated CAPTCHA image using PCA

I use the index of the image’s pixels whose color is not white, and then this problem is equivalent to the problem of finding the axis of a rectangle, so we just need to use PCA to find the first Principle Component, and then calculate the degree of this axis, or say this component, and then we can rotate the image according to the degree. Here is the result of my recovery:

b) Apply PCA to a face image dataset

i) The Eigenface is shown as below:

ii) The test error rate for each of these dimensionality is:

For the 8 eigen face, the test error rate is 0.245000

For the 16 eigen face, the test error rate is 0.200000

For the 32 eigen face, the test error rate is 0.180000

For the 64 eigen face, the test error rate is 0.150000

For the 128 eigen face, the test error rate is 0.150000

iii) The original face is:

Here is the result of reconstructing the face:

- When the number of eigenface is 8:

- When the number of eigenface is 16:

- When the number of eigenface is 32:

- When the number of eigenface is 64:

- When the number of eigenface is 128: