Introduction to Data Mining: 21191441

Zhejiang University June 4, 2019

Professor Deng Cai Homework 4

Homework 4

Collaborators:

Name: Mingkuan Tong

Student ID: 3170102904

Problem 4-1. Spectral Clustering

In this problem, we will try a dimensionality reduction based clustering algorithm –Spectral Clustering.

(a) We will first experiment Spectral Clustering on synthesis data

Answer:

Here I let k_in_knn_graph = 200 and threshold = 0.2. The result of spectral clustering is in the Figure 1 and K-means is in the Figure 2. Obviously from figures, the Spectral clustering works better.

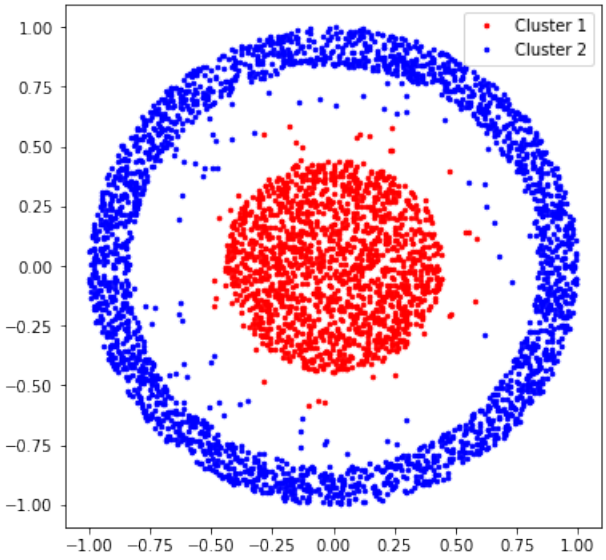

Figure 1

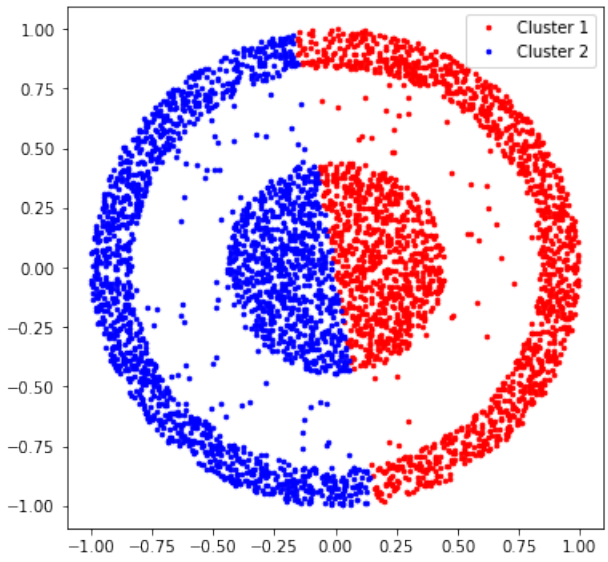

Figure 2

(b) Now let us try Spectral Clustering on real-world data.

Answer:

According to the below Table 1, we will see the spectral clustering has a higer accuracy and normalized mutual information than K-means clustering, which means spectral clustering works better. In the table, higher accuracy means the spectral clustering assigns more data to the correct clusters and higher normalized mutual information means has higher similarity between the result of clustering and truth.

	accuracy	normalized mutual information
spectral clustering	0.66	0.59
K-means clustering	0.52	0.34

Table 1: Accuracy and normalized mutual info of spectral and K-means clustering

Problem 4-2. Principal Component Analysis Let us deepen our understanding of PCA by the following problems.

(a) Your task is to implement *hack_pca.m* to recover the rotated CAPTCHA image using PCA.

Answer:

The results are in the Figure 3. Here is still a problem that the rotation of one result is not good.

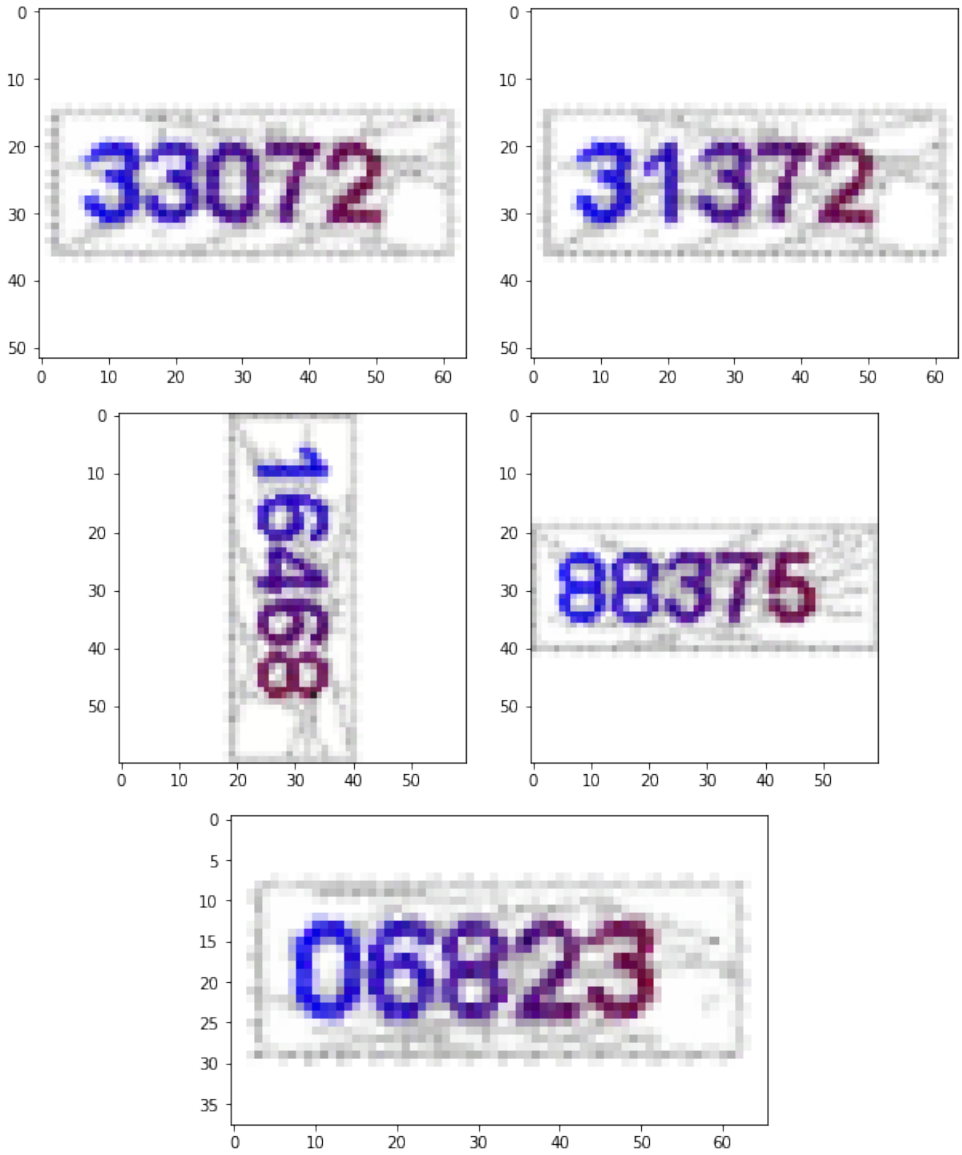

Figure 3

(b) Now let us apply PCA to a face image dataset.

Answer: The original image is in the Figure 4.

The EigenFace is in the Figure 5.

Figure 4: original image

Figure 5: EigenFace

The results of using PCA to do dimensionality reduction are in the Figure 6、7、8、9、10 below.

The testing error rates of different number of reduced dimensionality are in the table 2.

dimensionality	testing error rate
8	0.255
16	0.175
32	0.155
64	0.135
128	0.135

Table 2: Testing error rates of PCA with different number of reduced dimensionality

The testing error rates of different number of reduced dimensionality are in the table 3.

Figure 6: dimensionality = 8

Figure 7: dimensionality = 16

Figure 8: dimensionality = 32

Figure 9: dimensionality = 64

Figure 10: dimensionality = 128

dimensionality	testing error rate
8	0.115
16	0.055
32	0.045
64	0.03
128	0.03

Table 3: Testing error rates of LDA with different number of reduced dimensionality