

CSE 547: Machine Learning for Big Data

Homework 2

January Shen
April 30, 2019

Answer to Question 1(a)

see scanned part.

Answer to Question 1(b)

1(b)-1

sum of eigenvalues is 1084.2074349947675

$$\lambda_1 = 781.8126992600016$$

$$\lambda_2 = 161.15157496732692$$

$$\lambda_{10} = 3.339586754887817$$

$$\lambda_{30} = 0.8090877903777284$$

$$\lambda_{50} = 0.38957773951814617$$

1(b)-2

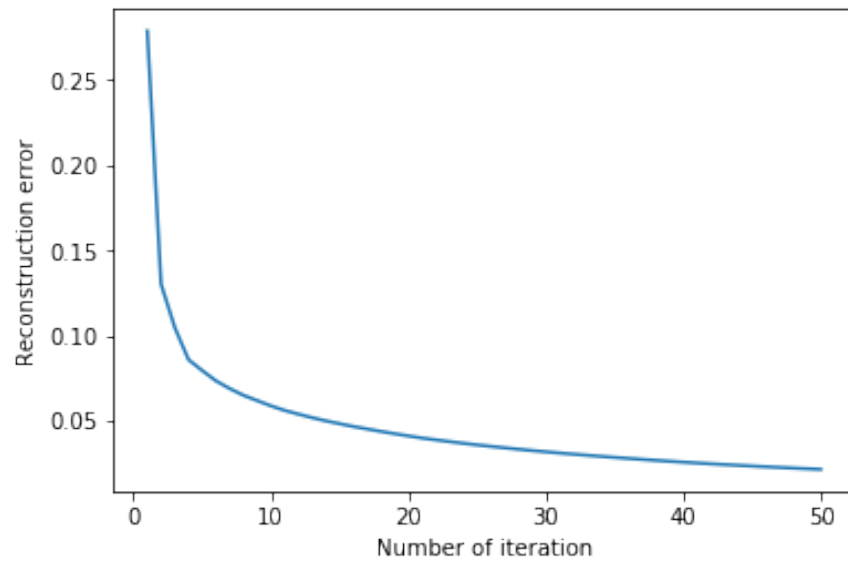


Figure 1: Figure for question 1 (b2)

1(b)-3

The principle eigenvalue captures the major theme of the pictures, which is the contour of a face. This feature is shared by all images. Other features, such as the shape of eyes and eyebrows, are less commonly shared so the eigenvalues are smaller.

Answer to Question 1(c)

1(c)-1

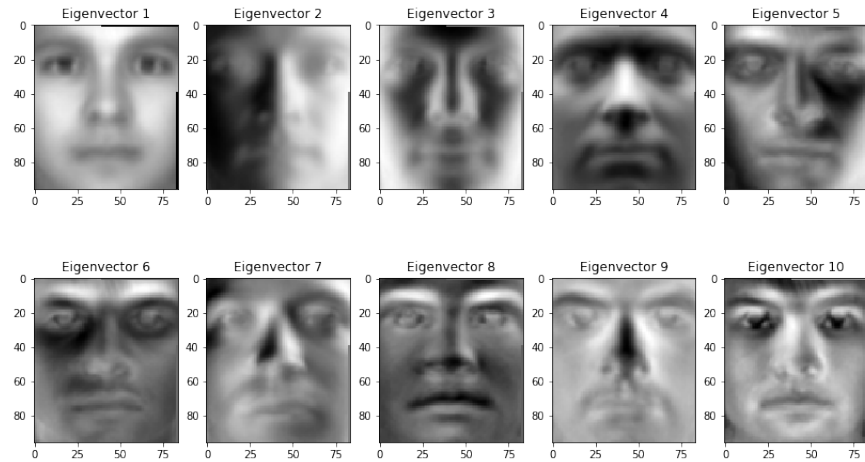


Figure 2: Figure for question 1 (c1)

1(c)-2

- 1: blurred image of a face
- 2: contour of a face with light from the right
- 3: contour of a face with light from the back
- 4: contour of a face with light from the front
- 5: contour of a face with light from the left
- 6: contour of a face with light from the top
- 7: contour of a face with lighter scale
- 8: contour of a face with darker scale
- 9: contour of a face with lighter scale
- 10: blurred image of a face

Answer to Question 1(d)

1(d)-1

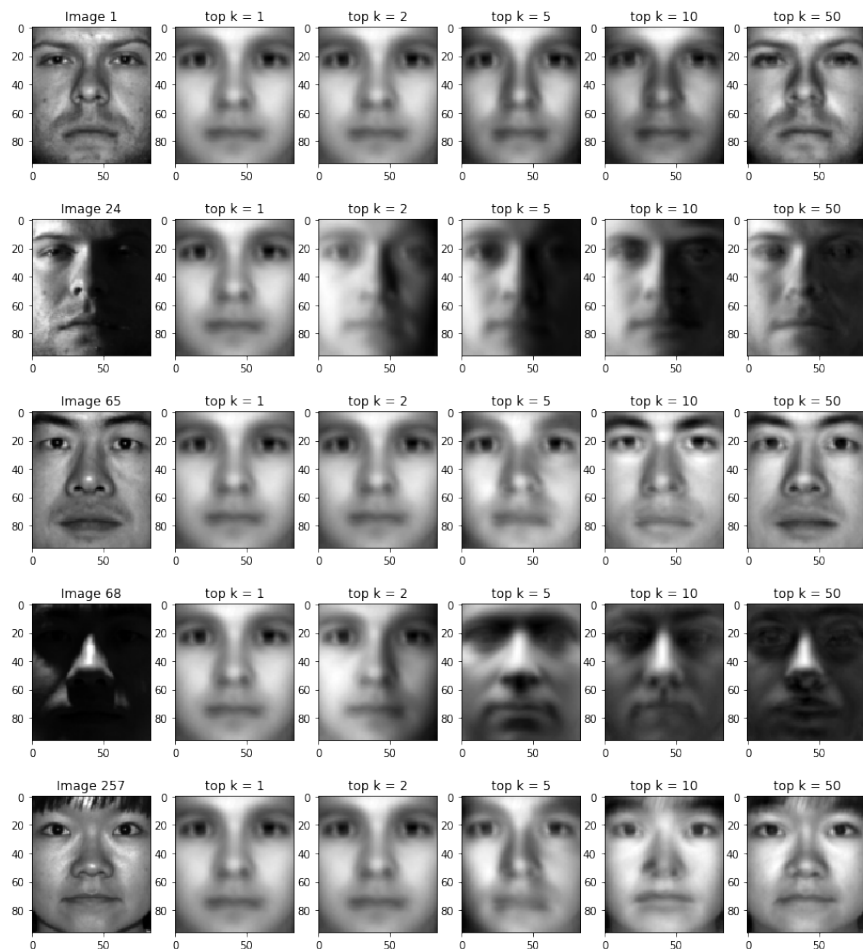


Figure 3: Figure for question 1 (d1)

1(d)-2

The more eigenvectors are applied, the clearer the image is. When we apply for more eigenvectors, the reconstruction error becomes less, so the reconstructed image becomes more like the original image.

Answer to Question 2(a)

2(a)-1

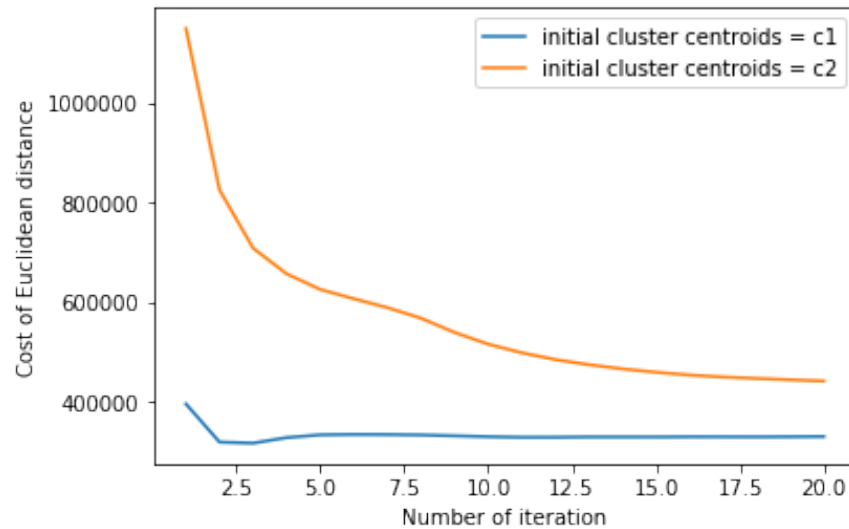


Figure 4: Figure for question 2 (a1)

2(a)-2

The cost change for Euclidean distance after 10 iterations is 20% for c1, and 123% for c2. The randomly chosen c1 has a lower cost at the beginning, but the cost in each iteration bounces back and forth sometimes. c2 has each dot positioned at the farthest distance. The initial setting may be far away from optimal, but it ensures improvement in each iteration.

Answer to Question 2(b)

2(b)-1

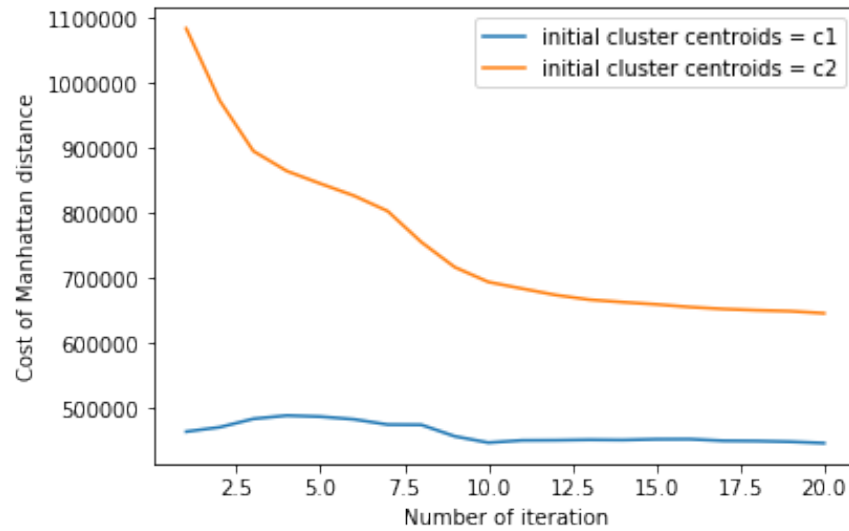


Figure 5: Figure for question 2 (b1)

2(b)-2

The cost change for Manhattan distance after 10 iterations is 5% for c1, and 28% for c2. The randomly chosen c1 has a lower cost at the beginning, but the cost in each iteration bounces back and forth sometimes. c2 has each dot positioned at the farthest distance. The initial setting may be far away from optimal, but it ensures improvement in each iteration.

Answer to Question 3(a)

see scanned part.

Answer to Question 3(b)

The learning rate here is 0.01.

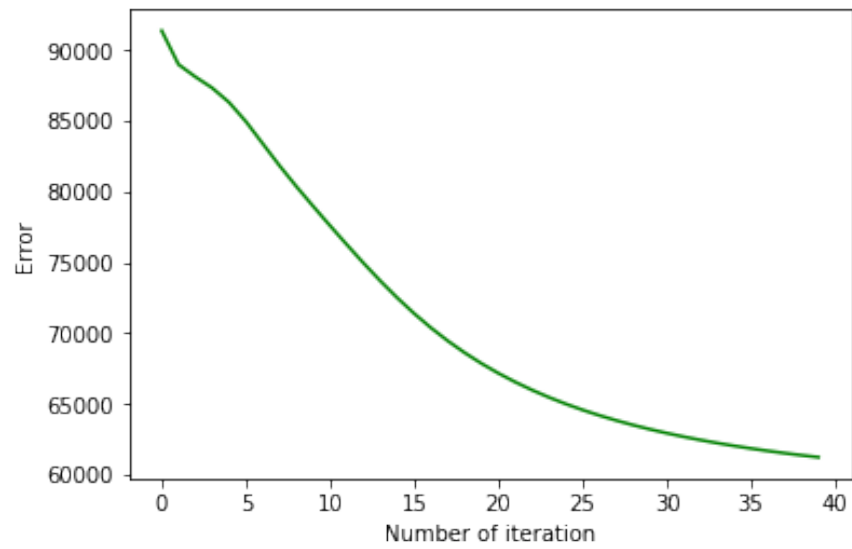


Figure 6: Figure for question 3 (b)

Answer to Question 4(a)

see scanned part.

Answer to Question 4(b)

see scanned part.

Answer to Question 4(c)

see scanned part.

Answer to Question 4(d)

User-User recommendation:

"FOX 28 News at 10pm", Similarity = 908.48

"Family Guy", Similarity = 861.18

"2009 NCAA Basketball Tournament", Similarity = 827.60

"NBC 4 at Eleven", Similarity = 784.78

"Two and a Half Men", Similarity = 757.60

Item-item recommendation:

"FOX 28 News at 10pm", Similarity = 31.36

"Family Guy", Similarity = 30.00

"NBC 4 at Eleven", Similarity = 29.40

"2009 NCAA Basketball Tournament", Similarity = 29.23

"Access Hollywood", Similarity = 28.97