Assignment 7: Unsupervised learning (PCA, K-Means)

November 26, 2017

Questions

In this assignment, you will need to compute the Principal Component Analysis and K-Means algorithms and use them on a dataset. The dataset is the set of images from MNIST database corresponding to the handwritten digit 7. Each image is $28px \times 28px$. The set is divided in a training set and a testing set of respective size 3133 and 3132.

As usual, the structure of the code is given to you and you need to fill the parts corresponding to the questions below.

Question 1. Complete the functions `pca(.)` and `pca_project(.)`. For information, the function `np.linalg.eigh` compute the eigenvalues and eigenvectors of a symmetric matrix. It returns two arrays, the first one contains the eigenvalues in ascending order and the second one the corresponding eigenvector.

Question 2. Using the decomposition learnt on the training test, compute the reconstruction error E on the testing test defined by:

$$E(D) = \frac{1}{N} \sum_{n=1}^{N} \| I_n - (\bar{\ } + \sum_{k=1}^{D} \omega_k^n \mathbf{u}_k) \|_2, \tag{1}$$

with N corresponding to the number of images in the TestSet, I_n denoting the n-th image of the testing set, $\bar{\ }$ is the mean digit learnt from the training set, $\mathbf{u}_k$ is the eigenvector with the k-th largest eigenvalue, and ω_k^n is the expansion coefficient of the n-th image on the k-th eigenvector. Finally, $\|.\|_2$ denotes the L_2 norm. Numpy has the method `np.linalg.norm(.)` that computes norms (check out the documentation for more infos).

Question 3. Plot the evolution of the error E for $D = 1, \ldots, 100$.

Question 4. Complete the function `distortion(.)` which computes the distortion cost F for a given clustering of the data:

$$F(m, c) = \frac{1}{N} \sum_{i=1}^{N} \| x^i - c^{m(i)} \|_2, \tag{2}$$

where N corresponds to the total number of images in the set and $m(i)$ denotes which cluster is assigned to the image x^i.

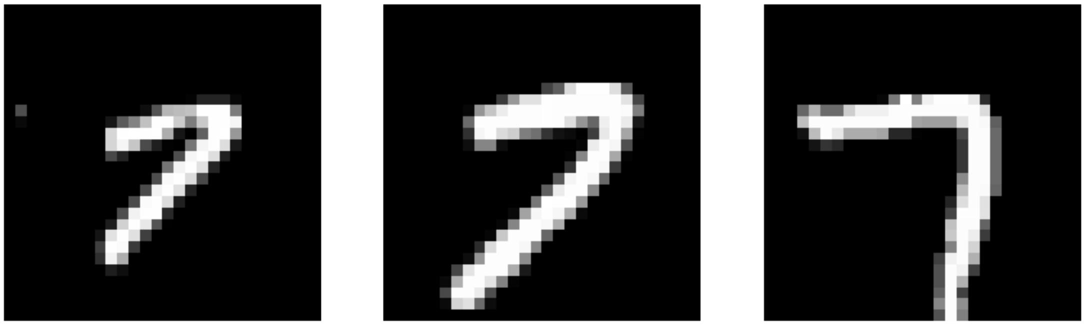
Question 5. Complete the function `kmeans(.)`, make sure that it computes the distortion after each update. Then use the function on your training set, the number of cluster $k = 2$. Check that the distortion decreases as the algorithm progresses.

Question 6. In order to mitigate the local minima problem of K-Means, repeat the algorithm 10 times, and keep the solution that yields the smallest distortion at the end. Show the resulting digit clusters.

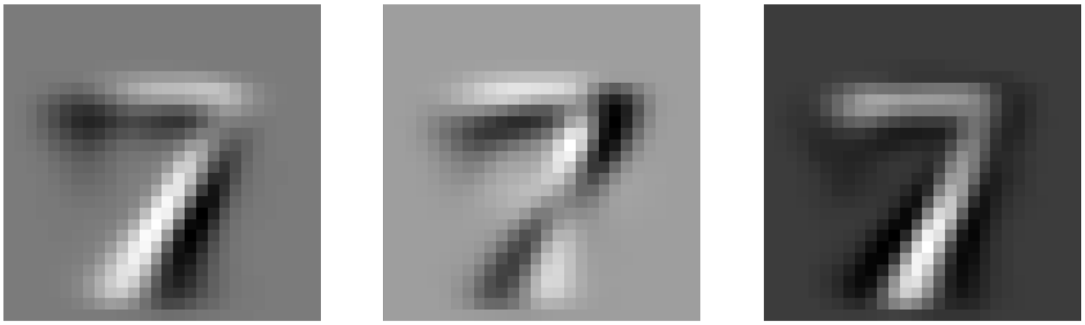
Question 7. Repeat the procedure of Question 6 for values of $k = 3, 4, 5, 10, 50, 100$ (allow for ~10min). For each such value report the distortion cost of the training and testing data.

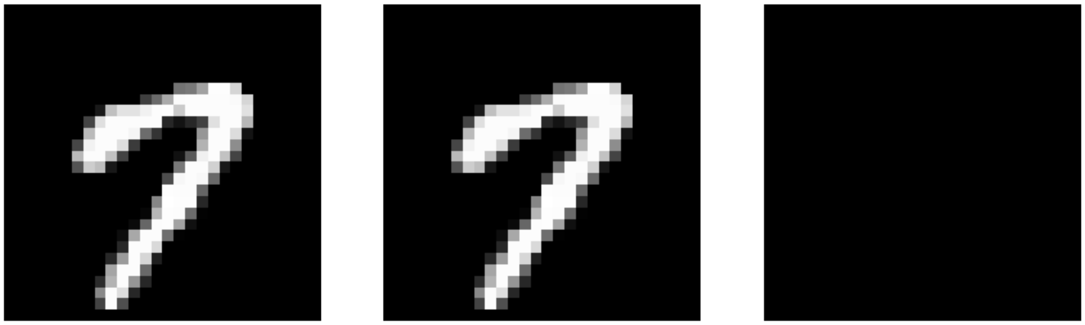
Question 8. Compare the results from PCA to the results of K-means on the test set by plotting on the same graph the reconstruction error $E(D)$ for $D3, 4, 5, 10, 50, 100$ and the distortion cost you just computed (remark that the two measures are simply L_2 norms thus the comparison is valid). To be clear, the first one measure the error in the reconstructed image from the projection on the components of PCA, the second measure the error between each image and the centroid of the cluster it is assigned to. Both correpond to the error made when approximating the original image to either its projection or its cluster's centroid.

Importing the data to form training and test sets

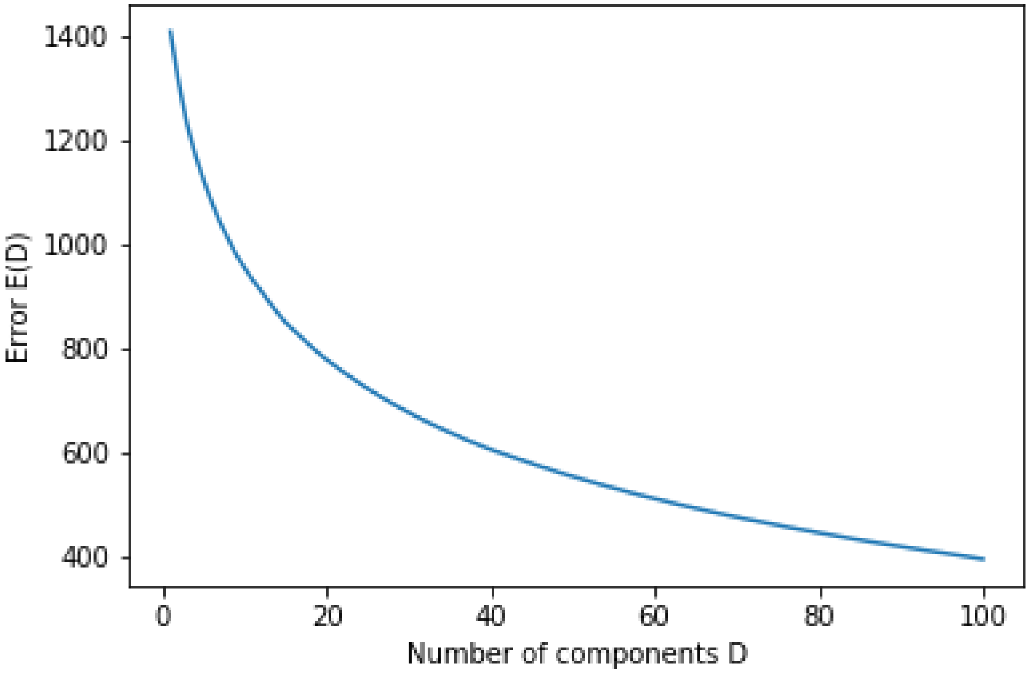


Test implementation





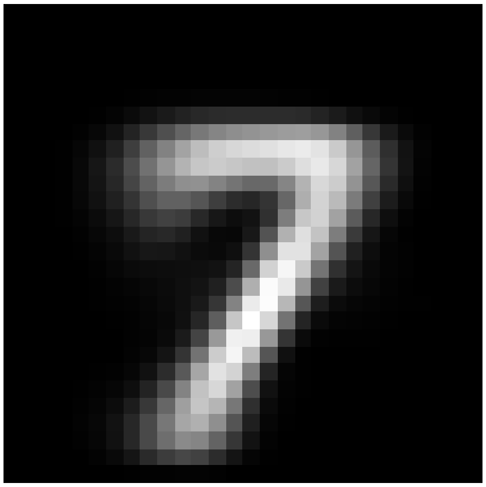
Testing PCA

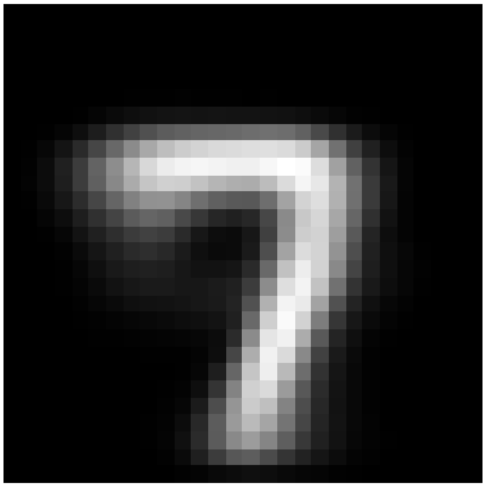


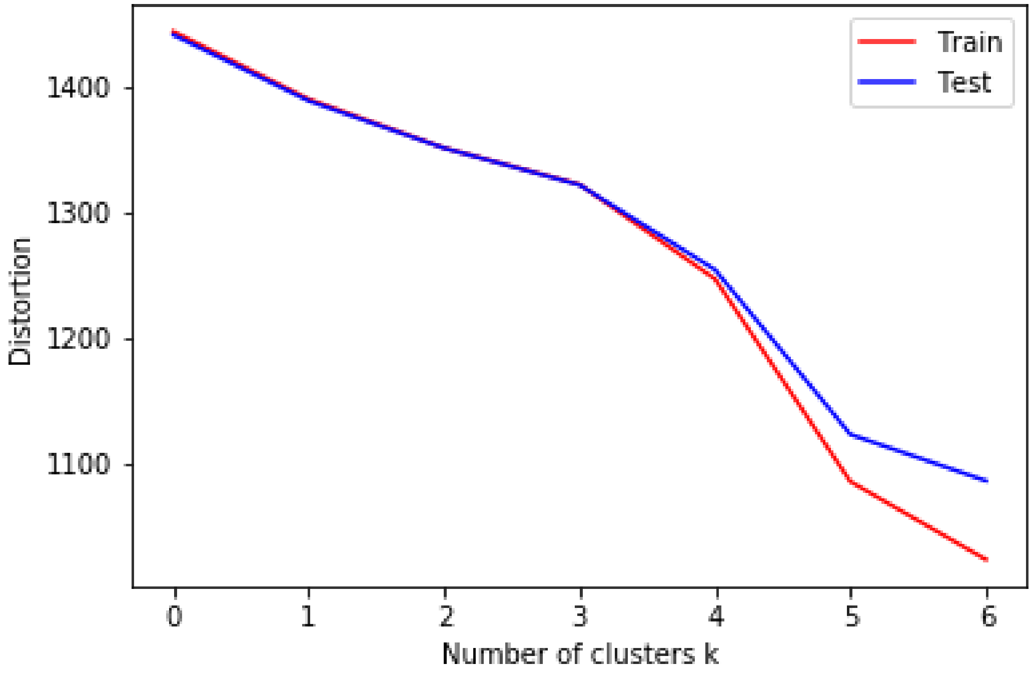
K-Means

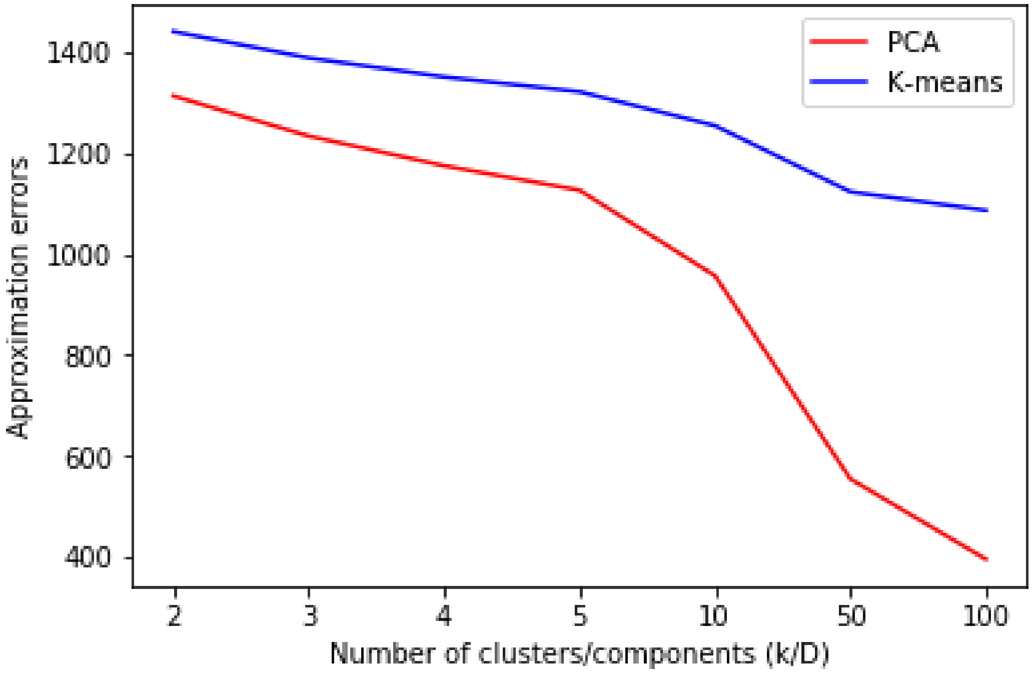
```
Iteration 1, distortion = 2090.74813897
Iteration 2, distortion = 1452.30911102
Iteration 3, distortion = 1447.64905431
Iteration 4, distortion = 1446.00589447
Iteration 5, distortion = 1444.89337228
Iteration 6, distortion = 1444.20812056
Iteration 7, distortion = 1443.80459365
Iteration 8, distortion = 1443.60184782
Iteration 9, distortion = 1443.50877501
Iteration 10, distortion = 1443.43447869
Iteration 11, distortion = 1443.38884229
Iteration 12, distortion = 1443.3651259
Iteration 13, distortion = 1443.33449997
Iteration 14, distortion = 1443.28489795
Iteration 15, distortion = 1443.26795045
Iteration 16, distortion = 1443.26274859
Terminates with difference: 0.0
```

Testing K-means






Train
Test
1400
1300
1200
1100
Distortion
0
1
2
3
4
5
6
Number of clusters k


PCA
K-means
1400
1200
1000
800
600
400
Approximation errors
2
3
4
5
10
50
100
Number of clusters/components (k/D)