Jordan Gilman

931942845

CS 434

Implementation Assignment #4

1. Non-hierarchical clustering – K-Means algorithm
   a.

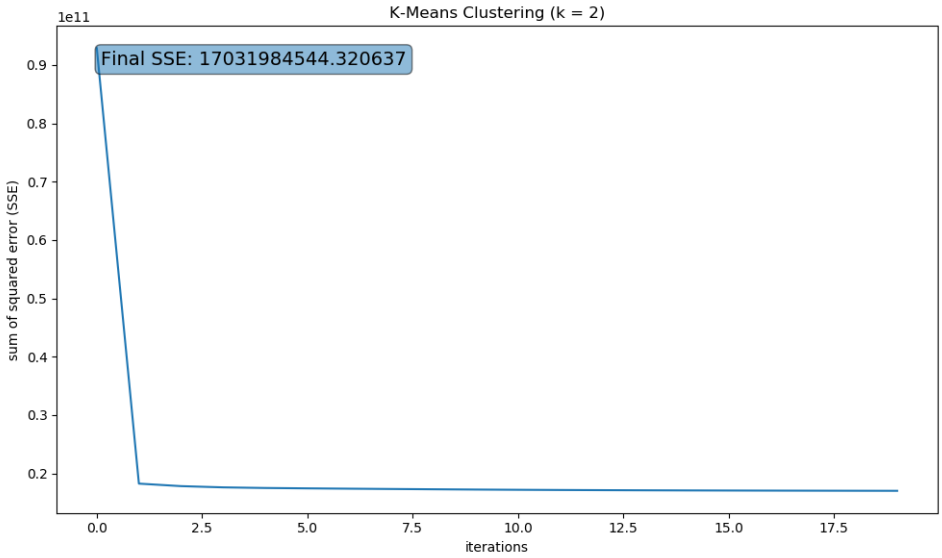


*Figure 1. Convergence of K-Means algorithm over 20 iterations*

   b. After running the K-Means algorithm with random initialization ten times for k = 2, 3, …, 10, the SSEs shown in figure 2 were obtained. Strictly based on the curve, k=10 would be the best value since it still performs noticeably better than any other values of k. For greater values of k, I would expect to see a saturation in performance that would become increasingly computationally expensive.

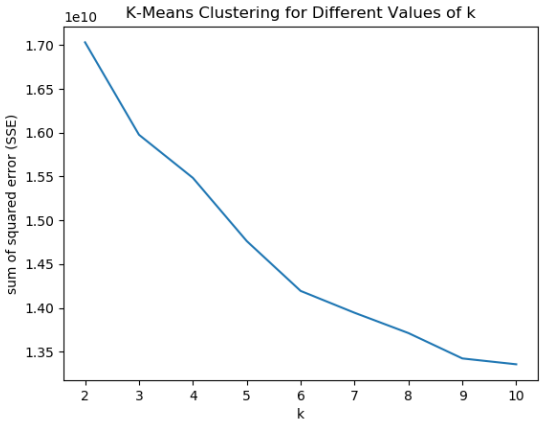


*Figure 2. Objective function results for different values of k*

2. Principal Component Analysis (PCA)
   a.

| Index | Eigenvalue |
|---|---|
| 1 | 352868 |
| 2 | 267895 |
| 3 | 227632 |
| 4 | 174703 |
| 5 | 130486 |
| 6 | 115542 |
| 7 | 99726 |
| 8 | 90576 |
| 9 | 85326 |
| 10 | 71547 |

*Table 1. Top 10 eigenvalues from PCA of handwritten digits*

b. The plotted eigenvectors show the ten most prominent components of the entire dataset and the corresponding eigenvalues represent the variance of the data with respect to the eigenvector's "direction."

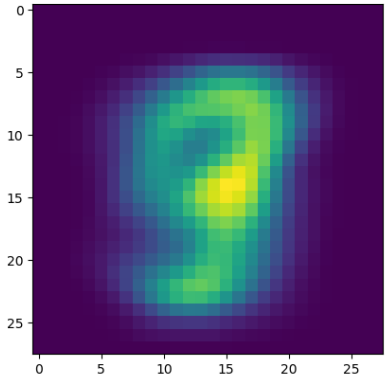


*Figure 3. Mean image from dataset*

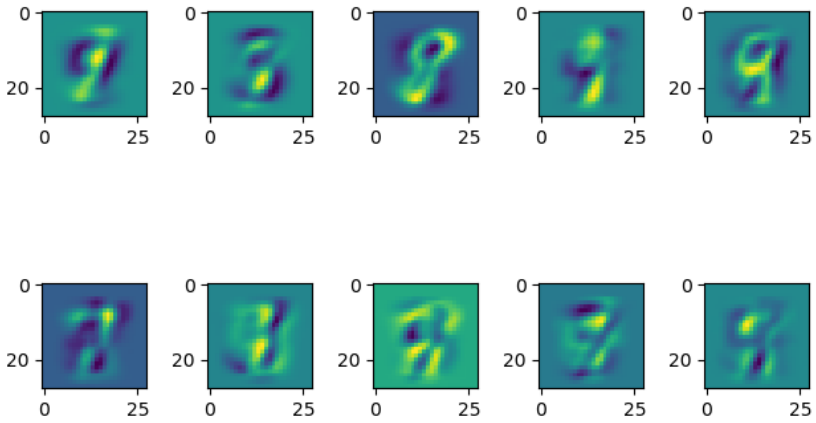


*Figure 4. Eigenvectors with the greatest corresponding eigenvalues*

c. The images with the greatest/least value in one of the ten dimensions looks very alike to its eigenvector counterpart in the positive or negative direction. The dimensions seem to serve as a sort of classification with greater absolute values corresponding to a "categorization" to one or more components.

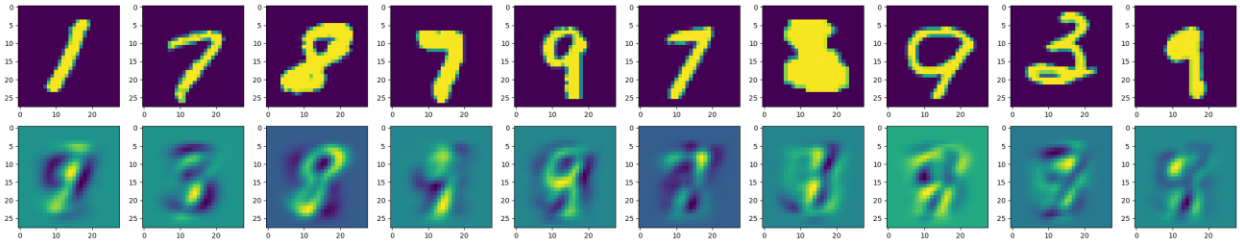

*Figure 5. Digits with greatest value in a particular dimension and the eigenvector corresponding to that dimension*

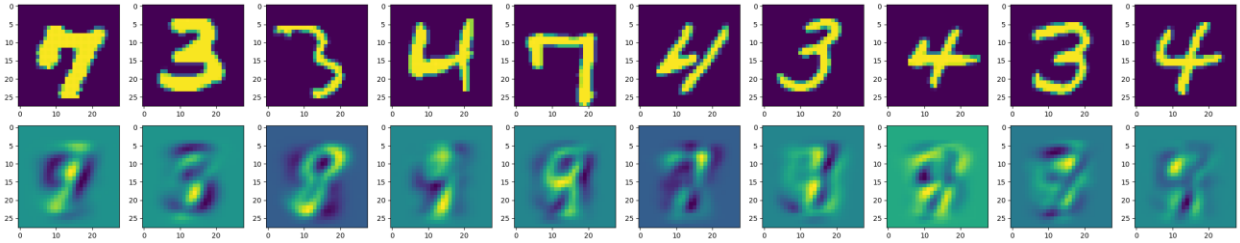

*Figure 6. Digits with least value in a particular dimension and the eigenvector corresponding to that dimension*