

ML Assignment 5

Implement k-means clustering. Analyze the clusters formed for various values of k. Display the centroids of the clusters.

- **K-Means Clustering:** Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.
- **K-Means Clustering Algorithm :**

The way K-Means algorithm works is as follows:

 1. Specify number of clusters K.
 2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
 3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
 - 3.1 Compute the sum of the squared distance between data points and all centroids.
 - 3.2 Assign each data point to the closest cluster (centroid).
 - 3.3 Compute the centroids for the clusters by taking the average of all data points that belong to each cluster.
- **Result:** MNIST data set of digit images is used to implement k mean clustering. Here we initialize K to {10,15,20}. Clustering is able to classify similar digit into one cluster but it fails in some cases where digit has a nearly similar shape. Following are the output with different value of K:

For K = 10

Centroid 1



Image from Cluster 1



Image from Cluster 1



Image from Cluster 1



Image from Cluster 1



Centroid 2



Image from Cluster 2



Image from Cluster 2



Image from Cluster 2



Image from Cluster 2



Centroid 3



Image from Cluster 3



Image from Cluster 3



Image from Cluster 3



Image from Cluster 3



Centroid 4



Image from Cluster 4



Image from Cluster 4



Image from Cluster 4



Image from Cluster 4



Centroid 5



Image from Cluster 5



Image from Cluster 5



Image from Cluster 5



Image from Cluster 5



Centroid 6



Image from Cluster 6



Image from Cluster 6



Image from Cluster 6



Image from Cluster 6



Centroid 7



Image from Cluster 7



Image from Cluster 7



Image from Cluster 7



Image from Cluster 7



Centroid 8



Image from Cluster 8



Image from Cluster 8



Image from Cluster 8



Image from Cluster 8



Centroid 9



Image from Cluster 9



Image from Cluster 9



Image from Cluster 9



Image from Cluster 9



Centroid 10



Image from Cluster 10



Image from Cluster 10



Image from Cluster 10



Image from Cluster 10



For K = 15

Centroid 1



Image from Cluster 1



Image from Cluster 1



Image from Cluster 1



Image from Cluster 1



Centroid 2



Image from Cluster 2



Image from Cluster 2



Image from Cluster 2



Image from Cluster 2



Centroid 3



Image from Cluster 3



Image from Cluster 3



Image from Cluster 3



Image from Cluster 3



Centroid 4



Image from Cluster 4



Image from Cluster 4



Image from Cluster 4



Image from Cluster 4



Centroid 5



Image from Cluster 5



Image from Cluster 5



Image from Cluster 5



Image from Cluster 5



Centroid 6



Image from Cluster 6



Image from Cluster 6



Image from Cluster 6



Image from Cluster 6



Centroid 7



Image from Cluster 7



Image from Cluster 7



Image from Cluster 7



Image from Cluster 7



Centroid 8



Image from Cluster 8



Image from Cluster 8



Image from Cluster 8



Image from Cluster 8



Centroid 9



Image from Cluster 9



Image from Cluster 9



Image from Cluster 9



Image from Cluster 9



Centroid 10



Image from Cluster 10



Image from Cluster 10



Image from Cluster 10



Image from Cluster 10



Centroid 11



Image from Cluster 11



Image from Cluster 11



Image from Cluster 11



Image from Cluster 11



Centroid 12



Image from Cluster 12



Image from Cluster 12



Image from Cluster 12



Image from Cluster 12



Centroid 13



Image from Cluster 13



Image from Cluster 13



Image from Cluster 13



Image from Cluster 13



Centroid 14



Image from Cluster 14



Image from Cluster 14



Image from Cluster 14



Image from Cluster 14



Centroid 15



Image from Cluster 15



Image from Cluster 15







































































































Image from Cluster 15



Image from Cluster 15



For K = 20

Centroid 1		Image from Cluster 1		Image from Cluster 1		Image from Cluster 1		Image from Cluster 1	
Centroid 2		Image from Cluster 2		Image from Cluster 2		Image from Cluster 2		Image from Cluster 2	
Centroid 3		Image from Cluster 3		Image from Cluster 3		Image from Cluster 3		Image from Cluster 3	
Centroid 4		Image from Cluster 4		Image from Cluster 4		Image from Cluster 4		Image from Cluster 4	
Centroid 5		Image from Cluster 5		Image from Cluster 5		Image from Cluster 5		Image from Cluster 5	
Centroid 6		Image from Cluster 6		Image from Cluster 6		Image from Cluster 6		Image from Cluster 6	
Centroid 7		Image from Cluster 7		Image from Cluster 7		Image from Cluster 7		Image from Cluster 7	
Centroid 8		Image from Cluster 8		Image from Cluster 8		Image from Cluster 8		Image from Cluster 8	
Centroid 9		Image from Cluster 9		Image from Cluster 9		Image from Cluster 9		Image from Cluster 9	
Centroid 10		Image from Cluster 10		Image from Cluster 10		Image from Cluster 10		Image from Cluster 10	
Centroid 11		Image from Cluster 11		Image from Cluster 11		Image from Cluster 11		Image from Cluster 11	
Centroid 12		Image from Cluster 12		Image from Cluster 12		Image from Cluster 12		Image from Cluster 12	
Centroid 13		Image from Cluster 13		Image from Cluster 13		Image from Cluster 13		Image from Cluster 13	
Centroid 14		Image from Cluster 14		Image from Cluster 14		Image from Cluster 14		Image from Cluster 14	
Centroid 15		Image from Cluster 15		Image from Cluster 15		Image from Cluster 15		Image from Cluster 15	
Centroid 16		Image from Cluster 16		Image from Cluster 16		Image from Cluster 16		Image from Cluster 16	
Centroid 17		Image from Cluster 17		Image from Cluster 17		Image from Cluster 17		Image from Cluster 17	
Centroid 18		Image from Cluster 18		Image from Cluster 18		Image from Cluster 18		Image from Cluster 18	
Centroid 19		Image from Cluster 19		Image from Cluster 19		Image from Cluster 19		Image from Cluster 19	
Centroid 20		Image from Cluster 20		Image from Cluster 20		Image from Cluster 20		Image from Cluster 20	

Implement Dimensionality reduction using PCA

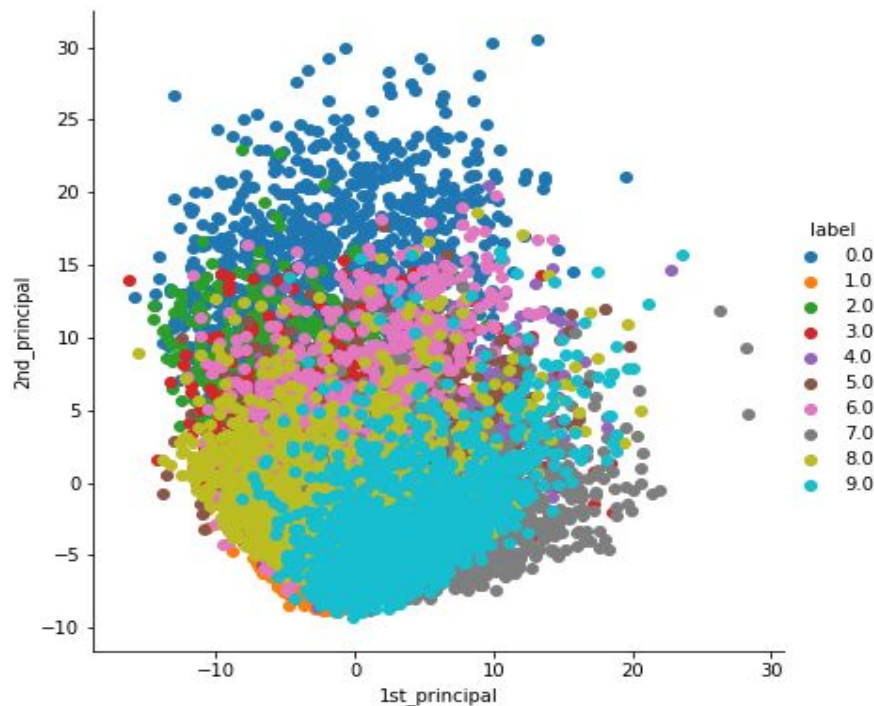
- **Principal Component Analysis:** Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

- **PCA Algorithm:**

The way PCA works is as follows:

1. **Standardize your data:** This is done by subtracting standard deviation and dividing by mean.
2. **Get covariance matrix of the data:** If the data matrix is M the covariance matrix of $M = M^T M$
3. **Calculate Eigen Vectors and Eigen Values:** Calculate eigenvalues and their corresponding eigenvectors of $M^T M$
4. **Sort the Eigen Values:** Sort eigenvalues from largest to smallest.
5. **Calculate the new feature:** If you have to project data to d dimensions then choose first d eigenvectors and form matrix and multiply this matrix with data matrix to get data into d dimension.

- **Result:** MNIST data set of digit images is used to implement PCA. Data has a dimension of 784 which is reduced to 2 using PCA. Following is the output in 2D space after applying PCA to data:



For Reconstruction data to original dimension we use:

$$\text{PCA reconstruction} = \text{PC scores} \cdot \text{Eigenvectors}^T$$

And then add mean and multiply it by the standard deviation to get reconstructed original data. The reconstructed output image is blurry due to reconstruction error.

