Implementing K-means Clustering and PCA Dimesionality Reduction On an Image Dataset

Bhagat Ram Labana, 22m1083, EE2; Shubham Dey, 22m1084, EE2;Mohanish Kumar, 22m1090, EE2
Under the Guidence of:-Prof. Amit Sethi, Electrical Engineering Department, IIT Bombay

***Abstract*—The unlabeled dataset is divided into k distinct clusters by an iterative process called "k-means clustering," which ensures that each dataset only belongs to one group with related properties. Here, K specifies how many pre-defined clusters must be produced as part of the process; for example, if K=2, there will be two clusters, if K=3, there will be three clusters, and so on.Principal Component Analysis (PCA) is a method of unsupervised linear transformation that is frequently employed in a variety of domains, most notably for feature extraction and dimensionality reduction.The fashion items in this dataset, which are scattered, include bags, t-shirts, shirts, and more. After creating a cluster of separate articles using k-means clustering, we will lower the dimensions of our dataset using PCA on our newly grouped dataset. This will enable us to do the necessary operations on our dataset much more quickly.**

***Index Terms*—K-means clustering, PCA, Image Dataset**

I. Introduction

In this project we have implemented PCA and K-means clustering on f-MNIST data to reduce the dimension of image data and cluster the images of the data set with some relative accuracy. K-means clustering and principal component analysis (PCA) are two popular algorithms used for data analysis and dimensionality reduction discussed in section-2 and section-3. In this report, we discuss the implementation of K-means clustering and PCA dimensionality reduction on an image dataset. The image dataset taken is of fashion articles. We go over the different steps of the implementation process of pricipal component analysis(PCA) and k-means clustering. After that section-5 we have shown Practical implementation on f-MNIST dataset, where PCA implemented and its result has been shown. and discuss the results obtained from the analysis. First we have performed PCA on the image dataset to reduce the dimensionality of the image dataset i.e that there is othogonal linear transformation of the dataset. It transforms the dataset into new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate the second greatest variance on the second coordinate, and so on.Then we perform k-means clustering, it works by assigning a number of centroids based on the number of clusters given. Each data point is assigned to the cluster whose centroid is nearest to it. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs. We also discuss the advantages and disadvantages of using both these algorithms on image datasets. Finally, we provide some tips for obtaining better results from the analysis.

II. Principle Component Analysis

Principal Component is a technique used for the application of dimension reduction, Data compression, feature extraction, and Data visualization.
PCA is defined as the orthogonal projection of the data on the lower dimension linear space which we call principle subspace such that the variance of the projected data is maximized. Equivalently, it can be defined as the linear projection that minimizes the average projection cost, defined as the mean squared distance between the data point and their projection.

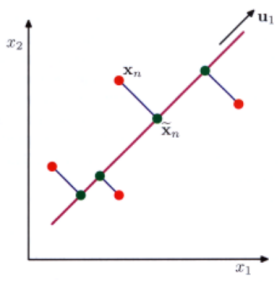

Fig. 1. Principal component analysis seeks a space of lower dimensionality,known as the principal subspace and denoted by the magneta line ,such that the orthogonal projection of the data points (red dots) onto this subspace maximizes the variance of the projected points(green dots). An alternative definition of PCA is based on minimizing the sum of squares of the projection errors, indicated by the blue lines.

A. Mathematics behind the PCA

PCA can be viewed as a challenge in unsupervised learning. Six steps can be used to sum up the entire process of extracting principle components from a raw dataset:

- Let's say our dataset is made up of m+1 dimensions and we ignore the label , now our resulting dimension is m.
- Calculate the mean for each dimension across the entire dataset. Lets say x_{train}=[x_1,x_2,x_3,x_4,.....,x_m] , where x_{train} is the training set and x_1,x_2,x_3,x_4,....x_m are the column of the x_{train} data. So we need to find the average of each column x_i as.

$$\overline{x}_i = \frac{\sum_{j=0}^{n} x_{ij}}{n+1} \tag{1}$$

here n+1 is the no. of data in each feature. And x_{ij} is the j_{th} data of i_{th} column.

- Create a matrix of covariance for the entire dataset. covariance of any two feature x_p and x_q is gievn by the following formula

$$COV(x_p, x_q) = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_i)(x_j - \overline{x}_j)}{n} \quad (2)$$

- Calculate the corresponding eigenvalues and eigenvectors.
  Now if Covariance matrix is given by M the eigenvalue eigenvector relation is given by

$$Mv = \lambda v \quad (3)$$

  where v is eigenvector corresponding to eigenvalue λ
- Now arrange the eigenvector from high dominance to low dominance. And after that choose the k high dominant vector to form a m × k dimensional matrix W.
- Now use this m × k, W eigenvector matrix to transform the samples onto the new subspace.

B. Advantages and disadvantages of PCA

a) Advantages:-:

- The PCA can counteract the issues of a high-dimensional data set
- Correlated features removed
- Speeds up other machine learning algorithms
- Improves visualization

b) Disadvantages:-:

- Data normalization required before performing the PCA
- We may lose some valuable information
- Major components may be difficult to understand

III. K-MEANS CLUSTERING

K-means clustering is one of the effortless and most favoured unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences about data sets using only input vectors without reference to known or labeled results.
A cluster refers to a set of data points grouped together due to certain similarities. We define a target number k, which refers to the number of centroids needed in the dataset. A centroid is an imaginary or real place that represents the center of a cluster. Each data point is assigned to each cluster by subtracting the within-bit sum of squares.
In other words, the K-means algorithm identifies k number of centroids and then separates each data point into the nearest cluster, keeping the centroi ll as possible. The 'mean' of K-means refers to averaging the data; i.e. find the center point.

K-means algorithm uses a first set of randomly chosen centroids as the starting points for each cluster to process the learning data, and iterative calculations are then performed to optimise the positions of the centroids.

When either of the following occurs:

- The centroids have stabilised — there is no change in their values as a result of the clustering being successful.
- Iterations have reached the predetermined number.

The K-Means Algorithm: How is it utilized? The following stages illustrate how the K-Means algorithm functions:

- To determine the number of clusters, choose K.
- Pick K locations or centroids at random. It might not be the input dataset.
- Assign each data point to its nearest centroid, which will create the K clusters that have been predetermined.
- Determine the variance and relocate each cluster's centroid.
- Re-assign each data point to the new centroid of each cluster by repeating the third step.
- move to step 4 if there is a reassignment; otherwise, move to FINISH.
- The finished model.

A. Mathematics behind K-means clustering

- Eucledian distance between two points. If a = (a_1, a_2) and b = (b_1, b_2) then the distance is given by
  $d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$

- Assigning each point to the nearest cluster:
  - If each cluster centroid is denoted by c_i, then each data point x is assigned to a cluster based on

$$\underset{c_i \in C}{argmin} dist(c_i, x)^2$$

    here dist() is the euclidean distance
- Finding the new centroid from the clustered group of points:

$$c_i = \frac{1}{S_i} \sum_{x_i \in S_i} x_i$$

  S_i is the set of all points assigned to the i^{th} cluster.

B. Advantages and Diadvantages of K-means Clustering

a) Advantages:

- It is very easy to put into practise.
- It adapts quickly to large datasets and is scalable to very large data sets.
- It routenly adapts to new cases.
- Cluster generalisation for various sizes and shapes

b) Disadvantages:

- It is attentive to anomalies.
- Manually selecting the k values is difficult.
- Its scalability reduces as the number of dimensions rises.

IV. 3-D Visualization of Clusters

Plotly will be used to create a 3D visualisation of the clusters. Python has a sophisticated visualisation module called Plotly. It facilitates the acquisition of a 3-D scatter plot of clustered data. Understanding how successfully the clusters have developed and how far a single cluster is dispersed into other clusters is made easier with the help of this visualisation.

V. Practical Implementation PCA and K-means Clustering

Here in this section we have shown the practical implementation of PCA and clustering using k-means.First we have implemented PCA where we have taken some input image data and applied the given above algorithm step by step.

A. PCA implementation

1) Input Data: In this project we did PCA analysis to reduce the dimension of the images on the f-MNIST data set.Here is below data-set fasion-MNIST in which we are displaying first ten images of each category.

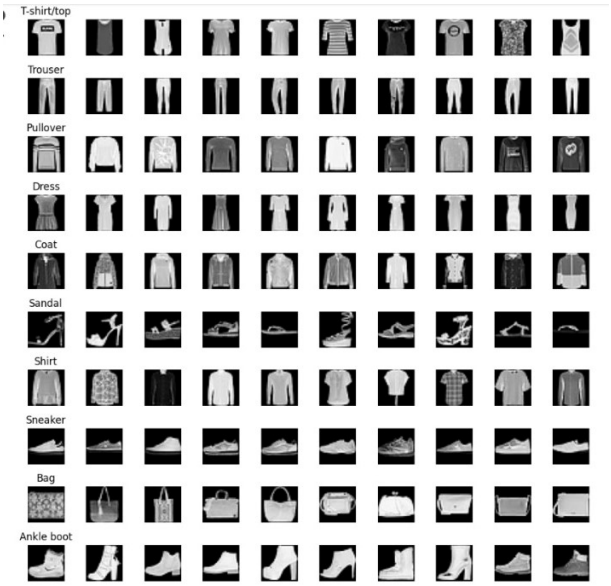

Fig. 2. Output of first ten images taken from the Fashion MNIST data set as input

2) Mean Output: As I have mentioned in the PCA algorithm above we have taken mean images of each category from the data set.

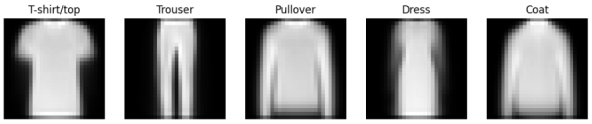

Fig. 3. Mean output images of each respective category

3) Zero-Centering: After that we made all our images zero centered by subtracting the mean image from each image in their respective category.

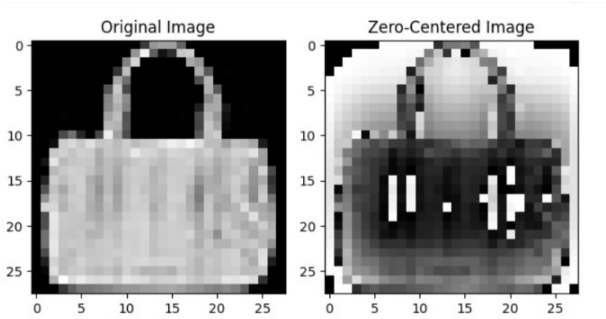

Fig. 4. Original and zero-centered image

After zero centering we calculated the covariance matrix as we have mentioned in the PCA algorithm. Then eigenvalues and the corresponding eigenvectors and sorted them high dominance to low dominance according to the magnitude of the eigenvalues.

4) Compression of image And Recontruction: After applying PCA we have reduced the dimension of the image from 784 to 400. Earlier, to store 60000 images, we required a 60000×784 matrix. Now after PCA we require 60000×400 matrix to store information. Here below we shown the reconstructed image using lesser dimension (less information) or using the principal component of the data matrix.

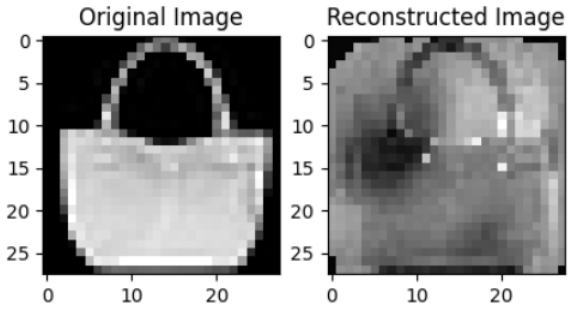

Fig. 5. Reconstructed image

B. K-means Clustering Implementation

we have fitted the dataset after implementing PCA to k-means clustering. In this, we have obtained the labels of

each dataset from k-mean clustering and displayed the list of unique cluster labels. Taking no. of clusters equal to ten, we performed k-means clustering for a hundred number of iterations. For visualization purpose, we are displaying a cluster of sneakers(clustering done on data output after implementing PCA to the original dataset) the given figure 6.

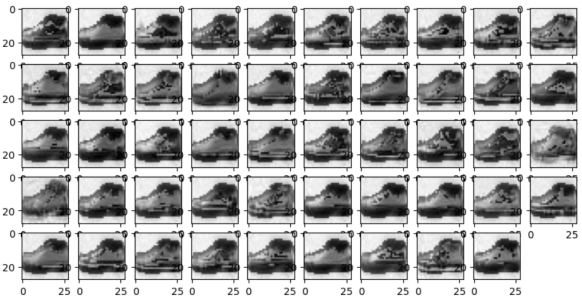

Fig. 6. Reconstructed image

Now here displaying 3-D visualization of the ten clusters in Fig.7.

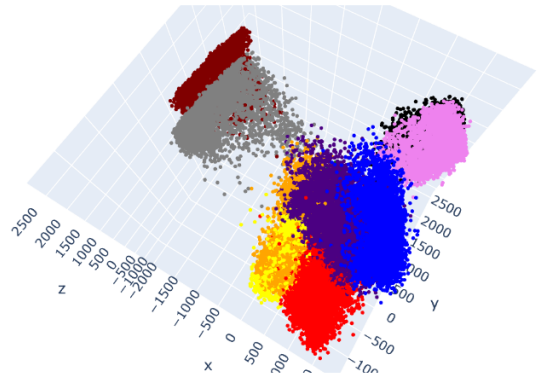

Fig. 7. Visualization after clustering

VI. CONCLUSION

In this project we have implemented PCA dimensinality reduction on an image dataset so that we can compress the dataset into lower size so that K-mean clustering will require less time to execute.PCA combined with K clustering is a basic starting step towards image segmentation which can be used in various fields.This dimensionality reduction and clustering plays a vital role espescially in medical field while brain tumour segmentation or identifying any irregularities in any section of importance.All the output while in the process of implementation of PCA and K clustering have been presented in this report.Future application can be to use this segment of code for image segmentation and any field of importance involving image processing.

VII. REFERENCES

1) Christopher M. Bishop ”Pattern Recognition and Machine Learning”
2) Introduction: https://towardsdatascience.com/k-means-and-pca-for-image-clustering-a-visual-analysis-8e10d4abba40
3) K-means clustering: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
4) K-means algorithm Steps: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning
5) Maths behind k-means clustring: https://muthu.co/mathematics-behind-k-mean-clustering-algorithm/
6) for learning pca: https://www.section.io/engineering-education/image-compression-using-pca/
7) Mathematics behind principal analysis:https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643
8) https://statisticsglobe.com/advantages-disadvantages-pca
9) for learning pca https://www.section.io/engineering-education/image-compression-using-pca/