HOMEWORK 3 -PART 1 & PART 2

NAME : ISHANA SHINDE
[redacted]
V MEASURE SCORE PART 1: 0.95
V MEASURE SCORE PART 2: 0.77
RANK PART 1: 27
RANK PART 2: 60

Introduction:

K-Means Clustering is a simple unsupervised machine learning model.The goal of this assignment is to cluster the iris data in part 1 and perform image clustering of digits in part 2 by implementing K-means algorithm.

Approach:

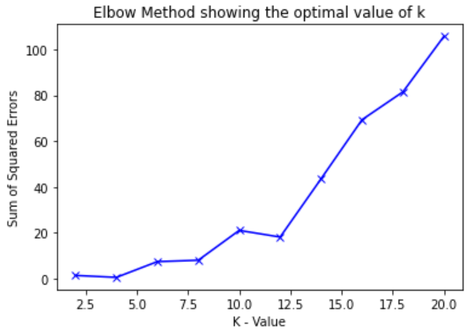
For this assignment as well I divided it into 2 parts -

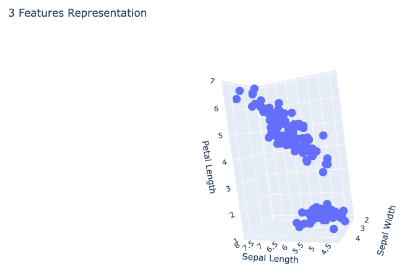
1. Data Preprocessing
2. Implementing K-Means

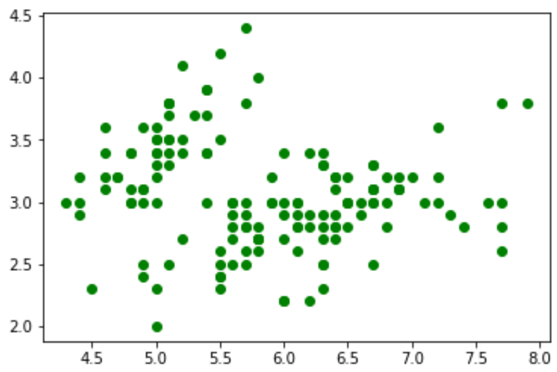
Data Preprocessing:

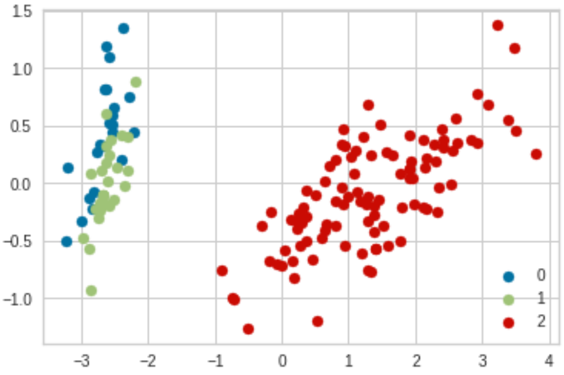
Part 1:

For part 1 of the homework the dataset to be used was the IRIS dataset. The dataset consists of 4 unlabeled features - Sepal Length, Sepal Width, Petal Length and Petal Width. This dataset was to practice the K-Means implementation.







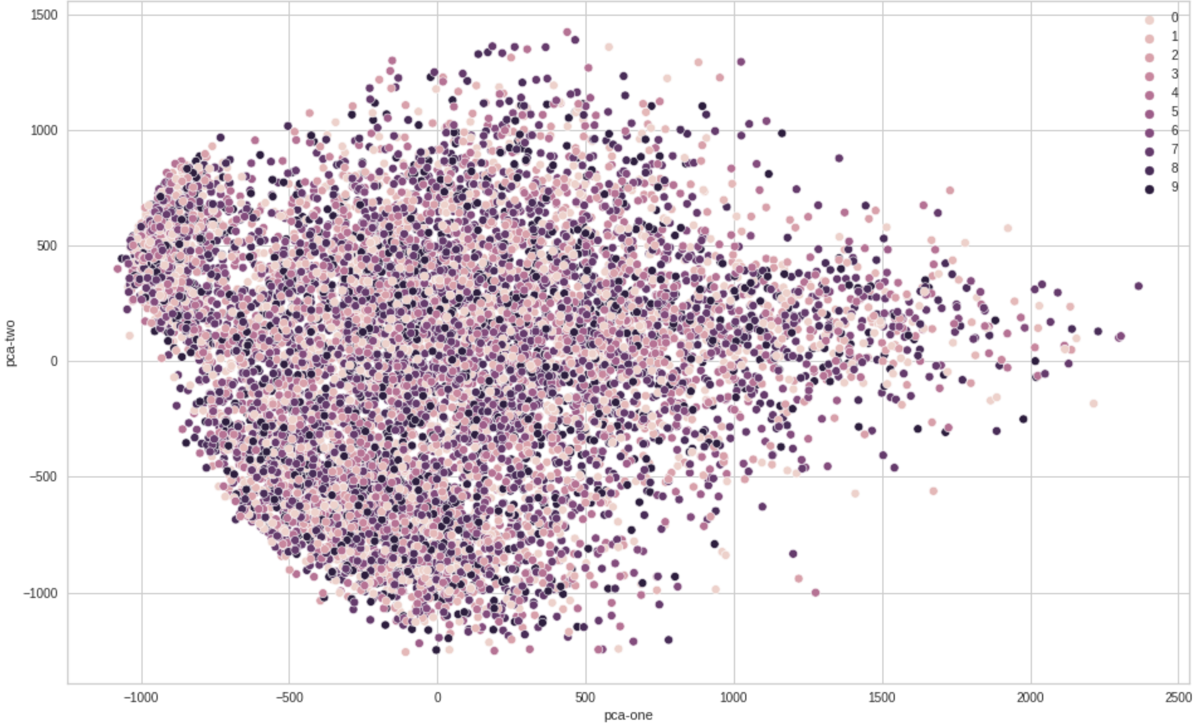


Part 2:

Part 2 of the assignment we use the dataset that consists of 10,000 images of handwritten digits (0-9). The images are scanned and scaled into 28X28 pixels. This gives a 28X28 matrix of integers for each digit when flattened gives a 1X784 vector. Since this dataset is large with data of high dimension, I tried to use PCA (Principal Component Analysis) and t-SNE (T - distributed Stochastic Neighbor Encoding).

1. *PCA (Principal Component Analysis)*

PCA is a technique that is used to reduce the number of dimensions in a dataset while retaining most of the information. It uses the correlation between some dimensions and tries to provide a minimum number of variables that keeps the maximum amount of variation or information about how the original data is distributed.

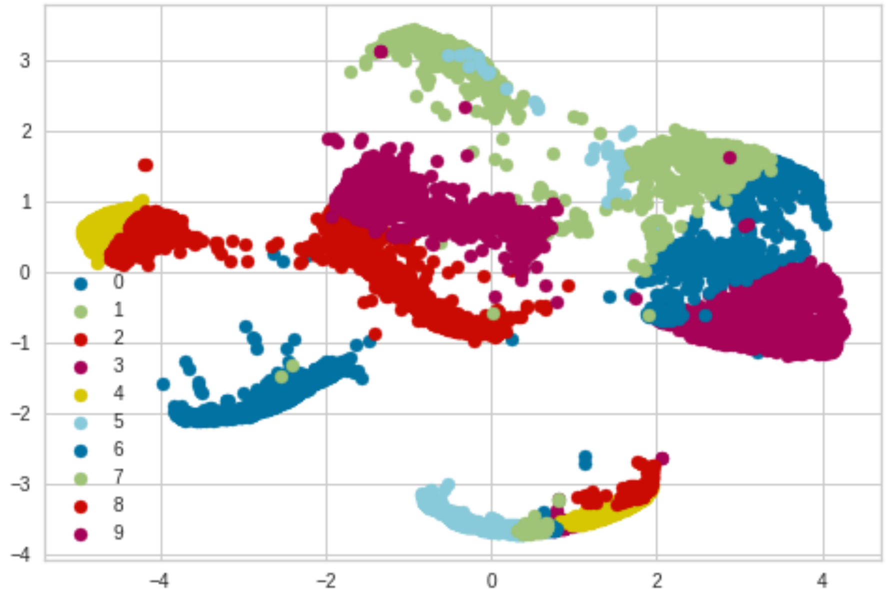


2. *t-SNE (T - distributed Stochastic Neighbor Encoding)*

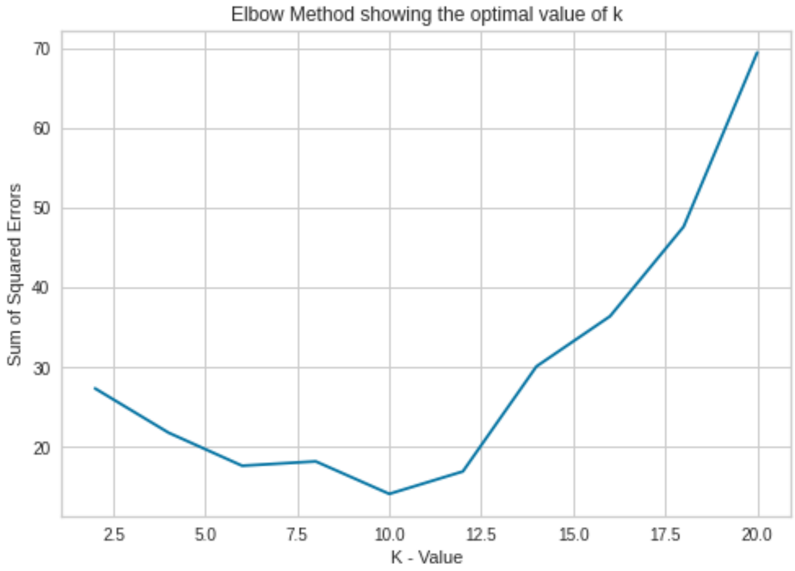
t-SNE stands for T-Distributed Stochastic Neighbor Embedding that is another method for dimensionality reduction and is well suited for the visualization of high-dimensional dataset. This is a probabilistic technique that works by minimizing the divergence between two distributions. This method is computationally heavy so for the high dimensional dataset provided, I first used another dimensionality reduction technique PCA before using t-SNE.

So while implementing PCA the n_components value was changed to 60 and the result generated was passed as input to the t-SNE method with n_components=2, perplexity=8, and n_iter=250.

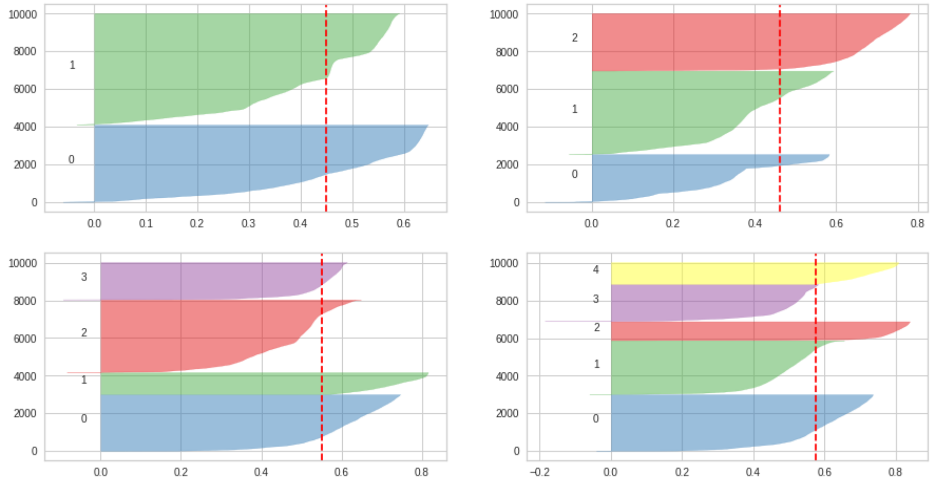
Both methods were implemented using the sklearn library - from sklearn.decomposition import PCA, from sklearn.manifold import TSNE. The result of t-SNE is passed as input to the K-Means code.



Internal Evaluation - Sum of Squared Error(SSE)



Silhouette Coefficient



K-Means Implementation

The simple K-Means algorithm is as follows -

Algorithm 1 *k*-means algorithm

```
Specify the number $k$ of clusters to assign.
Randomly initialize $k$ centroids.
repeat
    expectation: Assign each point to its closest centroid.
    maximization: Compute the new centroid (mean) of each cluster.
until The centroid positions do not change.
```

For my K-Means implementation I have taken random 10 points that will act as the initial centroids for the k-means algorithm. Then I have used the cdist function from scipy library - from scipy.spatial.distance import cdist to calculate the cosine distance of each point from the centroids. For every iteration up to the max iteration and for every cluster we find the new temporary centroid by calculating the mean. These centroids are then added to the centroid list. This loop continues till maximum iterations are achieved. The output clusters from K-Means are stored in the file HW3_pca_tsne.txt.

CONCLUSION:

The final result of this assignment is a clustering algorithm with 77% accuracy. As the dataset was very large and with a large amount of features there is still more room for improvement in terms of dimensionality reduction. As for the K-Means algorithm we could use the MiniBatch K-Means for the image clustering dataset to improve the external metric score. Also instead of randomly assigning the initial centroids, bisecting k-means could be used to overcome some of the initial centroid problems.