CMPE255 - DATA MINING - ASSIGNMENT#2 REPORT

Name: Ishwarya Varadarajan
SJSU ID: 011549473
RANK: 16
F1-score: 82.2

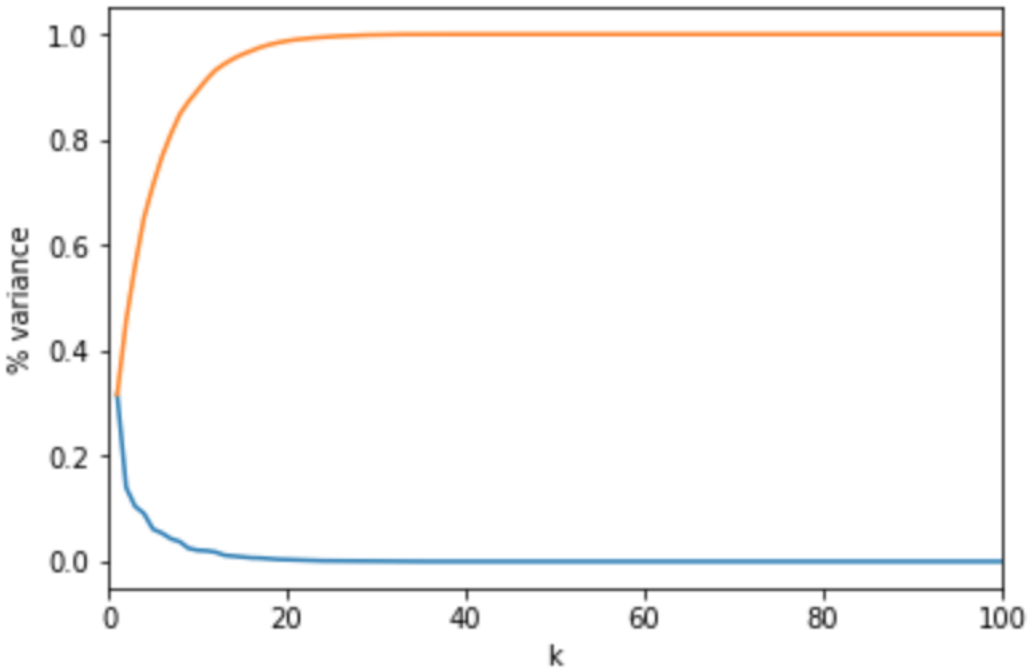
Problem Statement:
To classify images into classes 1 thru 11 using dimensionality reduction techniques and machine learning algorithms to yield high accuracy.

Solution:
The dimensionality reduction technique used is PCA and the machine learning algorithm is XGBoost.

Steps:
1. Data Preprocessing:
   Centered the data by calculating the mean of each feature and subtracting the mean from the original values
2. Dimensionality Reduction:

Principal Component Analysis is used to reduce the number of dimensions from 887 to 28.

- Covariance matrix is computed using the formula (X(Transpose).X)/n-1
- Eigen decomposition is done using the "np.linalg.eig" numpy function and the percentage variance between the 887 features and cumulative percentage variances are calculated
- A graph is plotted using matplotlib to find the number of dimensions exhibiting the maximum variance that is enough to classify an image correctly.
- Here, the above steps are repeated for the train and test data and 30 components as a result



3. Machine Learning Algorithm

- XGBoost classifier is used as the machine learning algorithm and is trained with the training data in train.dat and train.labels files.
- It is then tested with test.dat to predict the classes in the test data based on training data experience
- Parameters given for XGBoost are as follows
  learning_rate=0.3
  n_estimators=2000
  max_depth=7
  min_child_weight=1
  gamma=0
  subsample=0.8
  colsample_bytree=0.8
  objective= 'binary: logistic'
  nthread=4
  scale_pos_weight=1
  seed=27
- Increased the learning_rate from 0.1 to 0.3 to achieve higher F1-measure.
- N_estimators of from 1000 to 2000 to achieve higher F1-measure.
- Rest of the parameters have ideal values normally used for a XGB classifier