ELL409 - Assignment 1

Abhilash Soni 2015EE10422
Mohammad Ali Khan 2015EE30521
Aditi Narware 2015MT10586

I. Problem Statement

The assignment consists of the tasks of building up different kinds of classifiers on various example datasets which includes Fashion-MNIST, Blood Test and Train Selection. Following types of Classification schemes are implemented in Python for each problem:
1. Bayes Classifier (with different class conditional densities and estimation techniques)
2. Naive Bayes Classifier
3. K-means Clustering
4. K-Nearest Neighbor Classifier
5. Principal component analysis (where ever applicable)
Various Classification Performance Parameters such as *Accuracy, Recall, Precision, F1-Score, ROC-curve* also need to be computed.

II. Training and Test Data

A. Fashion-MNIST Dataset

This dataset consists of 60,000 training examples and 10,000 test examples. Each example is a 28x28 pixels gray-scale image. Each image is labeled with 10 class categories (t-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots).
Each image is considered to be 784 dimensional data sample. So Principal component analysis is used for selecting the important features and creating a lower dimensional feature vector for classification task. Here we illustrate few images corresponding to different number of principal components:

B. Blood Test

This dataset consist of outcomes of three Blood Tests (Test1, Test2 and Test3) for analyzing the condition of Heart of a patient. It also contains the doctors advise for whether the Heart is HEALTHY, MEDICATION and SURGERY based on the outcomes of the three tests.

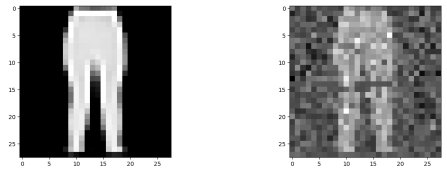
C. Train Selection

The dataset contains the field such as age, sex, fare paid, number of members traveling with, travel class etc. It also contains whether the person has boarded the train or not.

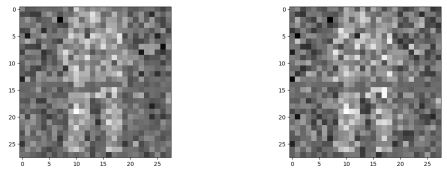
III. Experiments

A. Fashion-MNIST

1) Naive Bayes Classifier: We reduced the dimension of the feature space using PCA. The highest accuracy of 68.17% is obtained for 60 principal components. For no of principal components greater than 80, an overflow is encountered.



(a) 28x28=784 principal components (b) 24x24=576 principal components



(c) 20x20=400 principal components (d) 18x18=324 principal components

Fig. 1: Visualization of same object with different number of principal components

Variation of accuracy Vs number of principal components is shown in Fig. 2

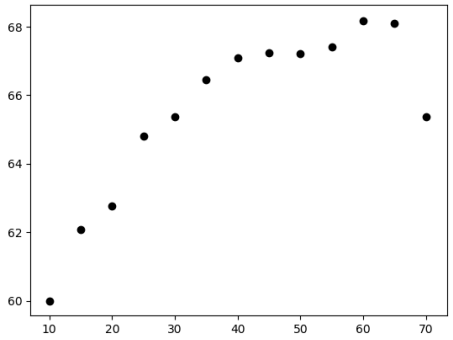


Fig. 2: Variation of Accuracy with different no of principal components

2) Bayes Classifier: Accuracy on test set for Bayes classifier is 81.02% for 65 principal components. Variation of accuracy Vs number of principal components in (3).

We have also computed the confusion matrix for this classifier as shown below:

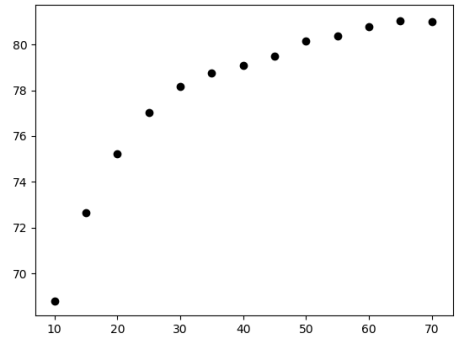


Fig. 3: Variation of Accuracy with different numbers of principal components

$$\begin{bmatrix} 760 & 1 & 7 & 54 & 3 & 4 & 112 & 0 & 59 & 0 \\ 4 & 881 & 3 & 63 & 0 & 0 & 32 & 0 & 17 & 0 \\ 13 & 0 & 680 & 10 & 108 & 1 & 136 & 0 & 52 & 0 \\ 25 & 1 & 6 & 830 & 24 & 2 & 64 & 0 & 48 & 0 \\ 0 & 0 & 101 & 50 & 688 & 0 & 123 & 0 & 38 & 0 \\ 0 & 0 & 0 & 0 & 0 & 962 & 0 & 14 & 17 & 7 \\ 163 & 0 & 96 & 35 & 78 & 0 & 536 & 0 & 92 & 0 \\ 0 & 0 & 0 & 0 & 0 & 131 & 0 & 804 & 0 & 65 \\ 3 & 0 & 1 & 4 & 1 & 8 & 10 & 2 & 971 & 0 \\ 0 & 0 & 0 & 0 & 0 & 50 & 0 & 27 & 2 & 921 \end{bmatrix}$$

3) Bayes Classifier with Bayesian Estimation: Maximum accuracy for this classifier was found to be 80.47% with 85 principal components

4) K-means Clustering: Initial centroids are chosen randomly from the data points. after the clustering is done, the class is assigned based on the majority voting. Accuracy for K-means Clustering was found to be 56.2% with 85 principal components. Variation of accuracy Vs number of principal components is shown in (4).

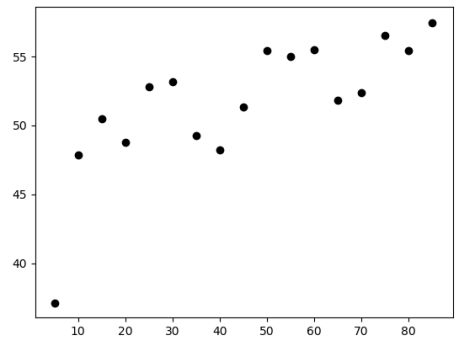


Fig. 4: Variation of Accuracy with different numbers of principal components

5) K-nearest Neighbour Classifier: We found the maximum accuracy 82.54% with 45 principal components. The variation of accuracy Vs different number of principal components is shown in (5). Since the execution of code was taking hours, it was difficult to get further observations. The

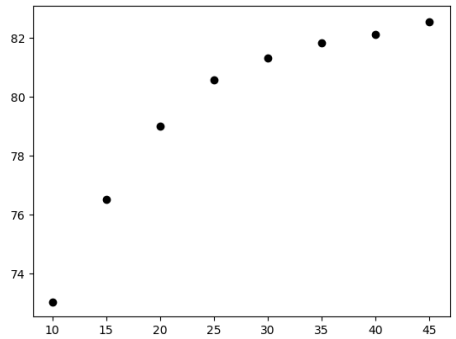


Fig. 5: Variation of Accuracy with different no of principal components

variation of accuracy Vs different values of K is plotted in (6). Maximum accuracy of 73.2% is found for k=20.

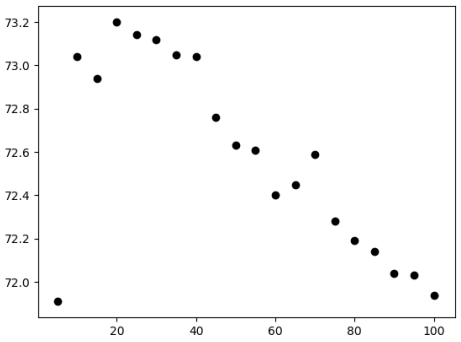


Fig. 6: Variation of Accuracy with k

B. Blood Test

1) Naive Bayes Classifier: This classifier gives 89.76% accuracy.

2) Bayes Classifier: This classifier gives 89.73% accuracy. We have also computed the confusion matrix for this classifier as shown below:

$$M = \begin{bmatrix} 887 & 46 & 67 \\ 13 & 883 & 104 \\ 20 & 58 & 922 \end{bmatrix}$$

3) Bayes Classifier with Bayesian Estimation: This classifier gives 89.7% accuracy.

4) K-means Clustering: Initial centroids are chosen randomly from the data points. after the clustering is done, the class is assigned based on the majority voting. Accuracy for K-means Clustering was found to be 56.07%.

5) K-nearest Neighbour Classifier: Optimum K-nearest neighbour was found at K= 40 and classifier with this K gives accuracy 90.03%. The variation of accuracy with K is shown in (7)

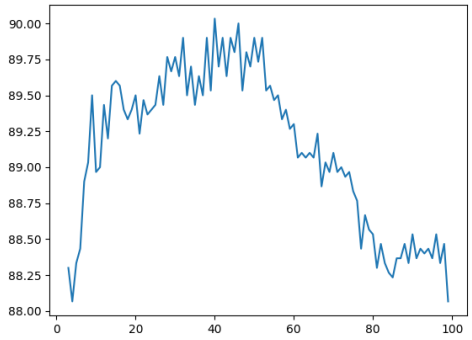


Fig. 7: Variation of Accuracy with different K

C. Train Selection

1) Naive Bayes Classifier: We found the accuracy corresponding to different ratio of training and test data as shown in Table I.

TABLE I: Accuracy corresponding to different ratio of training and test data

Ratio	Accuracy
2:1	74.1418764302
4:1	77.8625954198
5:1	78.5388127854

2) Bayes Classifier: We found the Accuracy corresponding to different ration of training and test data as shown in Table II. We have also computed the confusion matrix for this classifier as shown below:

$$M = \begin{bmatrix} 22 & 15 \\ 27 & 155 \end{bmatrix}$$

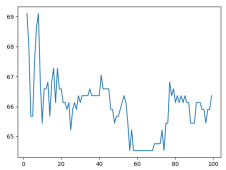
TABLE II: Accuracy corresponding to different ratio of training and test data

Ratio	Accuracy
2:1	74.828375286
4:1	79.7709923664
5:1	80.8219178082

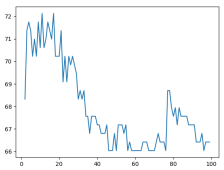
3) K-means Clustering: Initial centroids are chosen randomly from the data points. after the clustering is done, the class is assigned based on the majority voting. Accuracy for K-means Clustering was found to be 74.31% for the ratio of training set to test set as 2:1.

4) K-nearest Neighbour Classifier: We divided our data set into two parts one part for the training and one for the testing. Corresponding to each training and test data pair, a graph of K Vs the accuracy obtained is plotted (fig 8) and using which we found the optimal value of K (table III)

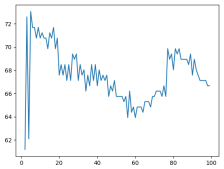
From the above trends, we can observe that as we take the higher ratio, i.e. more number of data points in our training dataset, the accuracy increases.



(a) ratio 2:1



(b) ratio 4:1



(c) ratio 5:1

Fig. 8: Variation of Accuracy with different K

TABLE III: Optimal 'K' and accuracy corresponding to different ratio of training and test dataset

Ratio	Optimal K	Accuracy
2:1	3	69.1075514874
4:1	12	72.1374045802
5:1	6	73.0593607306

IV. Performance Analysis

We have calculated following performance parameters for Baysian Classifier

A. Fashion-MNIST

Shown in Table IV

TABLE IV: Parameters for Fashion-MNIST dataset

Class	Precision	Recall	F-Score
Class1	0.7806841	0.776	0.77833501
Class 2	0.99774011	0.883	0.93687003
Class 3	0.81556886	0.681	0.74223433
Class 4	0.817560981	0.838	0.82765432
Class 5	0.78463329	0.674	0.72512103
Class 6	0.81314879	0.94	0.87198516
Class 7	0.5293578	0.577	0.55215311
Class 8	0.91464821	0.793	0.84949116
Class 9	0.74110522	0.979	0.8436019
Class 1	0.93595041	0.906	0.92073171

B. Blood Test

Shown in Table V

TABLE V: Parameters for Blood Test dataset

Class	Precision	Recall	F-Score
Class1	0.96413043	0.887	0.92395833
Class 2	0.89463019	0.883	0.88877705
Class 3	0.84354986	0.922	0.88103201

C. Train Selection

Shown in Table VI

TABLE VI: Parameters for Train Selection dataset

Class	Precision	Recall	F-Score
Class1	0.50806452	0.5625	0.53389831
Class 2	0.84294872	0.8117284	0.82704403

APPENDIX

The github link of the repository is https://github.com/makiit/ell409