Assignment 4 Report

Team-01

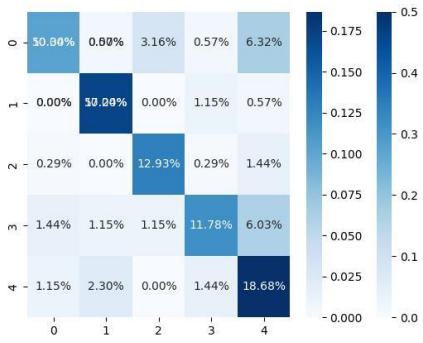
1.KNN

1.1 Synthetic Data

- When KNN was performed for the synthetic data , accuracy was 100% for k=5
- When the dimensions are reduced to 1 , using LDA accuracy was 90% and using PCA accuracy was 100%

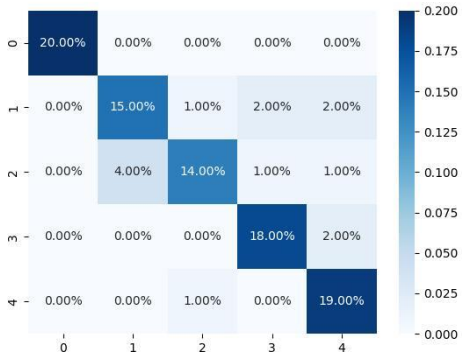
1.2 Image Data

- When KNN was performed on Image data , accuracy was 70% for k=15
- When LDA is used to reduce the dimension , accuracy was 55% with only 4 features used
- When PCA is used , accuracy was 69.9 % with 80 features used
- Confusion matrix for direct KNN for k = 15 is given below



1.3 Hand Written Data

- When KNN was performed on hand written data , accuracy was 86% for k = 10
- When LDA is used to reduce dimension , accuracy was 20% with 4 features
- When PCA is used , accuracy was 86% with only 10 features used
- Confusion matrix when PCA was performed with 10 features is given below



1.4 Spoken digits

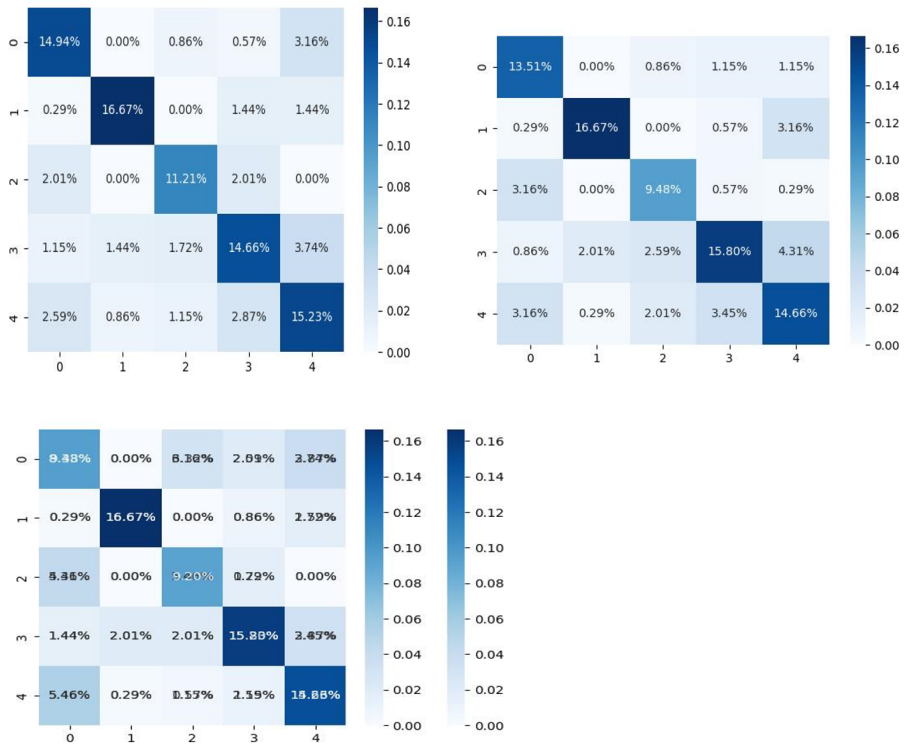
- When KNN was performed on Spoken digits , accuracy was 80% for k = 10
- When LDA is used to reduce dimension, accuracy was 20% with 4 features
- When PCA is used, accuracy was 78% with only 10 features used

2 Logistic Regression

- Logistic regression is implemented for multiple classes using softmax function and gradient descent to estimate parameters.
- Given below are results of logistic regression on different data

2.1 Real Image Data set

- When LR was performed on this dataset by keeping all the 828 features, this classifier gave an accuracy of 72%.
- When PCA was used to keep the accuracy of 70% only 35 components were required.The speed of the algorithm has improved by a large factor keeping the accuracy.
- When LDA was used to keep only c - 1 = 4 features, the accuracy was 66%.
- The confusion matrices for these 3 cases are:

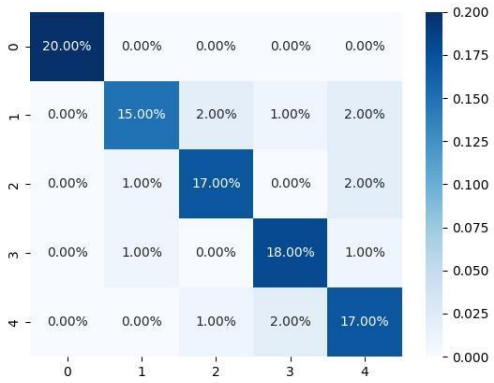


2.2 Synthetic Data

- When LR was performed on synthetic data with degree 2, accuracy was 90.3%.
- When PCA and LDA are used to reduce to 1 feature the accuracies were 88.5% and 88.9%.
- PCA and LDA had almost the same accuracy with one feature less in each example.

2.3 Handwritten Letters

- LR gave an accuracy of 87% on handwritten letters.
- Confusion matrix for this is as follows:



- When PCA was used to reduce features to 10, accuracy still was 87%.
- When LDA was followed by PCA to reduce features to 4, accuracy became 64%.

2.4 Spoken Digits

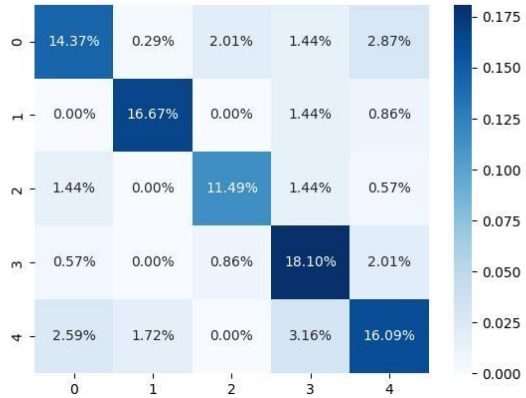
- When LR is trained on spoken digits accuracy of 85% is obtained.
- When PCA is used to reduce features to 10, accuracy became 81.6%
- When LDA is used to further reduce features to 4, accuracy of 81.6% is observed again.

3 SVM

- Inbuilt svm function from sklearn module is used to classify the datasets with radial basis kernel function
- The results for different datasets are as follows:

3.1 Real Image Data set

- SVM gave an accuracy of 76.7%.
- The confusion matrix for it is:



- When PCA is used to reduce features to 30, accuracy was 75.86% which is very close.
- When LDA is used to reduce features to 4, accuracy is 70%.

3.2 Synthetic Data set

- On synthetic data, SVM gave an accuracy of 100%.
- When PCA is used to reduce features to 1, accuracy became 88.8%
- When LDA is used accuracy was 88.4%.

3.3 Handwritten Letters

- SVM gave an accuracy of 93% when trained on this data.
- PCA and LDA neither improved accuracy nor improved speed.

3.4 Spoken Digits

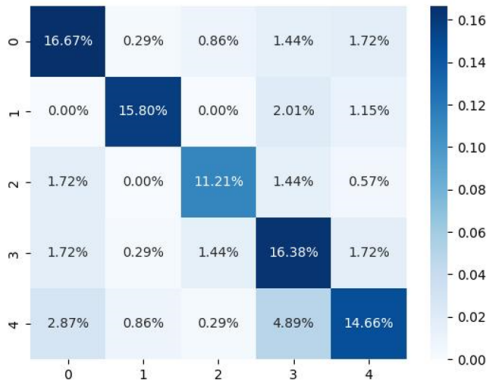
- SVM gave an accuracy of 92% when trained on this data.
- PCA and LDA neither improved accuracy nor improved speed.

4 ANN

- Multilayer perceptron Classifier from sklearn is used to classify given datasets.
- Activation functions used are different for different sets.

4.1 Real Image Data set

- Activation function tanh worked better when compared to others.
- Hidden layers of sizes 33 and 12 gave better results which had an accuracy of 74.7%



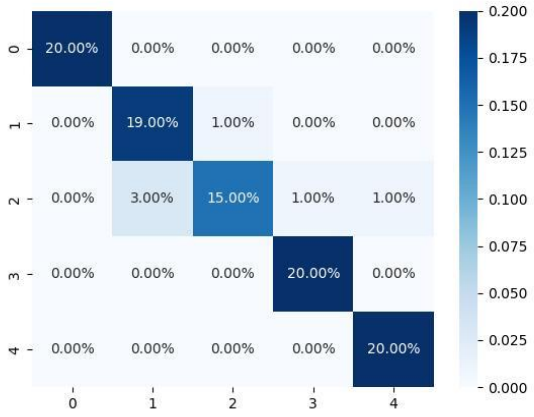
- Performing PCA or LDA neither improved accuracy nor improved speed in this case.

4.2 Synthetic Data set

- Activation function rectified linear function worked better.
- With hidden nodes in one layer = 70, accuracy of 99.9% was obtained on development data.

4.3 Handwritten Letters

- With activation function tanh and hidden layers of sizes 100 and 33, ANN gave an accuracy of 94% on this data.
- Confusion matrix is as follows:



- PCA and LDA neither improved accuracy nor improved speed.

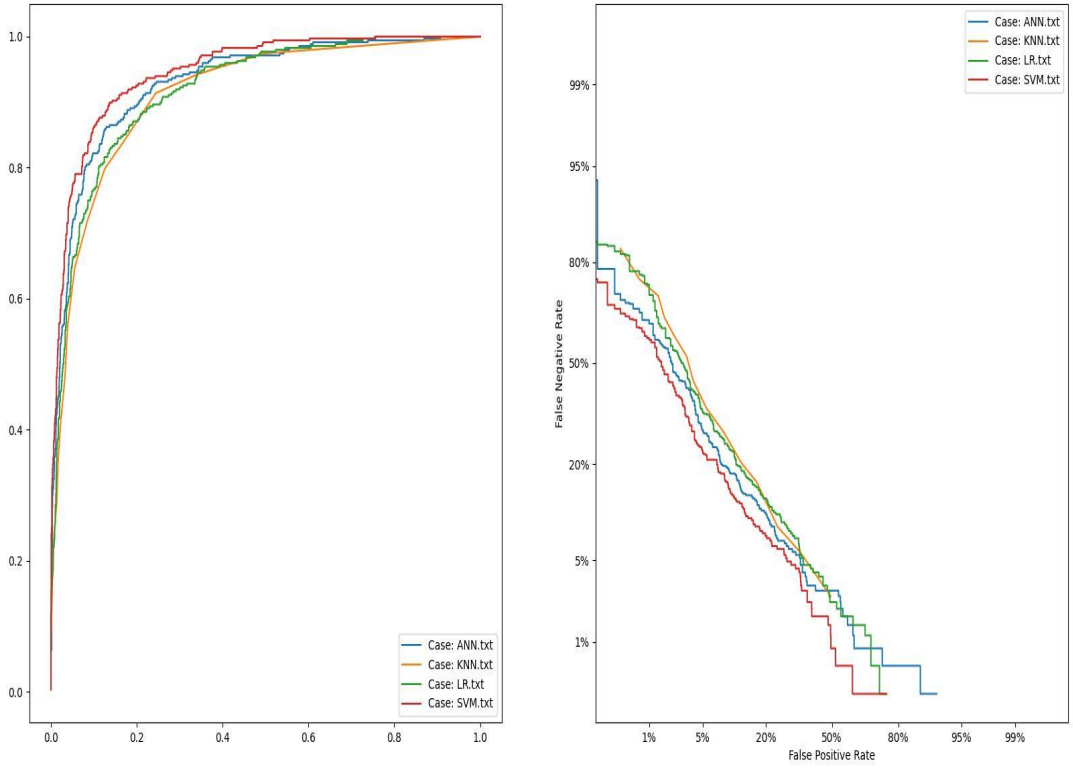
4.4Spoken Digits

- With activation function tanh and hidden layers of sizes 80,20, ANN gave an accuracy of 90% on this data.
- PCA and LDA neither improved accuracy nor improved speed.

5 Performance Comparison

5.1 Real Image Data

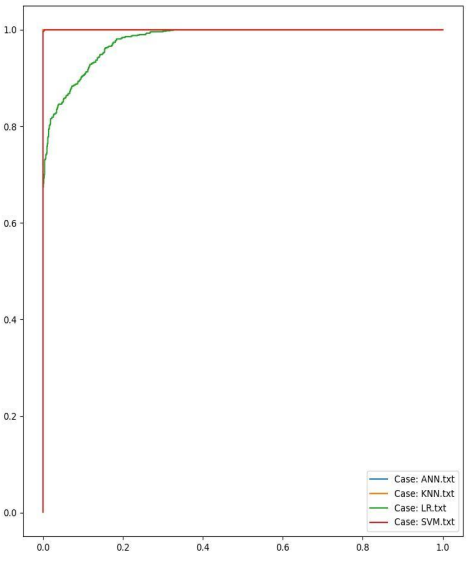
For this data the ROC and DET curves for various classifiers with their best case parameters is as follows:

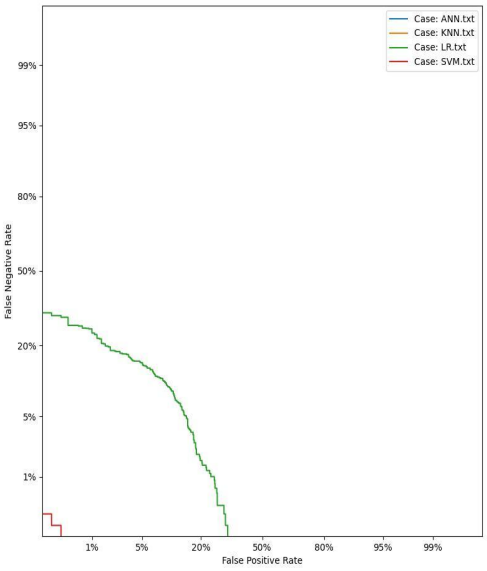


Performance of SVM > ANN > LR > KNN on this data.

5.2 Synthetic Data

ROC and DET:

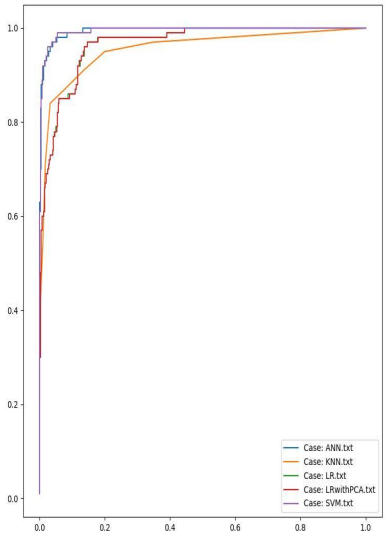


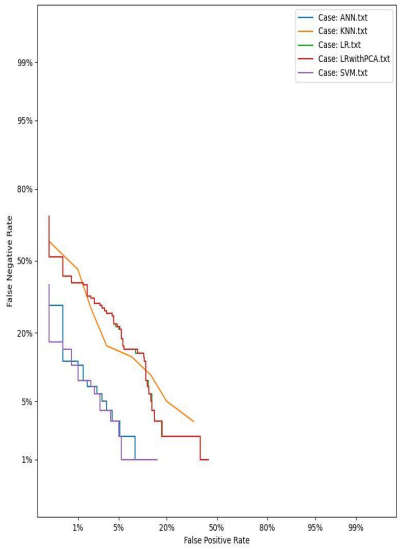


Performance ANN ~ KNN ~ SVM > LR.

5.3 Handwritten Data.

ROC and DET:

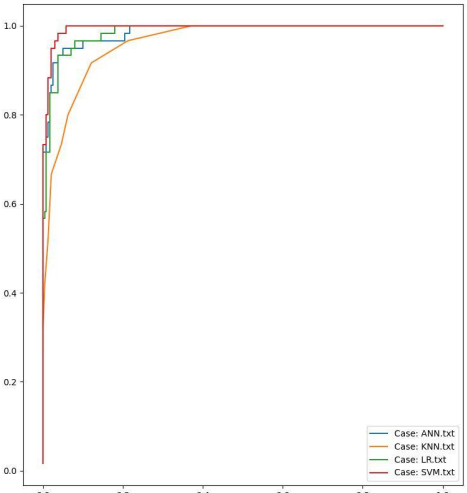


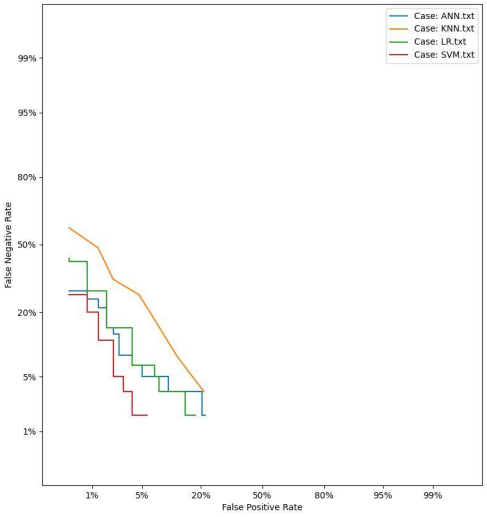


Performance: ANN ~ SVM > LDA with PCA ~ LDA > KNN.

5.4 Spoken Digit Data.

ROC and DET:





Performance SVM > ANN > LR > KNN.