CS5691 Assignment 4 Report - Team 14

Aswin Ramesh cs19b007, Ch Ajith Reddy cs19b014

Logistic Regression

1. **Synthetic dataset:**

   The logistic regression model for the synthetic dataset performs decently well on the test points(with accuracy around 90-92%, learning rate = 0.000001 and 20 iterations of gradient descent). The scatter of the points is shown in Fig-2.1, the colors of the points indicate the class of the points which is classified by the model. The linear boundary is cutting both the 3's. Since the 3's are slightly overlapping, we cannot obtain a 100% classifier with the logistic regression model. The ROC and DET curves are shown in Fig-2.2.

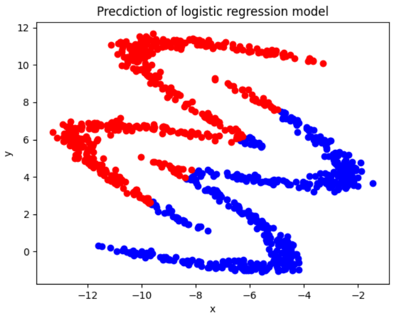

Fig 1.1 - Scatter of the points

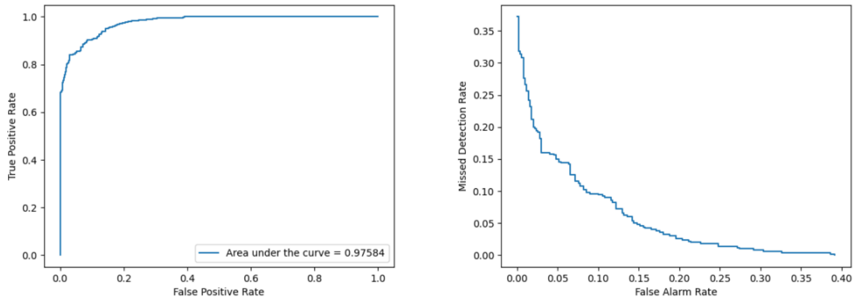

Fig 1.2 - ROC(left) and DET(right) for synthetic dataset

2. **Image dataset:**

   The logistic regression model for the image dataset was not performing so well(accuracy around 60-70%). For multiclass regression, one vs all classifier has been used in the code(Take one class as 0 and rest others as 1 and use a binary logistic regression classifier). Since we can't guess the misprediction class (i.e) if a point is wrongly predicted we cannot say the class predicted by the model(since wrong prediction implies the point belongs to all class). The ROC and DET curves are shown in Fig-2.3.

   Since this is one vs all, we cannot get a confusion matrix. One observation here is that area under ROC(all classes) is slightly higher than the accuracy. This is because for a one vs all case, the accuracy of predicting it belonging to the class is lesser but the accuracy of it predicting a point not belonging to the class is higher and hence the reason for high areas under the ROC curve. The ROC, DETs and accuracies for all 3 possible cases(without PCA/LDA, with PCA and with LDA) are taken into account, the graphs shown are for the best predicting model.

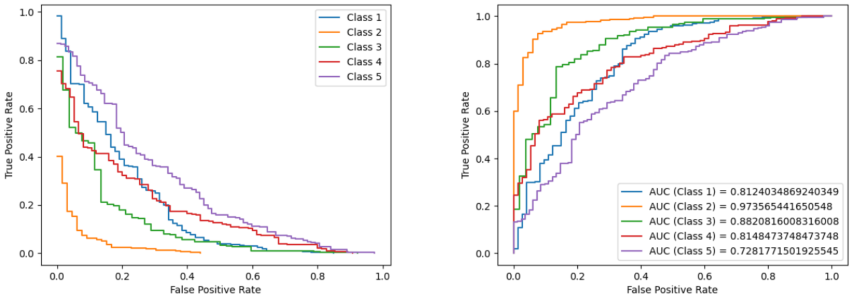

Fig 1.3 - ROC(right) and DET(left, labelling wrong - false alarm rate vs missed detection rate) for individual classes

3. **Isolated Spoken Digit dataset:**

   The logistic regression model for the Isolated Spoken Digit dataset is around 80-85%. This same amount of accuracy is obtained even by reducing this data to 10 dimensions using PCA. Hence PCA performs well on this dataset(since it can reduce a lot of space and provide the same amount of accuracy). The ROC and DET curves are shown in Fig-2.4.

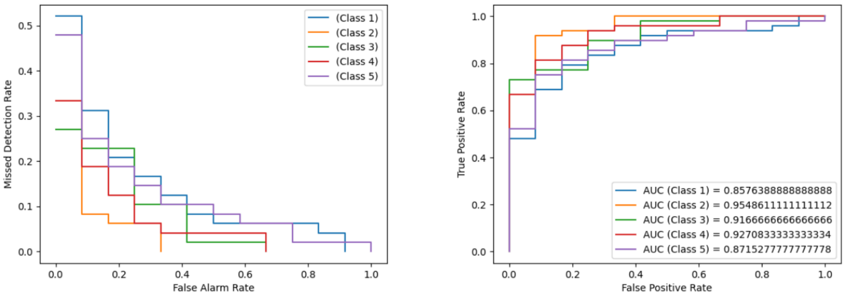

Fig 1.4 - ROC(left) and DET(right)

4. **Online Handwritten-Character Dataset:**

   The logistic regression model for the Online Handwritten-Character Dataset performs very well. The accuracy is about 90-95%. The same case here as well, PCA performance is high(similar accuracy for a less dimensions data). The ROC and DET curves are shown in Fig-2.5.

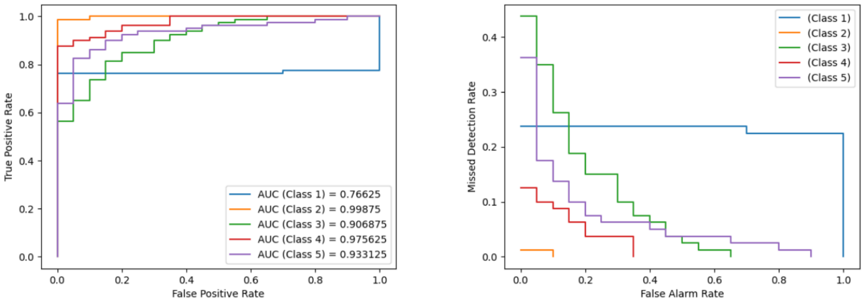

Fig 1.5 - ROC(right) and DET(left)

Support Vector Machine

1. **Synthetic dataset:**

   The SVM model works well on the synthetic datset. The accuracy varies for different choices of kernel function. RBF(Radial Basis Function) kernel has higher accuracy(99.7%), degree 10 polynomial kernel has

98.8% accuracy and the linear kernel has 92% accuracy. There is also this factor of penalty for every wrong prediction on the model(we can set this parameter), higher penalty results in a much better model.(This does not affect the linear model since the data cannot be linearly seperable).

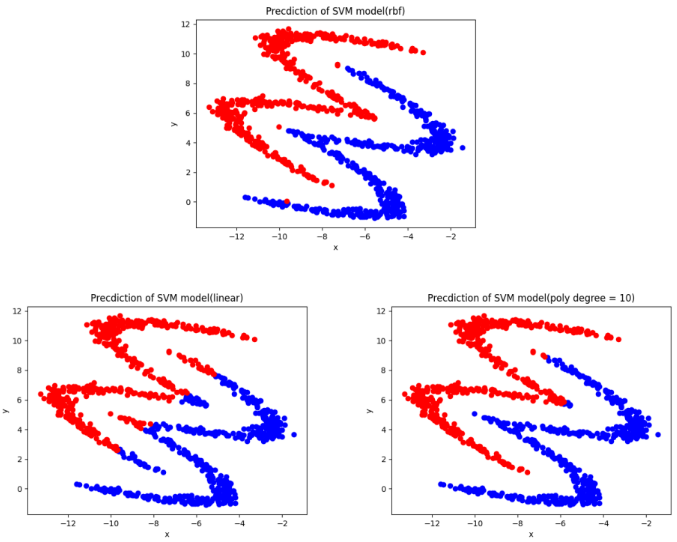

Fig 2.1 - Scatter of different SVM models

ROC and DET plots for the RBF kernel SVM are shown in Fig-3.2.

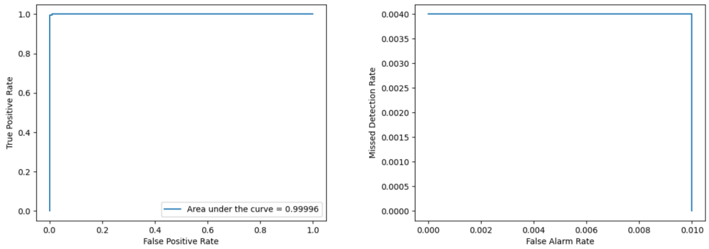

Fig 2.2 - ROC(left), DET(right)

2. **Image dataset:**

The accuracy of the SVM model on the image dataset is not so well(around $60 - 65\%$). The RBF kernel has been chosen to classify for the SVM model. The confusion matrix is shown in Fig-3.3 and the ROC DET plots are shown in Fig-3.4. Applying PCA and LDA did not help the model much.

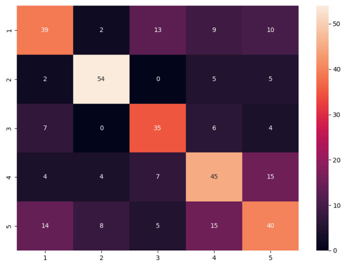

Fig 2.3 - Confusion Matrix (x-axis is actual class, y-axis is predicted class)

3. **Isolated Spoken Digit dataset:**

The accuracy of the model on the Isolated Spoken Digit data is around 90%. The confusion matrix is shown in Fig-3.5 and the ROC and DET plots are shown in Fig-3.6. PCA did not reduce the accuracy by much.

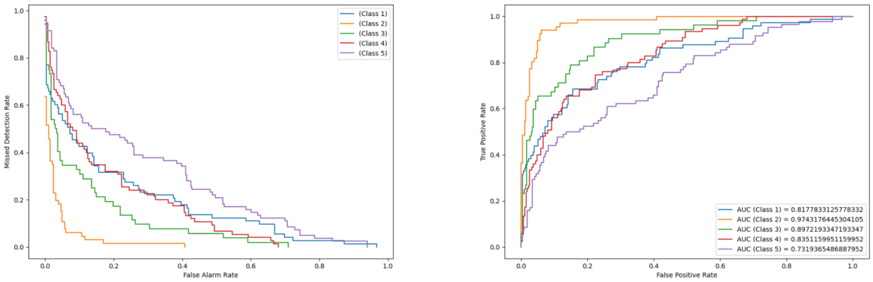

Fig 2.4 - ROC(right), DET(left)

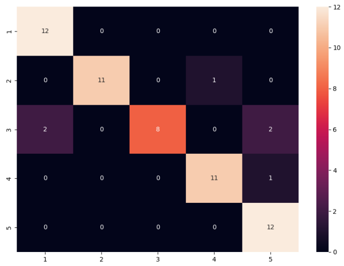

Fig 2.5 - Confusion Matrix (x-axis is actual class, y-axis is predicted class)

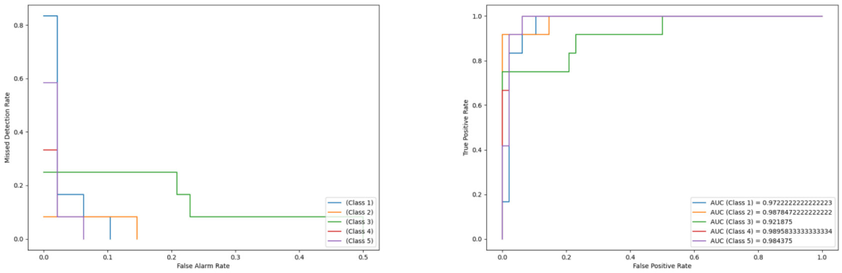

Fig 2.6 - ROC(right), DET(left)

4. **Online Handwritten-Character dataset:**

   The accuracy of the model on the Online Handwritten-Character data is extremely good 98%. The confusion matrix is shown in Fig-3.5 and the ROC and DET plots are shown in Fig-3.6. PCA did not reduce the accuracy by much.

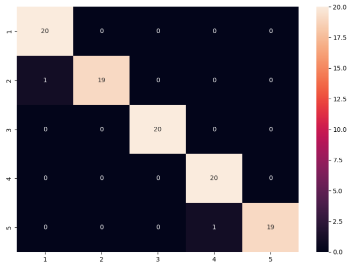

Fig 2.7 - Confusion Matrix (x-axis is actual class, y-axis is predicted class)

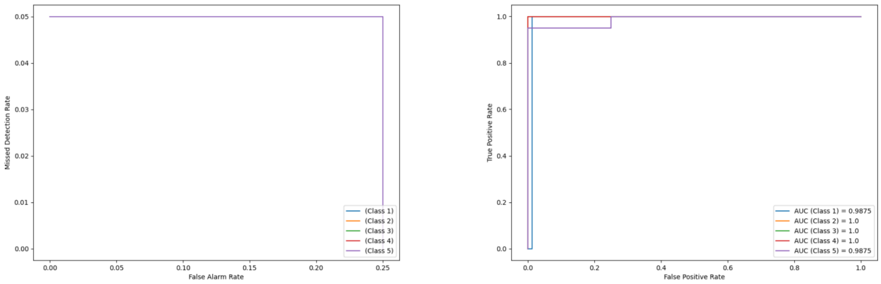

Fig 2.8 - ROC(right), DET(left)

Artificial Neural Networks

1. **Synthetic dataset:** Accuracy of 100% was found
2. **Image dataset:** Accuracy of 64% was found
3. **Isolated Spoken Digit dataset:** Accuracy of 82% was found
4. **Online Handwritten-Character dataset:** Accuracy of 97% was found

K Nearest Neighbour

1. **Synthetic dataset:** Accuracy of 100% was found
2. **Image dataset:** Accuracy of 63% was found
3. **Isolated Spoken Digit dataset:** Accuracy of 82% was found
4. **Online Handwritten-Characted dataset:** Accuracy of 74% was found

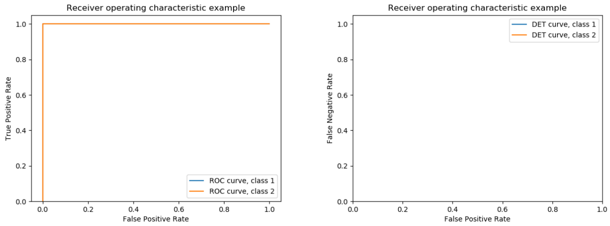

Fig 3.1 - ROC(left) and DET(right) curves for synthetic data using ANN

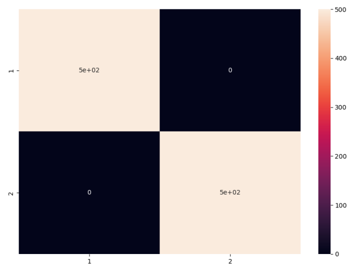

Fig 3.2 - Confusion matrix for synthetic data using ANN

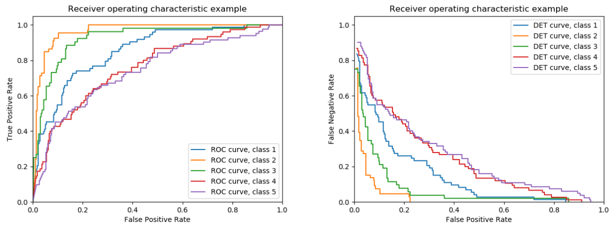

Fig 3.3 - ROC(left) and DET(right) curves for image data using ANN

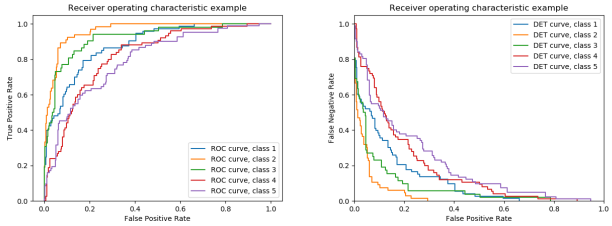

Fig 3.4 - ROC(left) and DET(right) curves for image data using ANN and PCA

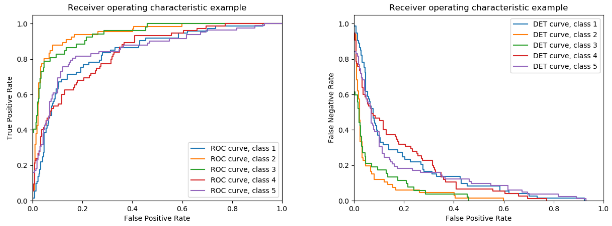

Fig 3.5 - ROC(left) and DET(right) curves for image data using ANN and LDA

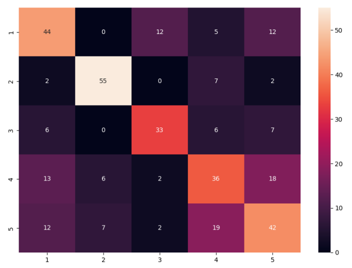

Fig 3.6 - Confusion matrix for image data using ANN

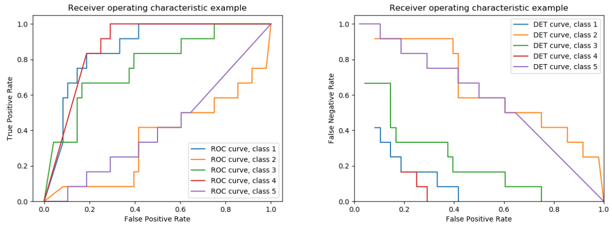

Fig 3.7 - ROC(left) and DET(right) curves for digits data using ANN

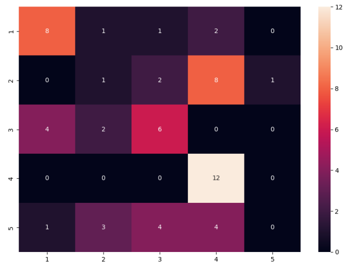

Fig 3.8 - Confusion matrix for digits data using ANN

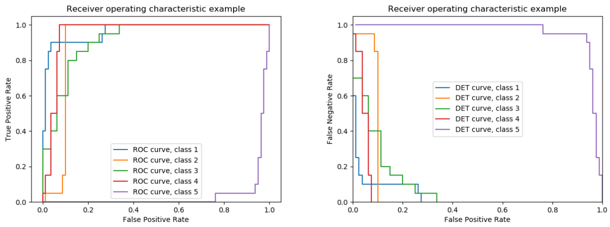

Fig 3.9 - ROC(left) and DET(right) curves for handwritten data using ANN

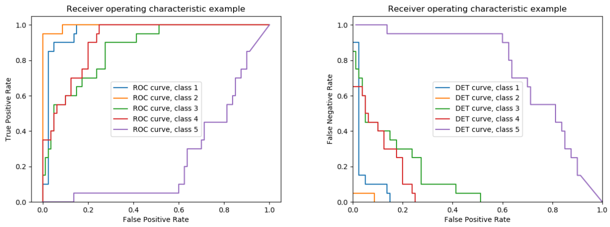

Fig 3.10 - ROC(left) and DET(right) curves for handwritten data using ANN and LDA

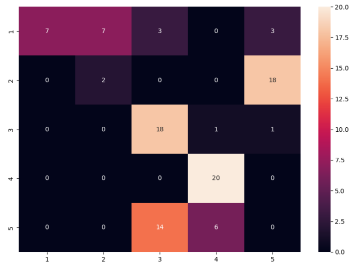

Fig 3.11 - Confusion matrix for handwritten data using ANN

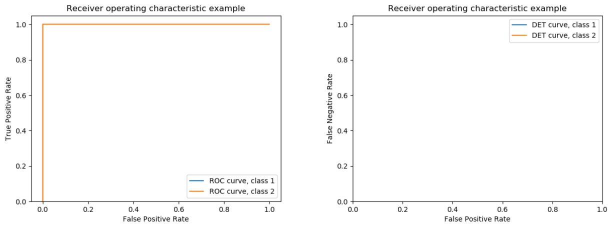

Fig 4.1 - ROC(left) and DET(right) curves for synthetic data using KNN

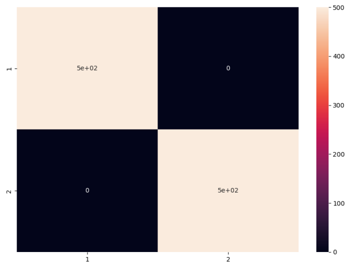

Fig 4.2 - Confusion matrix for synthetic data using KNN

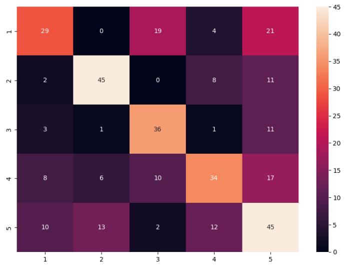

Fig 4.3- Confusion matrix for image data using KNN

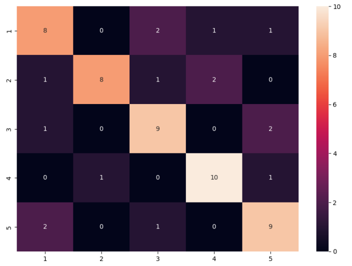

Fig 4.4 - Confusion matrix for digit data using KNN

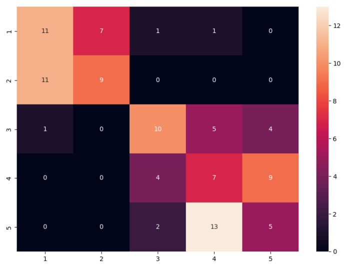

Fig 4.5 - Confusion matrix for handwritten data using KNN