CS5691 - ASSIGNMENT 4

Maddula Jaya Kamal (CS19B028), Vedaant Alok Arya (CS19B046)

May 2, 2022

Data Processing: The data given to us in its direct form was not suitable to use for the required algorithms. So it had to be properly processed to make it suitable. Synthetic Data was good and was ready to use in it's given form. But the audio and the hand-written characters data needed to be made into single feature vector, as they had multiple feature vectors and whose count is different for every case of data. And also the hand-written characters needed to be normalised as scale of every character was different. You can use either z normalisation or min-max scaling, but the data from z-normalisation was smooth and properly handled outliers. After this the number of feature vectors was needed to make equal for all cases in a model to properly compute cost functions. So you pick the case with least number of feature vectors and then do **Rolling Average** with window a specific size to make all cases of same length.

For image data, we have run PCA and LDA to reduce the 828 dimensions down to manageable sizes. Prior to running PCA, we need to ensure the mean is zero - so we have also normalized the given data (to zero mean unit variance) over the training and development batches individually - this limits the accuracy of the model when individual samples are given, yet there was a significant difference across multiple models for this data. Normalizing separately provided a much better accuracy (KNNs, SVMs, ANNs).

KNN: KNN is the K-nearest neighbour Algorithm. After processing the datasets, you can implement KNN algorithm with a suitable cost function to get desired results. In our cases the frobenius norm can be used as the cost function for all datasets. While choosing a value of K make sure, you don't pick really large K nor an extremely small K. By ensuring these we can avoid a lot of misclassifications as small K would behave badly in case of outliers and large K might include unnecessary points into consideration.

For synthetic data most values of K gave good results with a hundred percent accuracy. But when the value of K approaches total dataset size then there might be slight misclassification but most cases any value of K would gave good results for synthetic data.

For image data the best accuracy achieved was 42%, for K=13 with individually normalized train and dev sets, and considering only the patches 10 and 11 of given images (ignoring the rest). Changing any of the given settings yielded a drop in performance of

at least 10% and in the worst going down to 15-20%. However, this accuracy is competent when compared to other models.

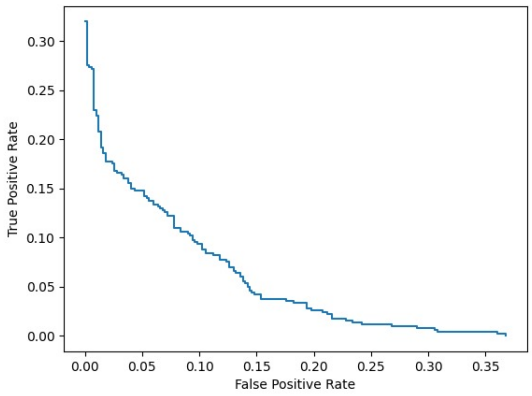
For Hand-written characters KNN performed good with an accuracy of greater than 85 percent accuracy. The variation K vs Accuracy has no direct trend and is purely dependent on the nature of the data.

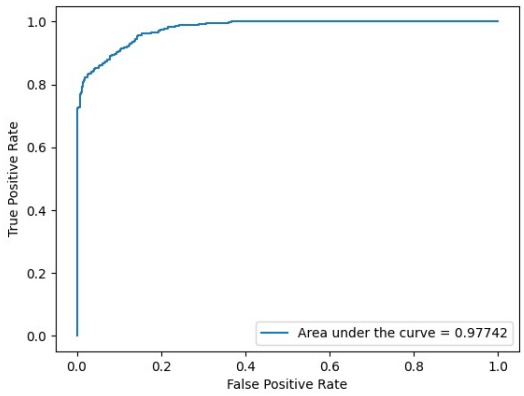
For audio dataset the accuracy obtained for different values of K are in the range of 70 to 80 percent, without any direct trend. So we just had to randomly try multiple values of K and then pick the optimal value.

PUT GRAPHS HERE

SVM: An SVM helps you to classify data by fitting in a hyper-plane between classes.We use the inbuilt function of SKLearn. With these we can try various kernels like 'linear', 'poly' and 'rbf' to increase accuracy.

For synthetic data the choice of the kernel didn't matter much as the accuracy was almost same with 90.7 percent for all kernel modes. LDA or PCA were not useful as the data was already of lower dimension and was clearly separable.





For image data, rbf performed the best with other configuration as described for KNN (normalizing, etc) - except taking all the patches instead of 10-11. Accuracy achieved is 79%, with rbf $>$ linear $>$ poly.

For Hand-written characters SVM performed good with an accuracy of greater than 90 percent accuracy. All the 3 kernels performed good but the 'rbf' one stood out with an accuracy of 97 percent.

For the audio dataset the best accuracy we managed to get is 88.33 percent using a 'linear' kernel. The accuracy of the other kenels was also in the similarr range.

Logistic Regression: Below are the ROC and DET graphs for logistic regression of synthetic data.