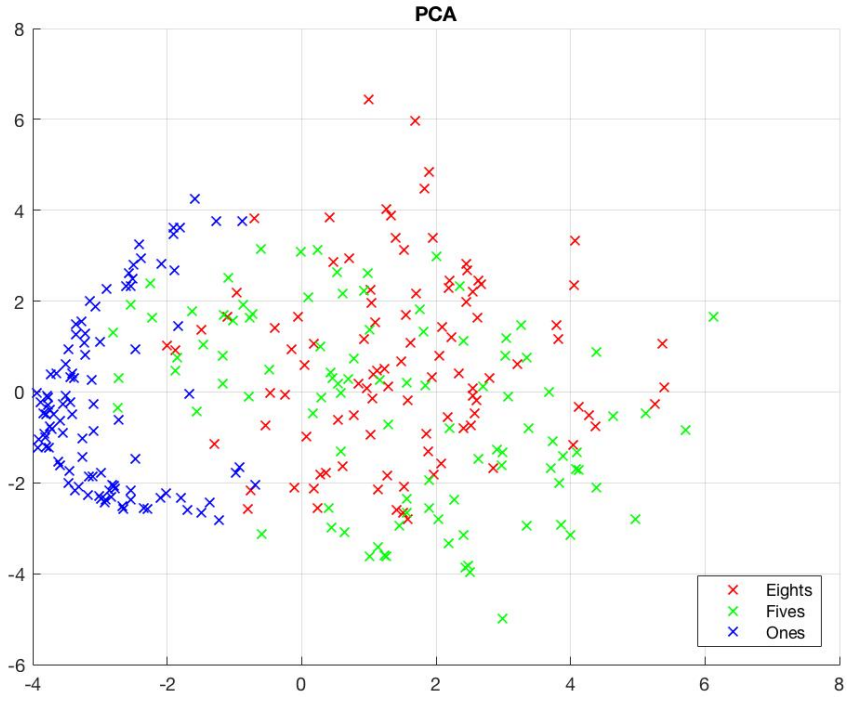
AC50001 Introduction to Data Mining and Machine Learning
Assignment 2
Julia Gratsova

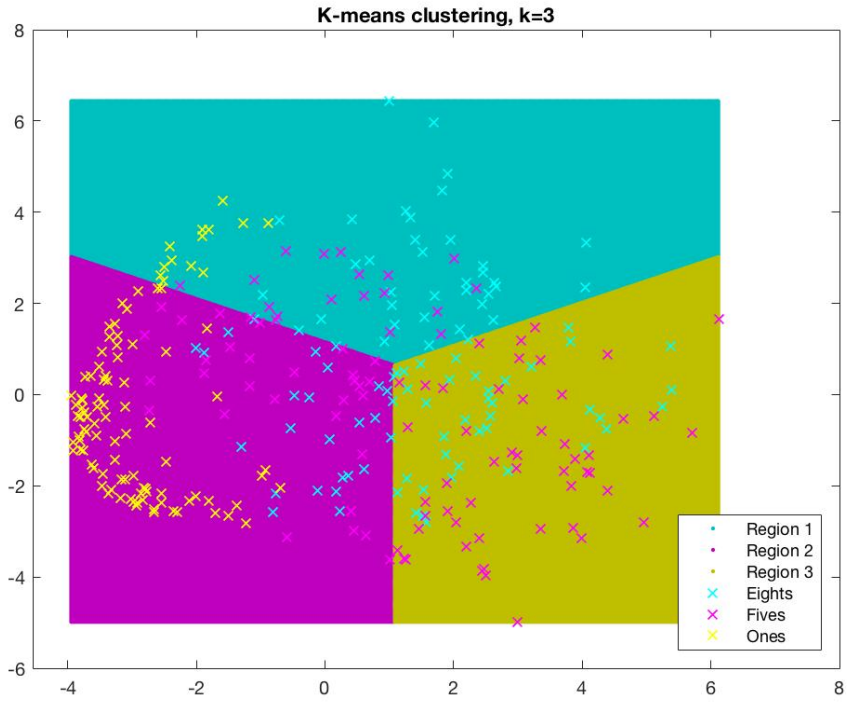
1. PCA

The purpose of this example is to show the PCA dimensionality reduction using the Matlab's built-in PCA function. K-means and Hierarchical clustering algorithms are then used to compare the PCA scores.

K-means clustering

K-means clustering was chosen to separate the data into 3 clusters and performed with the following results:

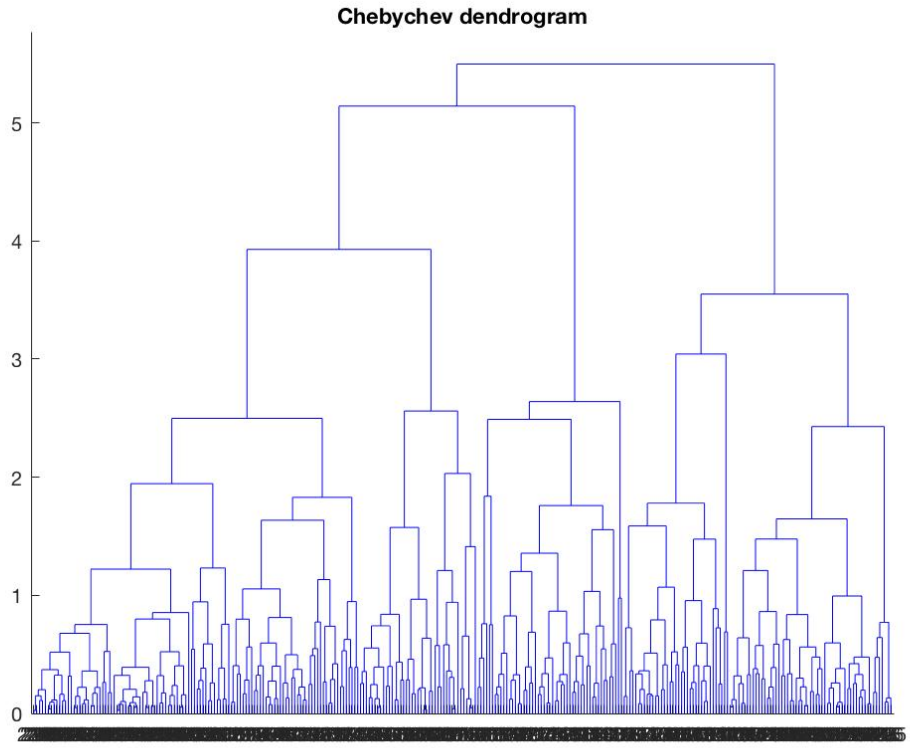


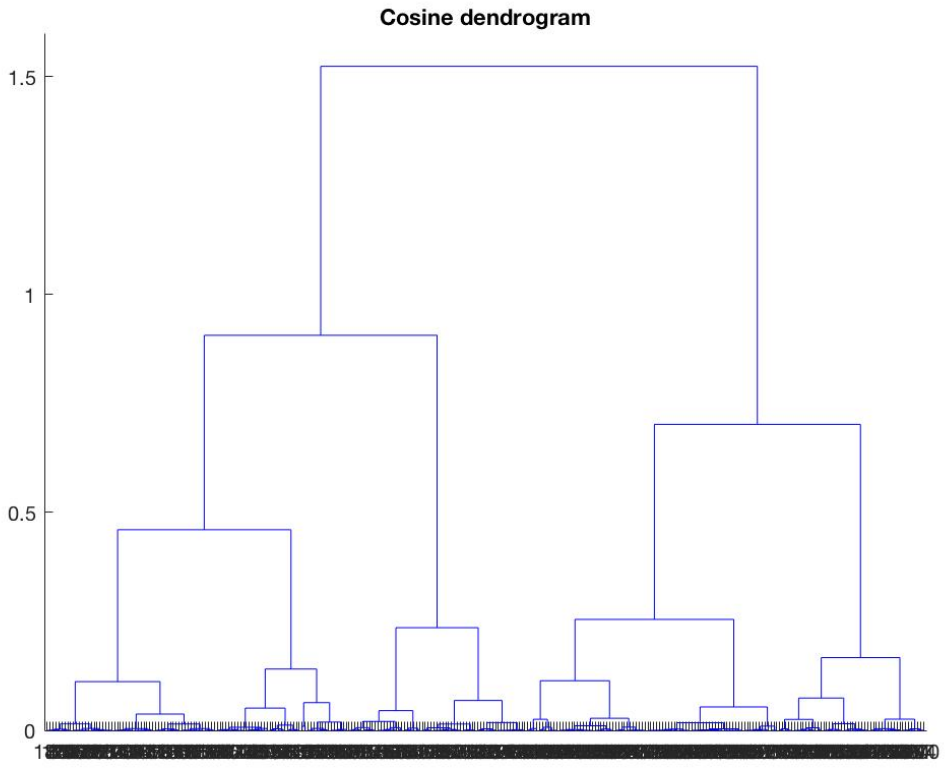


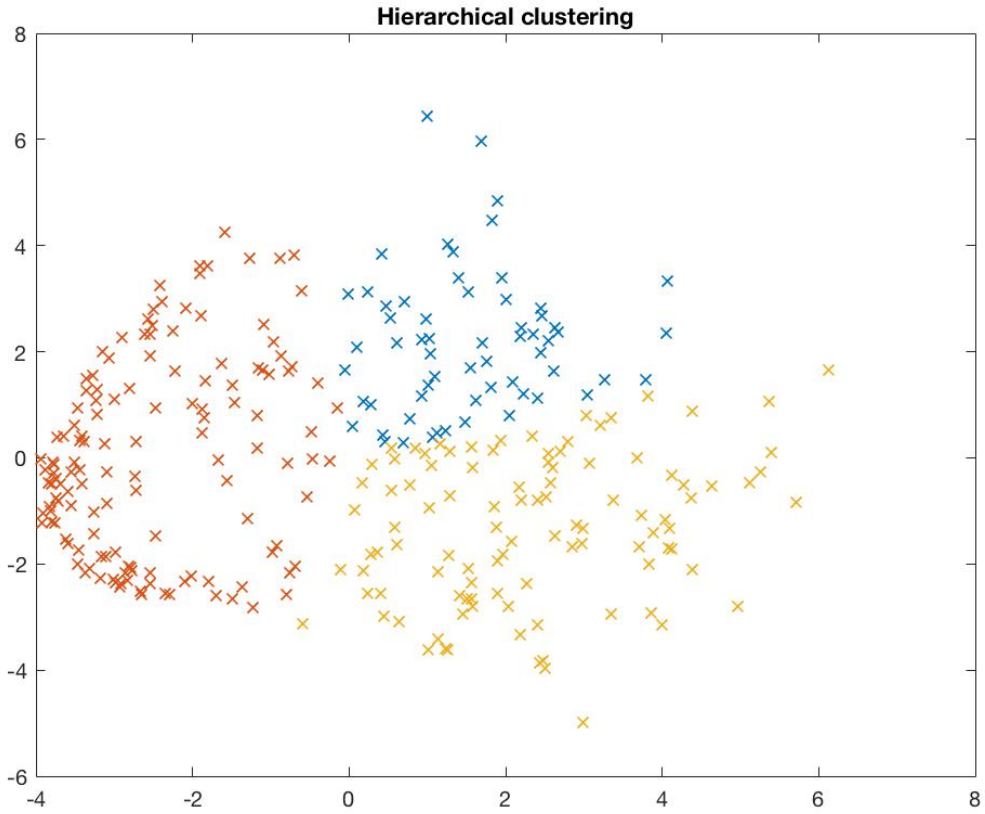
It can be seen, that using PCA to reduce the dimensionality from 100 to 2 dimensions did not perform very well: Eights and Fives have a low degree of separation, Ones, however, produced a better result.

Hierarchical Clustering

Two hierarchical clustering algorithms were used to test the PCA results. In this case, a dendrogram using the Chebychev distance was produced and then the cosine distance method was tested as well, the results can be seen below:



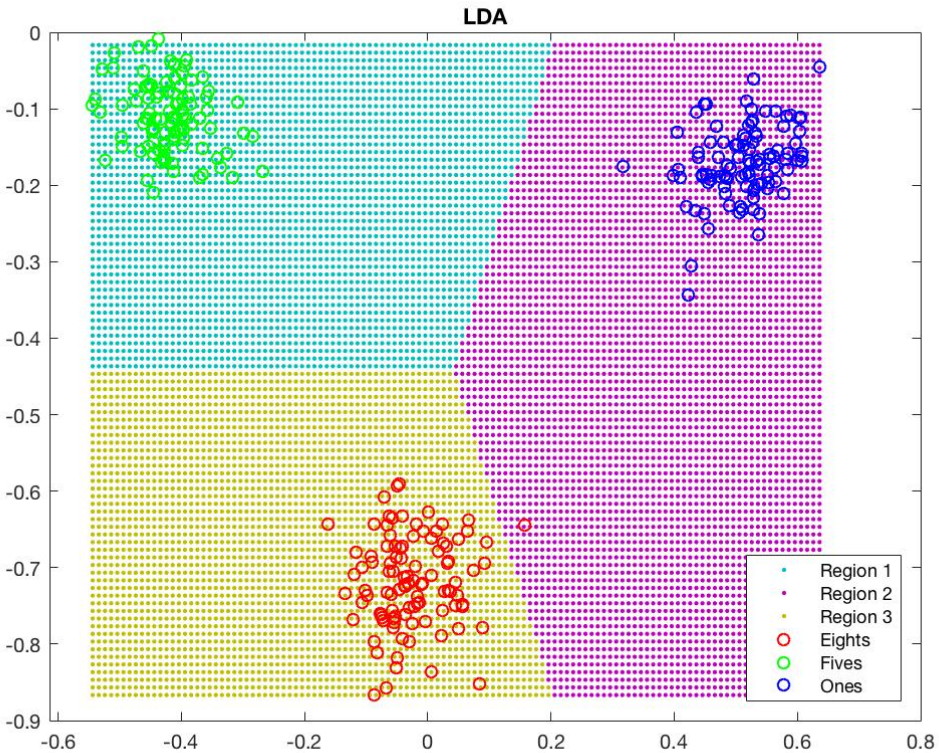




The dendrograms produced above show the hierarchical clustering of the distance matrices and each has a place where the tree can be cut in order to produce 3 clusters (3.8 and 0.6 respectively). The usage of the cosine distance has produced much clearer results.
The plotted results show a much better distinction of the each cluster comparing to K-means result.

2. LDA

LDA was used in an attempt to Improve the results of the dimensionality reduction with k-means clustering.

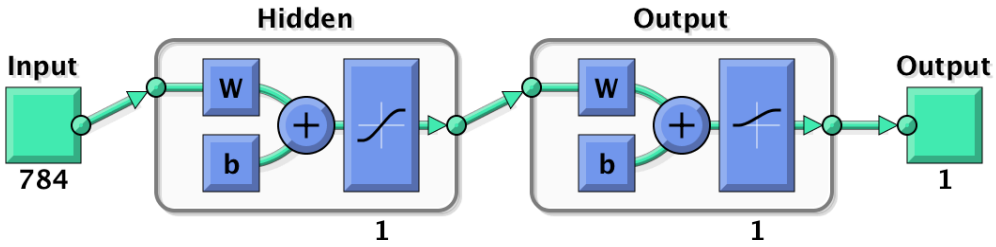


This algorithm shows a much better results comparing with PCA and the data is clearly separated into 3 distinct clusters and in my opinion, LDA is the preferred method for dimensionality reduction in the case of this particular dataset.

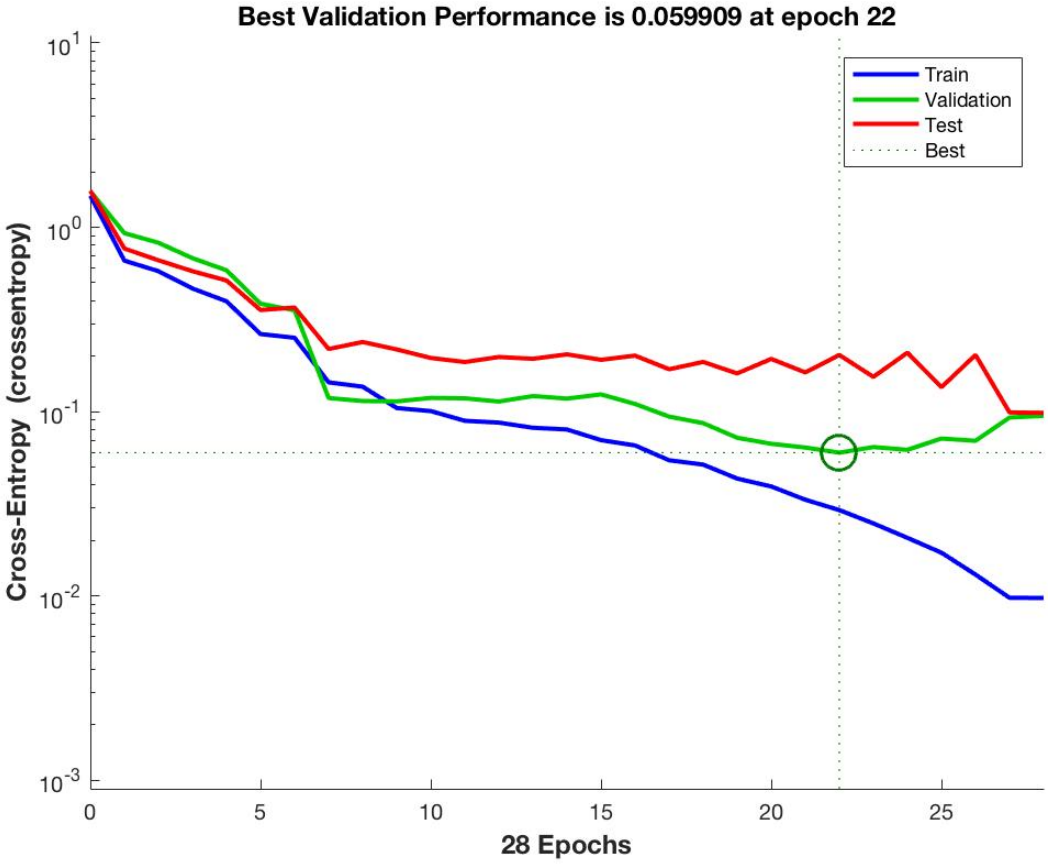
3. Neural Network Classification

Pattern Recognition Network was used in order to perform this task and produced the following results:

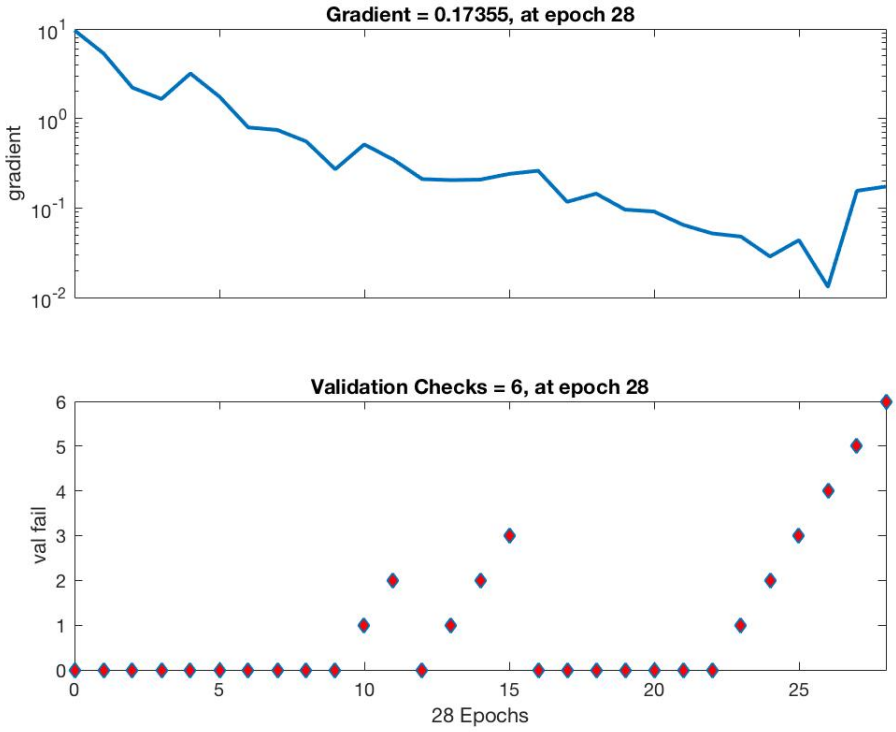
NN Scheme:



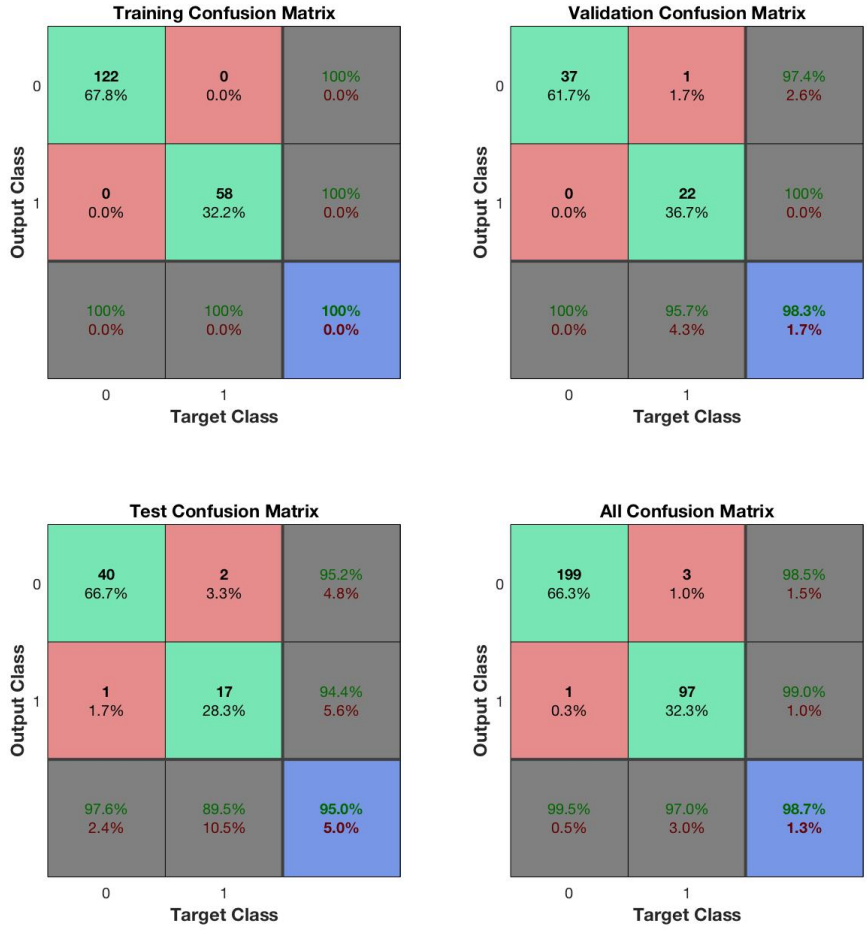
NN training performance:



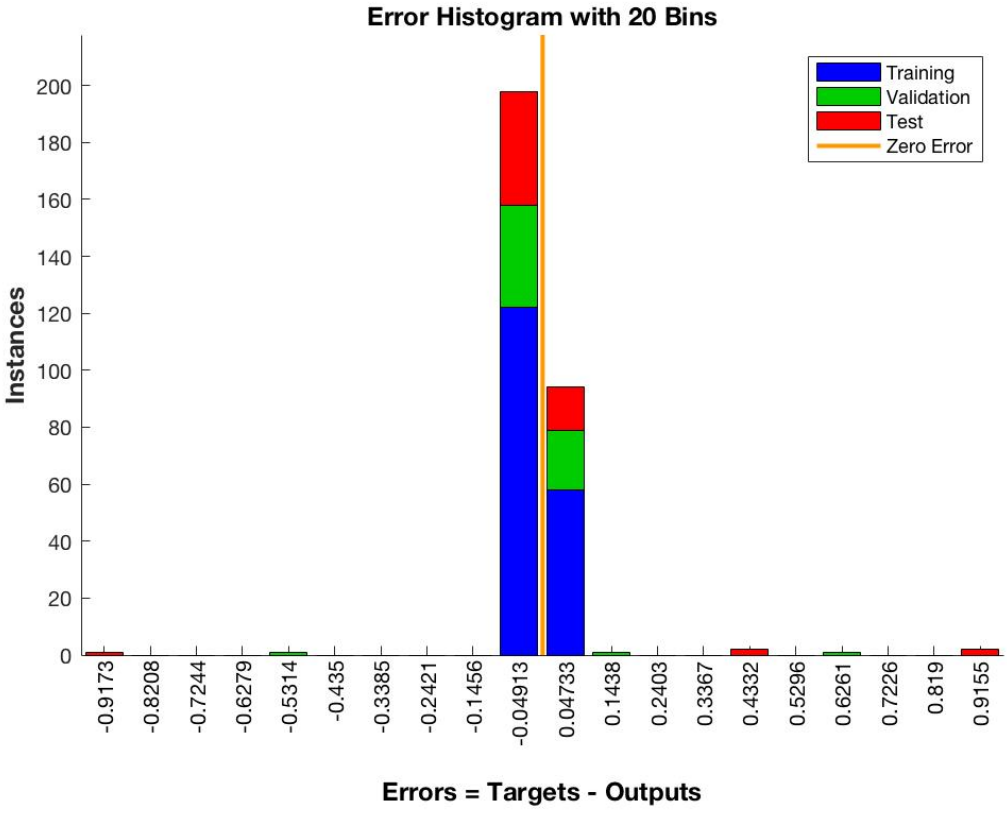
NN training training state:



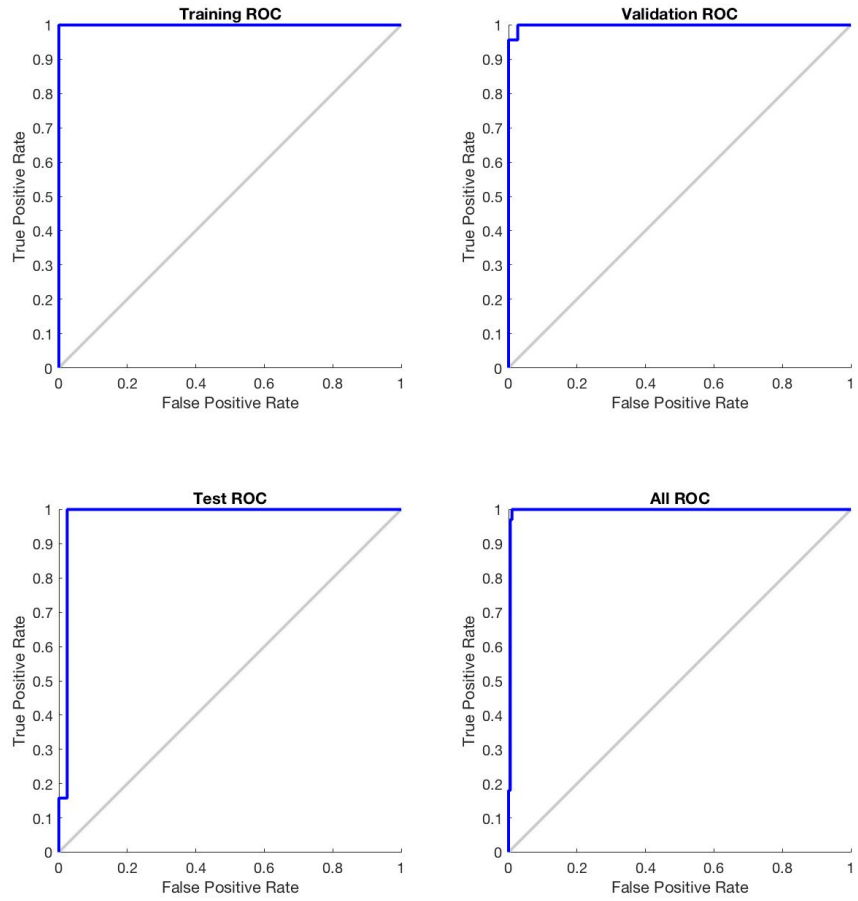
NN training confusion:



NN training error histogram:



NN training Receiver Operating Characteristic:



In order to perform the classification, the dataset was divided and class Fives was separated from the rest. The NN showed an overall good performance with the mean accuracy of 92%.

The ROC curve plotting was used to check the relationship between the TPR and FPR, where the 45 degree grey line represents the hypothetical model's ROC curve (completely random predictions). It can be seen that the blue line, which is our model's curve, sits very far away from the grey line, and that represents a good accuracy result.

SVM

The SVM classification was performed using a Linear kernel and a RBF kernel. Due to not having an access to relevant version of Matlab later into assignment, the accuracy results were checked using Python (SVM MNIST digital classification). The Linear kernel classification produced accuracy of 93% and RBF kernel classification model produced accuracy of 95%.

If we compare the results of Linear, RBF and NN algorithms we can see that all three performed well with a minimal result variation, with RBF kernel algorithm being the most accurate of the three.

Software Instructions.

The zipped folder contains the given dataset, libsvm library which must be compiled through running 'make.m' file. Run each corresponding file -
PCA_JG.m - for PCA, K-means and hierarchical clustering
LDA_JG.m - for LDA dimensionality reduction
NN_JG.m - for Neural Network classification