HANDWRITING COMPARISION USING MACHINE LEARNING

INTRODUCTION

- The objective of the project is to use machine learning techniques to compare handwritten samples.
- Given two samples of handwriting, the system must come to the conclusion whether both samples belong to the same writer or they are sample from a different writer
- The output is in the form of a scalar value. If the output is 1, then both samples are said to be from the same writer.
- If the output is 0, then the values are said to be from different writers.

EXPERIMENT

- The word samples of the word 'AND' have been extracted from the CEDAR dataset.
- There are two main datasets provided for the purpose of classification and identification.
- The human observed dataset where for each ample provided, the human expert has entered numerical values for 9 features corresponding to the sample.
- The set contains 791 same writers and 293,032 different writer pairs.
- The other dataset is the GSC (Gradient Structural Concavity) dataset which provides much more in-depth analysis of each sample providing a set of 512 features for each sample.
- It contains 71,531 same writer pairs and 762,557 different writer pairs.
- For each dataset, the techniques to be applied are Linear regression, Logistic Regression and Neural networks.

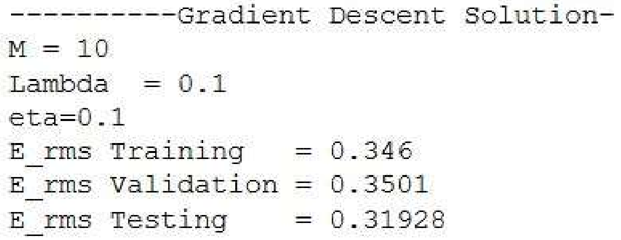
TASK

- For each dataset, following task has been performed:
- The sample writer pair entries from 'same pair' files have been taken and a random sample of the same size from the different writer pairs (which has a lot more entries) has been taken
- The entries have been grouped into one data-frame using PANDAS and the resulting data-frame is shuffled.

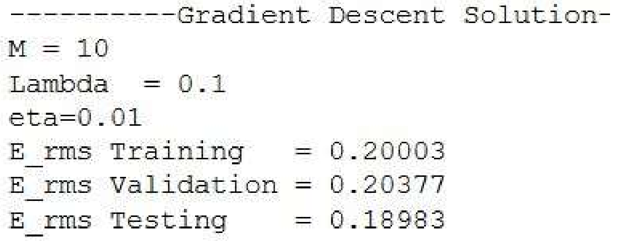
- From this data-frame, two data-frames are created. One file has all the features for each entry while the other data-frame has the respective target values.
- These data-frames are processed into CSV files which are then used in each of the ML technique as the input data and comparison standard
- Number of rows for human observed dataset: 791 + 791 = 1524
- Number of rows for GSC dataset: 2000 (taken as a sample)
- Two sets of features are obtained for each type of operation formed:
- Namely, Concatenation and subtraction. For each operation, the number of features in file are 18 and 9 respectively.
- Naming of files: type of dataset_type of technique perfomed_natureof dataset
- Ex: human_concat_features_set

LINEAR REGRESSION

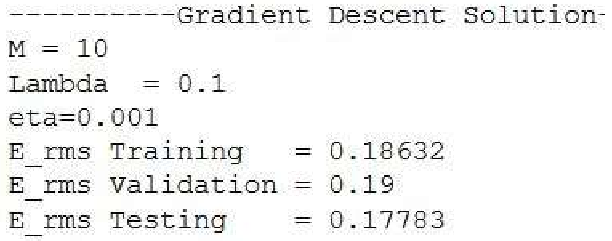
- For linear regression, the E_rms values are as follows:



```
----------Gradient Descent Solution-
M = 10
Lambda  = 0.1
eta=0.1
E_rms Training   = 0.346
E_rms Validation = 0.3501
E_rms Testing    = 0.31928
```



```
----------Gradient Descent Solution-
M = 10
Lambda  = 0.1
eta=0.01
E_rms Training   = 0.20003
E_rms Validation = 0.20377
E_rms Testing    = 0.18983
```

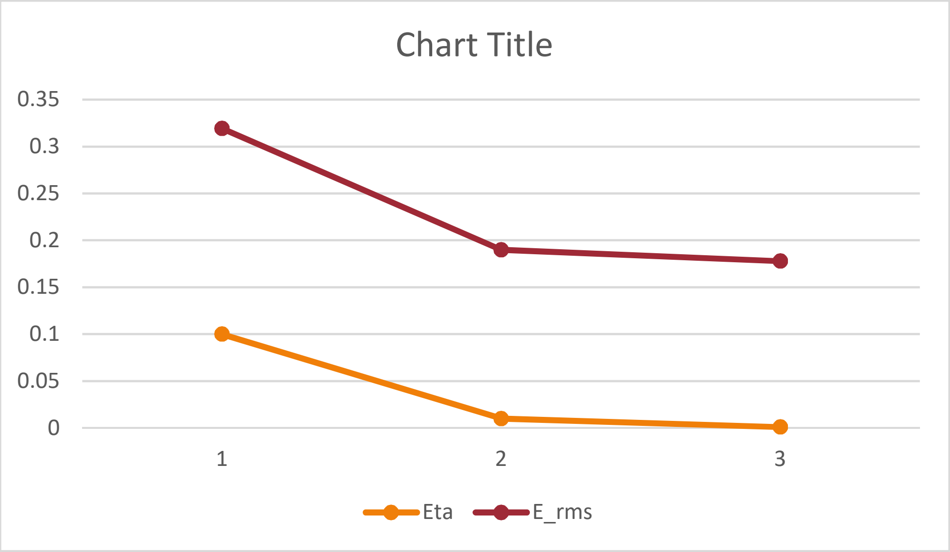


```
----------Gradient Descent Solution-
M = 10
Lambda  = 0.1
eta=0.001
E_rms Training   = 0.18632
E_rms Validation = 0.19
E_rms Testing    = 0.17783
```

OBSERVATION:

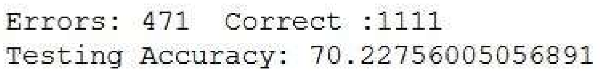
- When the number of clusters and lambda (trade-off) are kept constant and the learning rate (eta) is varied, the performance of the system is impacted significantly.

- Plot of learning rate vs Erms shows that if the learning rate is slightly decreased, then the overall performance is increased in the form of lower Root mean square error

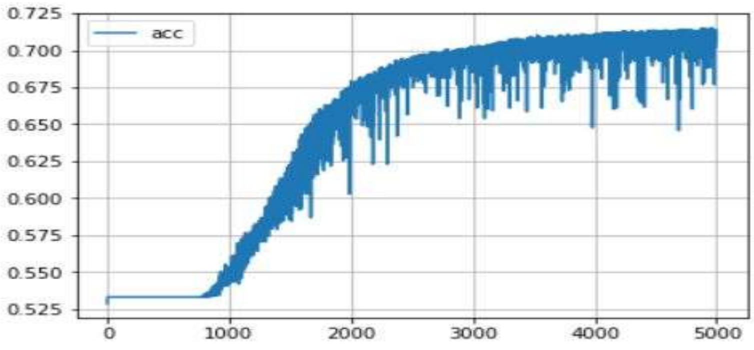


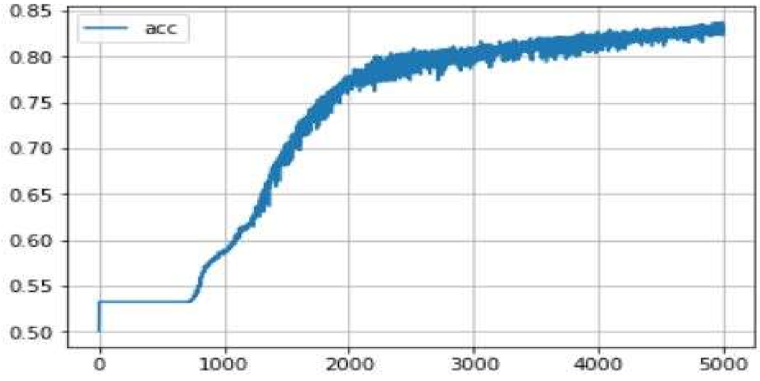
NEURAL NETWORKS

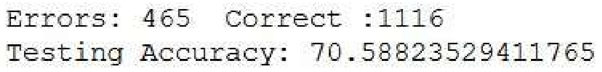
- The neural networks algorithm utilizes a network of layers connected with weights wherein with each iteration, the weights are updated and the efficiency is increased.
- The concatenation and subtraction feature sets are both run through the neural network separately to produce the classification based on various parameters such as number of hidden layers, number of epochs, batch size etc.



```
Errors: 471  Correct :1111
Testing Accuracy: 70.22756005056891
```



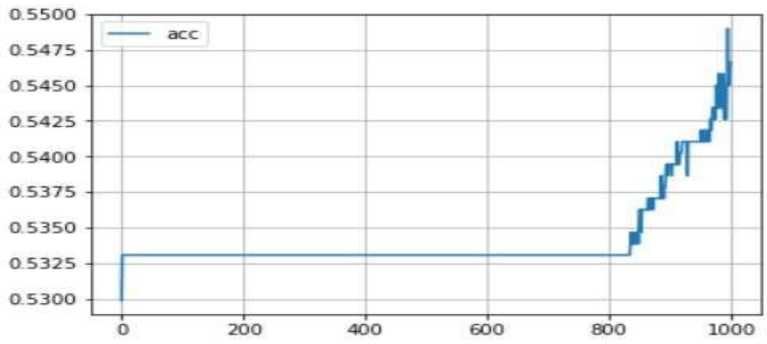


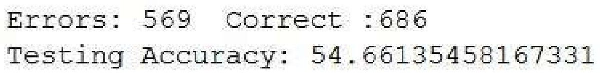


```
Errors: 465  Correct :1116
Testing Accuracy: 70.58823529411765
```

PERFORMANCE ANALYSIS:

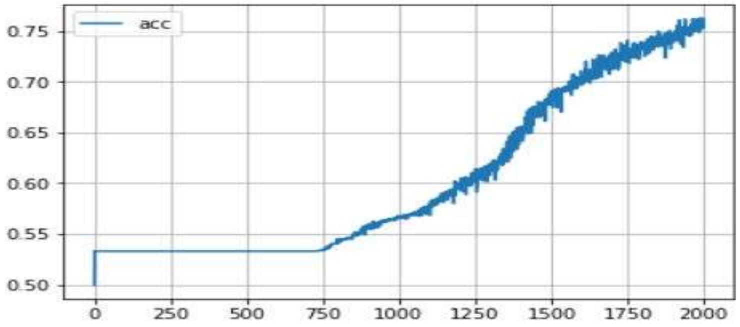
- Number of epochs: 1000

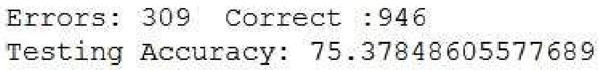




```
Errors: 569  Correct :686
Testing Accuracy: 54.66135458167331
```

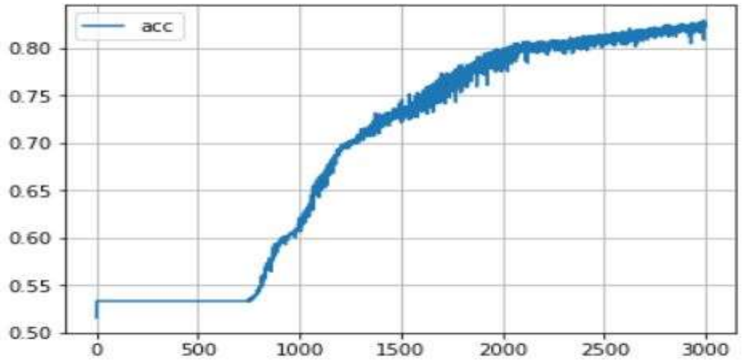
- Number of epochs: 2000

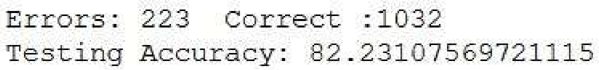




```
Errors: 309  Correct :946
Testing Accuracy: 75.37848605577689
```

- Number of epochs: 3000





```
Errors: 223  Correct :1032
Testing Accuracy: 82.23107569721115
```

CONCLUSION

- We find that on the concatenation dataset, the accuracy is that of around 70%
- This shows that the neural network (at least in certain cases) performs better than other techniques
- As the number of epochs are increased i.e number of iterations for training is increased, the performance of the system increases.