Machine learning to solve the handwriting comparison task in Forensics.

Linear regression, Logistic regression, Neural Network

Machine Learning | Project 2 |1st November, 2018

Monica Vashu Kherajani
Person #50290424
mkheraja@buffalo.edu

Problem Statement

The problem statement is broadly classified into three major tasks:

1. Performing linear regression
2. Performing logistic regression
3. Using neural network for prediction

on two different datasets with two combinations of each dataset, i.e., 4 datasets.

Solution

The major steps involved in the solution are as follows:

1. Pre-process given datasets.
2. Partition them into training, validation and testing sets.
3. Obtain two different combinations of each dataset:
    a. Concatenated feature form
    b. Subtracted feature form

4. Define linear regression, logistic regression model and a neural network for the four dataset combinations obtained in step 3.
5. Train these models.
6. Tune hyper-parameters for the models defined in step 4 using validation set.
7. Test the fixed model in step 6 on testing set.

Conceptual and Technical understanding

Conceptual understanding:

1. Linear regression
    - Used k-means clustering for getting the cluster centers for using Gaussian radial basis function on input data.
    - Used the weights from closed form solution as initial weight vector for Gradient Descent.
    - Used Gradient descent to train the model for about 1000 samples.
    - Calculated root mean square error and accuracy for training, validation and testing

set.

2. Logistic regression
    - For a fixed set of epochs, the weight updates are done using vectorization.

      The genesis equation for Logistic regression:

      $$y = \sigma(W^T X)$$

      The gradient of weights is:

      $$\Delta w = X(A - Y)$$

      Where A = $\sigma(W^T X)$

    - Final accuracy is calculated using the weight vector obtained after the end of last epoch.
    - Also, the accuracy of the designed model is compared with the sklearns logistic model defined already.
3. Neural Network
    - Used tensorflow for neural network implementation.
    - Used one hidden layer with 200 neurons in hidden layer.
    - Accuracy is calculated for evaluation of the designed model.

Technical understanding of the three models in demonstrated in the code submitted along with appropriate and sufficient comments.

Experiments and Results

Part I - Human observed features Dataset

1. Linear Regression with feature concatenation

Using unshuffled Data set

Linear Regression for Input appended features
E_rms Training = 0.6099
E_rms Validation = 0.70453
E_rms Testing = 0.70904
Training accuracy = 50.0
Validation accuracy = 50.0
Testing accuracy = 49.358974358974365

Using shuffled Data set

Linear Regression for Input appended features SHUFFLED DATA SET
E_rms Training = 0.5922
E_rms Validation = 0.60421
E_rms Testing = 0.56572
Training accuracy = 49.5260663507109
Validation accuracy = 49.36708860759494
Testing accuracy = 55.12820512820513

2. Linear Regression with feature subtraction

Using unshuffled Data set

Linear Regression for Input subtracted features
E_rms Training = 0.52303
E_rms Validation = 0.56265
E_rms Testing = 0.6117
Training accuracy = 50.0
Validation accuracy = 50.0
Testing accuracy = 49.358974358974365

Using shuffled Data set

Linear Regression for Input subtracted features SHUFFLED DATA SET
E_rms Training = 0.53636
E_rms Validation = 0.50263
E_rms Testing = 0.53325
Training accuracy = 50.39494470774092
Validation accuracy = 46.835443037974684
Testing accuracy = 50.0

3. Logistic Regression with feature concatenation

Logistic Regression(Own Implementation)
Testing Accuracy = 57.32484076

Logistic Regression(SKLEARN) for Input appended features
E_rms Training = 0.27823
E_rms Validation = 0.13779
E_rms Testing = 0.11323
Training Accuracy = 92.25908372827804
Validation Accuracy = 98.10126582278481
Testing Accuracy = 98.71794871794873

4. Logistic Regression with feature subtraction

Logistic Regression(Own Implementation)
Testing Accuracy = 48.40764331

Logistic Regression(SKLEARN) for Input subtracted features
E_rms Training = 0.45665
E_rms Validation = 0.29767
E_rms Testing = 0.38397
Training Accuracy = 79.14691943127961
Validation Accuracy = 91.13924050632912
Testing Accuracy = 85.25641025641025

5. Neural Network with feature concatenation

Parameters

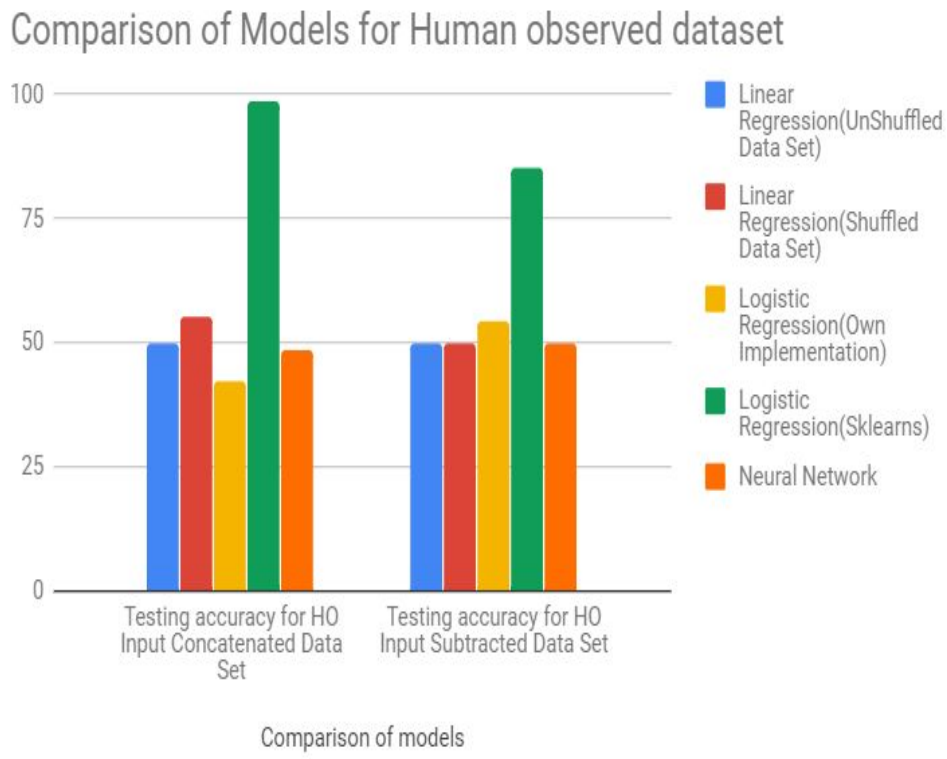
Number of hidden layers = 1
Number of Neurons in hidden layer = 200
NUM_OF_EPOCHS = 1500
BATCH_SIZE = 128
LEARNING_RATE = 0.05

Neural Networks for HOF input concatenated features SHUFFLED DATA SET
E_rms Testing = 0.71827819602086
Testing Accuracy = 48.40764331210191

6. Neural Network with feature subtraction

Neural Networks for HOF input subtracted features SHUFFLED DATA SET
E_rms Testing = 0.7093551391057911
Testing Accuracy = 49.681528662420384

Comparison between the models for both datasets:



Part II - Gradient Structural Concavity(GSC) Dataset

7. Linear Regression with feature concatenation

Linear Regression for Input appended features SHUFFLED DATA SET
E_rms Training = 0.70652
E_rms Validation = 0.70552
E_rms Testing = 0.7078
Training accuracy = 50.082713949532845
Validation accuracy = 50.22368237103313
Testing accuracy = 49.90213896267301

8. Linear Regression with feature subtraction

Linear Regression for Input subtracted features SHUFFLED DATA SET
E_rms Training = 0.70754
E_rms Validation = 0.70562
E_rms Testing = 0.70696
Training accuracy = 49.93825578415154
Validation accuracy = 50.20970222284357
Testing accuracy = 50.020970222284355

9. Logistic Regression with feature concatenation

Using shuffled Data set

Logistic Regression(SKLEARN) for Input Appended features
E_rms Training = 0.09056
E_rms Validation = 0.7113
E_rms Testing = 0.71301
Training Accuracy = 99.17986952469711
Validation Accuracy = 49.40584370194324
Testing Accuracy = 49.16119110862575

10. Logistic Regression with feature subtraction

Using shuffled Data set

Logistic Regression(SKLEARN) for Input Subtracted features

E_rms Training = 0.31751
E_rms Validation = 0.5111
E_rms Testing = 0.51157
Training Accuracy = 89.91845293569432
Validation Accuracy = 73.87809310778695
Testing Accuracy = 73.82916258912344

11. Neural Network with feature concatenation

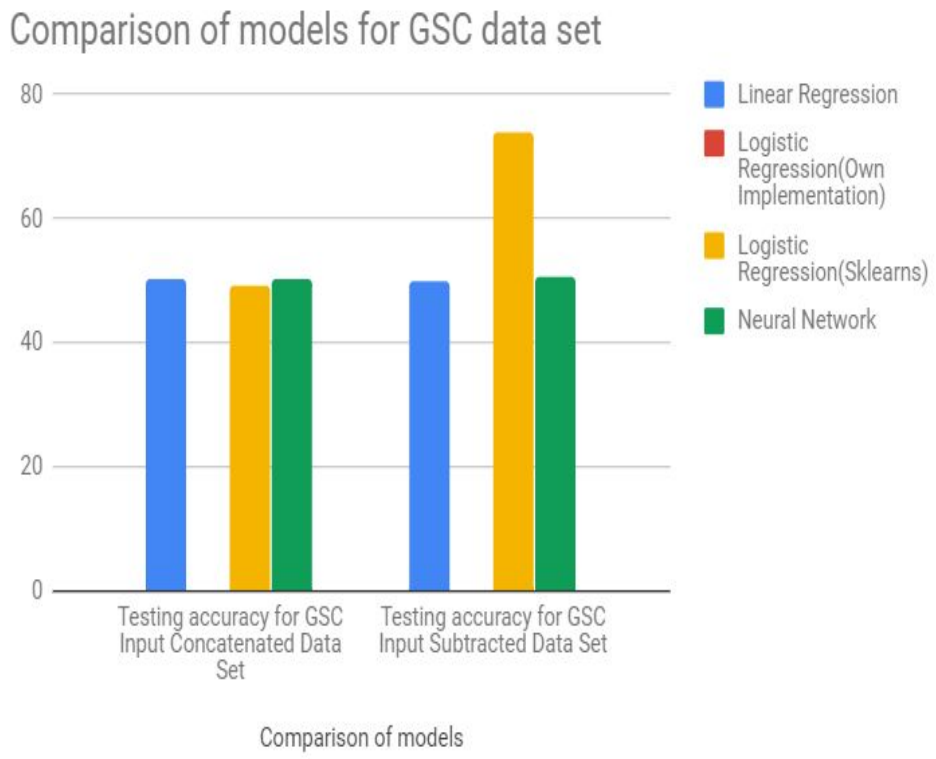
Parameters
Number of hidden layers = 1
Number of Neurons in hidden layer = 200
NUM_OF_EPOCHS = 50
BATCH_SIZE = 32
LEARNING_RATE = 0.05

Neural Networks for GSC input concatenated features SHUFFLED DATA SET
E_rms Testing = 0.7060185434527532
Testing Accuracy = 50.153781630085284

12. Neural Network with feature subtraction

Neural Networks for GSC input subtracted features SHUFFLED DATA SET
E_rms Testing = 0.7025447543272937
Testing Accuracy = 50.64308681672026

Comparison between the models for both datasets:



Inferences and Conclusions

1. Linear Regression isn't giving good results for the given data sets. Only about 50% accurate.
2. Logistic regression and Neural network are performing comparatively better than Linear regression for these particular datasets.
3. However, unable to process the entire GSC data set due to low processing power. Only used about 30% as training after taking equal number of samples for both the output classes.
4. There is not much difference in the accuracies given by both the datasets for the proposed three models, i.e, both data sets perform almost similar.
5. Logistic regression proposed model is not working for GSC Data set.

REFERENCES

1. Project 1.1 code (Neural Network)
2. Project 1.2 code (Linear regression)