Introduction to Machine Learning

Project 2 Report

UBID: sppandey@buffalo.edu

PERSON NO: 50288608

Overview :

The goal of this project is to apply machine learning to solve the handwriting comparison task in forensics. We formulate this as a problem of linear regression where we map a set of input features x to a real-valued scalar target y(x,w).

Data to be worked with :

Each instance in the CEDAR “AND” training data consists of set of input features for each handwritten “AND” sample. The features are obtained from two different sources:

1. Human Observed features: Features entered by human document examiners manually
2. GSC features: Features extracted using Gradient Structural Concavity (GSC) algorithm.

Project Content:

The Report highlights the process of concatenating extracted features, subtracting them and then training them using linear and Logistic Regression. A fair comparison of mean squared values, loss values has been presented in the report obtained over training, validating and testing data over the said models. Hyperparameter such as learning rate, epochs, number of basis functions have been tuned to obtain optimal output and least Erms.

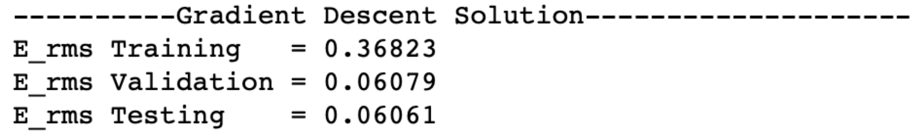
Scenarios:

a) **Human Observed Dataset with feature concatenation** :

Linear Regression :

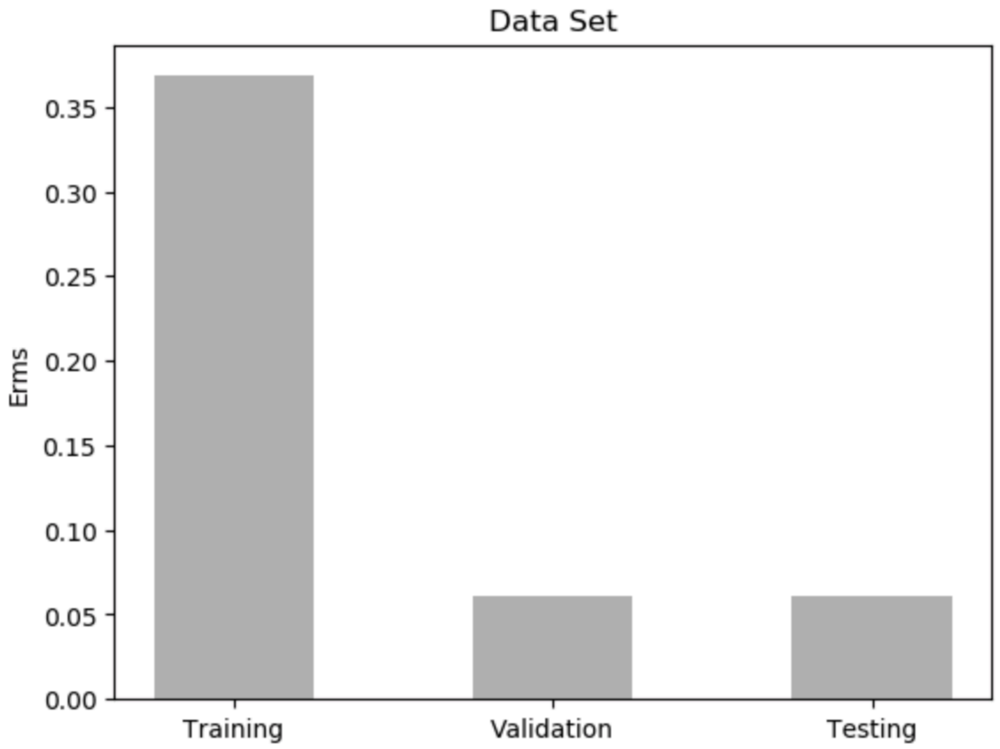
Number of Features for each writer : 9
Hyperparameters :
Learning Rate : 0.02
Number of Epochs : 800

Erms Values :



```
----------Gradient Descent Solution--------------------
E_rms Training   = 0.36823
E_rms Validation = 0.06079
E_rms Testing    = 0.06061
```

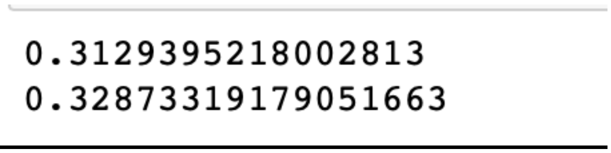
Graphical Representation of Erms over the different data sets for Human Observed Data :



Logistic Regression:

Learning rate: 0.00001
Number of Iterations : 100000

Prediction values:



```
0.3129395218002813
0.32873319179051663
```

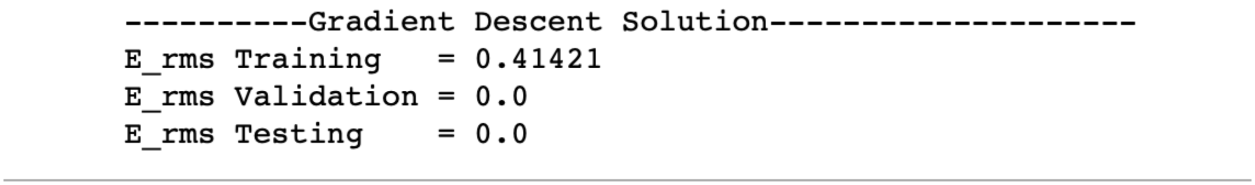
In the given setting of regression model an increase in number of iterations at a constant learning rate caused the model to give out lower prediction values.

b) Human Observed Dataset with feature Subtraction :

Linear Regression :

Number of Features for each writer : 9
Hyperparameters :
Learning Rate : 0.01
Number of Epochs : 900

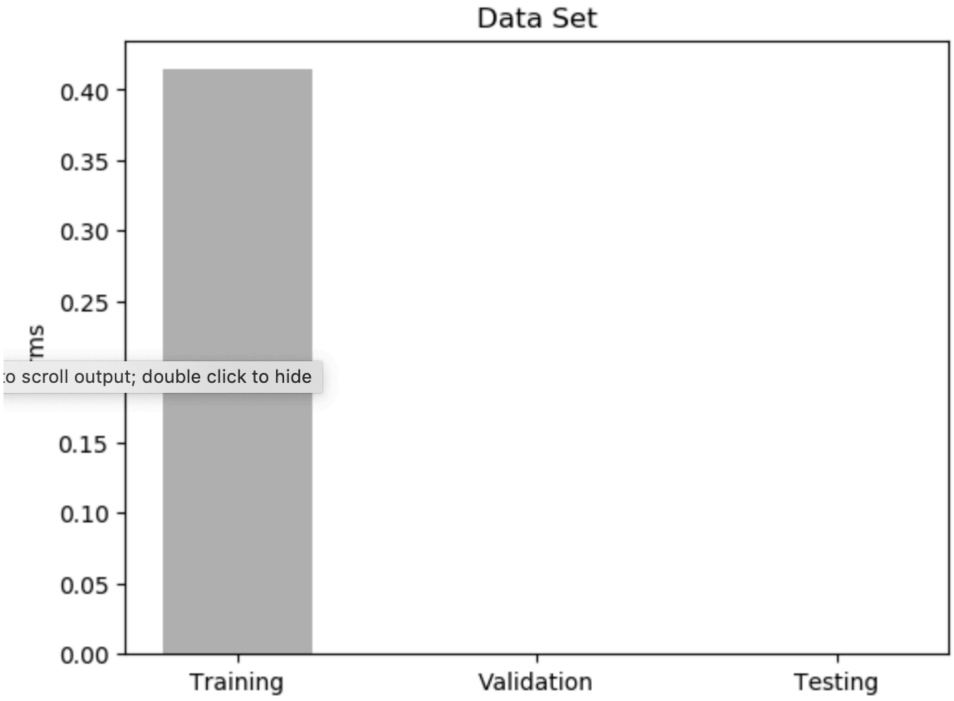
Erms Values :



```
----------Gradient Descent Solution--------------------
E_rms Training   = 0.41421
E_rms Validation = 0.0
E_rms Testing    = 0.0
```

For the given setting Erms reached and ideal value of Zero.

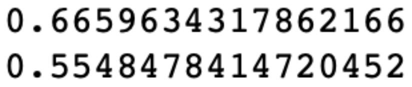
Graphical Representation of Erms over the different data sets for Human Observed Data :



Logistic Regression:

Learning rate: 0.007/0.7/0.8
Number of Iterations : 100000

Prediction values: Keeping the iterations constant an improved prediction was observed for values 0.07 and further increasing the values to 0.7 and above gave a constant prediction value as below.



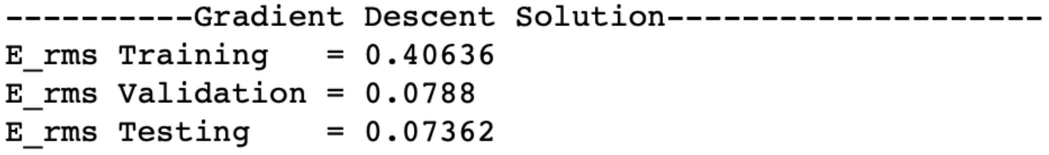
```
0.6659634317862166
0.5548478414720452
```

c) **GSC dataset with feature Concatenation**:

Linear Regression :

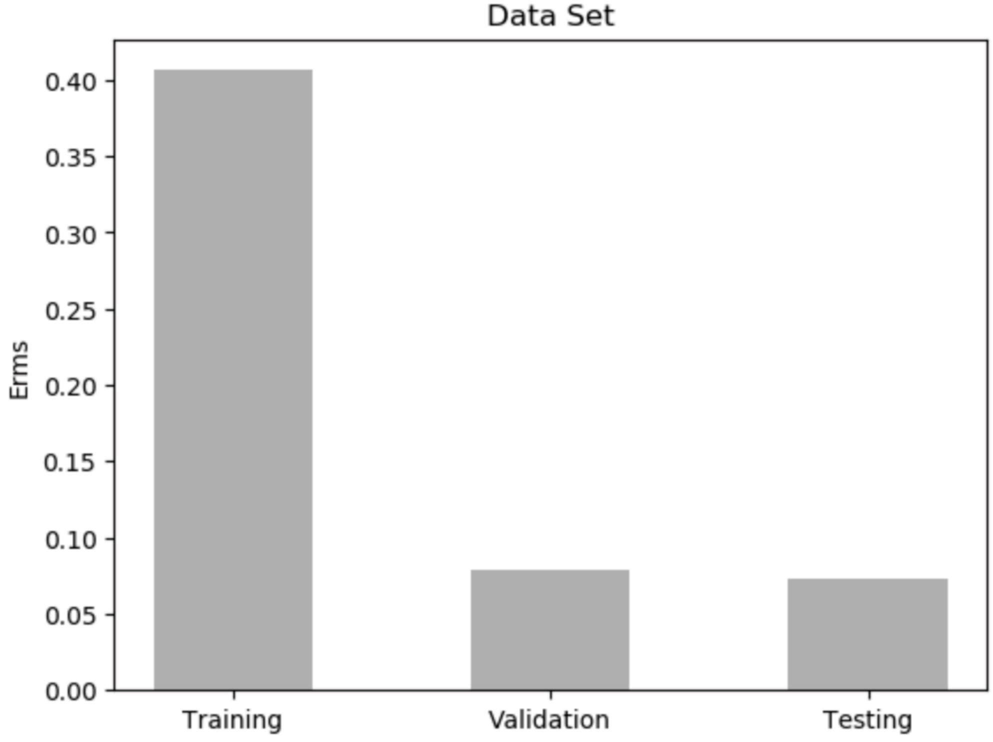
Number of Features for each writer : 512
Hyperparameters :
Learning Rate : 0.1
Number of Epochs : 800

Erms Values :



```
----------Gradient Descent Solution--------------------
E_rms Training   = 0.40636
E_rms Validation = 0.0788
E_rms Testing    = 0.07362
```

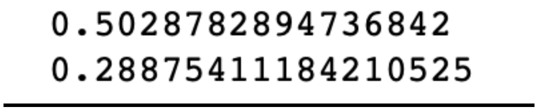
Graphical Representation of Erms over the different data sets for Human Observed Data :



Logistic Regression:

Learning rate: 0.01/0.08
Number of Iterations : 10000/100000

Prediction values: Trying different values for learning rate and iterations showed a very slow increase in the prediction values given the sheer volume of data and learning curve. Increased iterations helped in bumping up the values by 0.3 to 0.5 but kept overall growth slower.



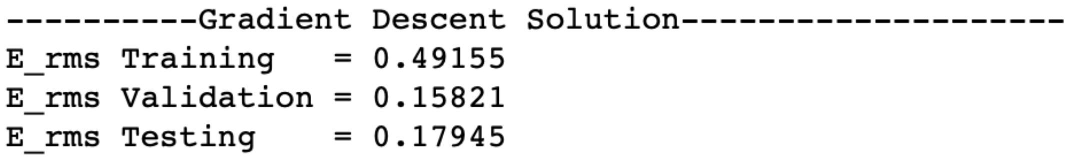
```
0.5028782894736842
0.28875411184210525
```

d) **GSC dataset with feature Subtraction**:

Linear Regression :

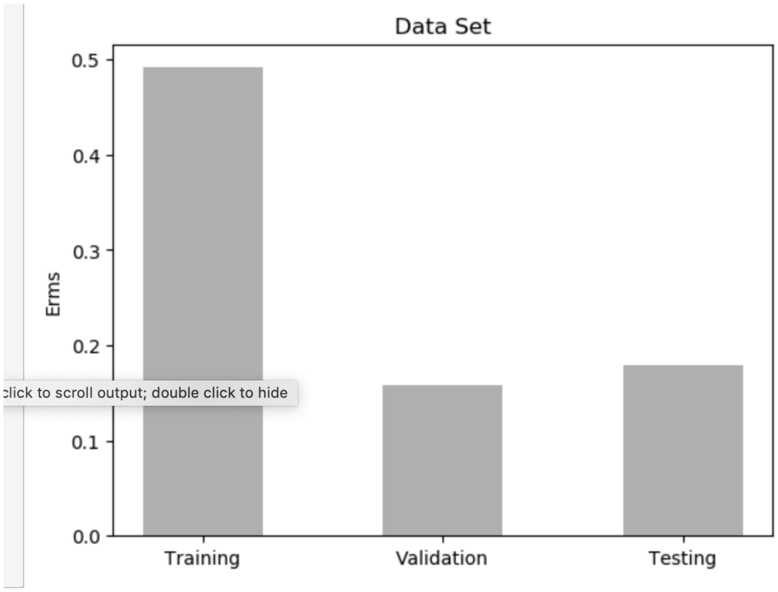
Number of Features for each writer : 512
Hyperparameters :
Learning Rate : 0.01
Number of Epochs : 810

Erms Values :



```
----------Gradient Descent Solution--------------------
E_rms Training   = 0.49155
E_rms Validation = 0.15821
E_rms Testing    = 0.17945
```

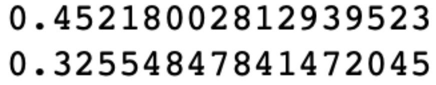
Graphical Representation of Erms over the different data sets for Human Observed Data :



<u>Logistic Regression:</u>

Learning rate: 0.01/0.06
Number of Iterations : 10000/100000

Prediction values: A close to stagnant growth rate was seen for the GSC subtraction data even with some substantial change in the values.



```
0.45218002812939523
0.32554847841472045
```