Report On : Project – 2.0 Handwriting Analysis & Comparison Using Linear Regression and Logistic Regression

Neeraj Ajit Abhyankar **UB Person No. : 50290958** **UBID : nabyank**

Objective — The main objective of this project is to find similarity between the handwritten samples of the known and the questioned writer by using linear regression and logistic regression. For this purpose we are given the instances of word 'AND' written by various writers. The Data consists of Human Observed Data and GSC Data (Computer Observed Data) which has to be processed for effective output and then find the similarities using the regression techniques.

I. Linear Regression

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

II. Logistic Regression

Binary Logistic Regression is a special type of regression where ***binary response variable*** is related to a ***set of explanatory variables***, which can be discrete and/or continuous. The important point here to note is that in linear regression, the expected values of the response variable are modeled based on combination of values taken by the predictors. In logistic regression ***Probability*** or ***Odds*** of the response taking a particular value is modeled based on combination of values taken by the predictors.

III. Human Observed Data

The Human Observed dataset shows only the cursive samples in the data set, where for each image the features are entered by the human document examiner. There are total of 18 features for a pair of handwritten \AND" sample (9 features for each sample).

Question : What is in the Human Observed Data?
Answer : Feature Description for Handwriting of various individual writers

Initial stroke of formation of a (x_1)	Formation of staff of a (x_2)	Number of arches of n(x_3)	Shape of arches of n (x_4)	Location of mid-point of n(x_5)	Formation of staff of d (x_6)	Formation of initial stroke of d (x_7)	Formation of terminal stroke of d (x_8)	Symbol in place of the word *and* (x_9)
Right of staff (0)	Tented (0)	One (0)	Pointed (0)	Above baseline (0)	Tented (0)	Overhand (0)	Curved up (0)	Formation (0)
Left of staff (1)	Retraced (1)	Two (1)	Rounded (1)	Below baseline (1)	Retraced (1)	Underhand (1)	Straight across (1)	Symbol (1)
Center of staff (2)	Looped (2)	No fixed pattern (2)	Retraced (2)	At baseline (2)	Looped (2)	Straight across (2)	Curved down (2)	None (2)
No fixed pattern (3)	No staff (3)		Combination (3)	No fixed pattern (3)	No fixed pattern (3)	No fixed pattern (3)	No obvious ending stroke (3)	
	No fixed pattern (4)		No fixed pattern (4)				No fixed pattern (4)	

Table 1a) Description of Features in Human Observed Data

Question : What does the data look like?

Answer : Example of the given Human Observed Dataset given as below

img_id_A	img_id_B	f_{A1}	f_{A2}	f_{A3}	f_{A4}	f_{A5}	f_{A6}	f_{A7}	f_{A8}	f_{A9}	f_{B1}	f_{B2}	f_{B3}	f_{B4}	f_{B5}	f_{B6}	f_{B7}	f_{B8}	f_{B9}	t
1121a_num1	1121b_num2	2	1	1	3	2	2	0	1	2	2	1	1	0	2	2	0	3	2	1
1121a_num1	1386b_num1	2	1	1	3	2	2	0	1	2	3	1	1	0	2	2	0	1	2	0

Table 1b) Description of Data in Human Observed Data csv file

IV. GSC Data

In the GSC data :
1) The Gradient and Structural features encode local structure while the concavity features are global descriptors extracted from binarized images. A gradient map is constructed from the normalized digit image.
2) The Structural features are computed from the gradient map by examining the similarities in gradient direction in local neighborhood of each pixel.
3) The Concavity features are coarse global descriptors. They are of three kinds :
a) pixel density, b) large strokes and c) true concavity.

Question : What is in the GSC Data?
Answer : GSC algorithm generates 512 sized feature vector for an input handwritten image. GSC algorithm extracts 192 binary gradient features, 192 binary structural features and 128 concavity features. There are total of 1024 features for a pair of handwritten sample (512 features for each sample).

Question : What does the data look like?
Answer : Example of the given Human Observed Dataset given as below

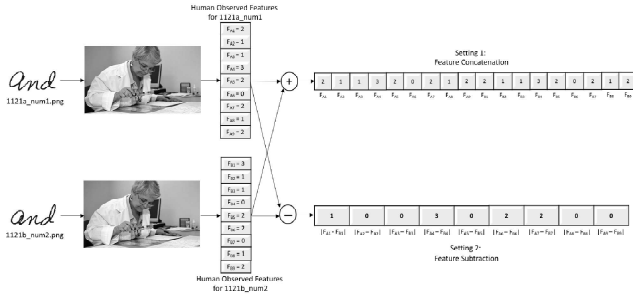
img_id_A	img_id_B	f_{A1}	f_{A2}	f_{A3}	f_{A4}	f_{A5}	f_{A6}	...	f_{A512}	f_{B1}	f_{B2}	f_{B3}	f_{B4}	f_{B5}	f_{B6}	...	f_{B512}	t
1121a_num1	1121b_num2	0	1	1	0	1	0	...	0	0	1	1	0	0	1	...	1	1
1121a_num1	1386b_num1	0	1	1	0	1	0	...	0	1	1	1	0	1	0	...	0	0

Table 2a) Description of Features in Human Observed Data

V. Feature Extraction for Human Data

The description of each of the Human Observed features are given in Table 1a) where there are total of 18 features for a pair of handwritten "AND" sample (9 features for each sample). The dataset is named as "HumanObservedDataset.csv". The entire dataset consists of 512,345 handwritten sample pairs (rows), each having 2 **image_id's**, **18 features** and a **target** value. Table 1b) shows

two sample rows from human observed dataset and the below Figure a) shows the process of extraction of features from the Human Observed Data Set in two settings for Regression.
The Settings are as Follows
i) Setting 1: Feature Concatenation [18 features],
ii) Setting 2: Feature subtraction [9 features]

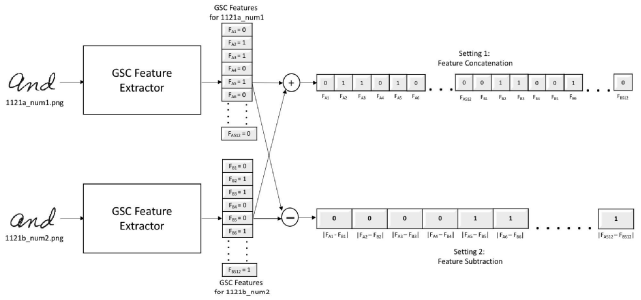


.Figure a) Human Observed Data FeaturesExtraction

VI. FEATURE EXTRACTION FOR HUMAN DATA

GSC algorithm generates 512 sized feature vectors for an input handwritten "AND" image. The entire dataset consists of 512,345 handwritten sample pairs (rows), each having 2 **image id's**, **1024 features** and a **target** value. Table 2b) shows two sample rows from GSC dataset and the below Figure b) shows the process of extraction of features from the GSC Data Set in two settings for Regression.

.



VII. KEY CONCEPTS

A. Training Dataset

The model is initially fit on a training dataset, that is a set of examples used to fit the parameters e.g. weights of connections between neurons in artificial neural networks of the mode. In practice, the training dataset often consist of pairs of an input vector and the corresponding answer vector or scalar, which is commonly denoted as the target. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation.

B. Testing Dataset

The testing dataset is used to provide an unbiased evaluation of a final model fit on the training dataset. A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. A test set is therefore a set of examples used only to assess the performance (i.e. generalization) of a fully specified classifier.

C. Validation Dataset

The Validation Dataset is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper-parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

D. .csv File

CSV is a simple file format used to store tabular data, such as a spreadsheet or database. Files in the CSV format can be imported to and exported from programs that store data in tables, such as Microsoft Excel or Open-Office Calc. CSV stands for "comma-separated values". Its data fields are most often separated, or delimited, by a comma. For example, let's say you had a spreadsheet containing the following data.

VIII. KEY IMPORTS

1) **from sklearn.cluster import KMeans** - import clustering from sklearn.cluster using KMeans.
2) **from sklearn.utils import shuffle** - import shuffle utility from sklearn utilities.
3) **import numpy as np** - import numpy library as np.
4) **import csv** - import csv import and export format for spreadsheets and databases.
5) **import math** - import math library.
6) **import matplotlib.pyplot** - import pythons math plot library.
7) **from matplotlib import pyplot as plt** - import plot from math plot library.
8) **import pandas as pd** - imports pandas methods as pd since python built-in methods can overlap panda methods.

IX. ALGORITHM FOR HUMAN OBSERVED DATA BASED REGRESSION

A. For Human Observed Data :
1) Initialize/ Import the required libraries.
2) Read the given .csv files and get the required Human Observed Dataset
3) Get equal number of ***same pairs*** *and* ***different pairs*** *from the Dataset*
4) Pre-process the Human Observed Dataset with the settings for
i) Feature Concatenation and ii) Feature Subtraction
5) Write the pre-processed data to new.csv files.
6) Initialize the Parameters required for the calculation of Linear Regression & Logistic Regression. For Logistic regression Calculate the Sigmoid Function for regression
7) Calculate the required values for regression using various library functions such as Target Vector, Raw Data

Generation, Training Target, Training Matrix, Validation Data, Validation Target Vector.
8) Calculate the Big-Sigma, Phi Matrix, Radial Basis Function, Required Weights for Closed Form Solution and ERMS values for Training, Testing and Validation Datasets.
9) Get the new written .csv files from their location.
10) Split and Prepare the Training, Testing and Validation Datasets.
11) Calculate the Closed Form Solution for the above parameters.
12) Calculate the Gradient Descent Solution for above parameters. For Logistic Regression Pass the data through Sigmoid Function.
13) Plot the respective graphs for Training, Testing and Validation Dataset Accuracy and ERMS values.

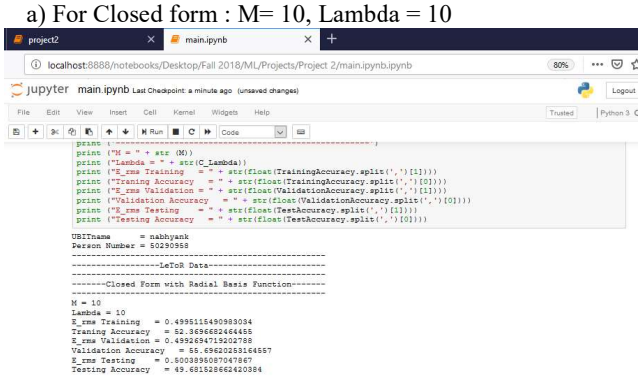
B. For GSC Observed Data :
1) Initialize/ Import the required libraries.
2) Read the given .csv files and get the required GSC Observed Dataset
*3) Get equal number of **same pairs** and **different pairs** from the Dataset*
*4) Pre-process the GSC Observed Dataset with the settings for **i) Feature Concatenation and ii) Feature Subtraction***
5) Write the pre-processed data to new.csv files.
6) Initialize the Parameters required for the calculation of Linear Regression & Logistic Regression. For Logistic regression Calculate the Sigmoid Function for regression.
7) Calculate the required values for regression using various library functions such as Target Vector, Raw Data Generation, Training Target, Training Matrix, Validation Data, Validation Target Vector.
8) Calculate the Big-Sigma and add some noise say 0.00001 to it while calculating the BigSigma Inverse, Phi Matrix, Radial Basis Function, Required Weights for Closed Form Solution and ERMS values for Training, Testing and Validation Datasets.
9) Get the new written .csv files from their location.
10) Split and Prepare the Training, Testing and Validation Datasets.
11) Calculate the Closed Form Solution for the above parameters.
12) Calculate the Gradient Descent Solution for above parameters. For Logistic Regression Pass the data through Sigmoid Function.
13) Plot the respective graphs for Training, Testing and Validation Dataset Accuracy and ERMS values.

X. CHANGES IN HYPER_PARAMETERS WITH PLOTS AND GRAPHS

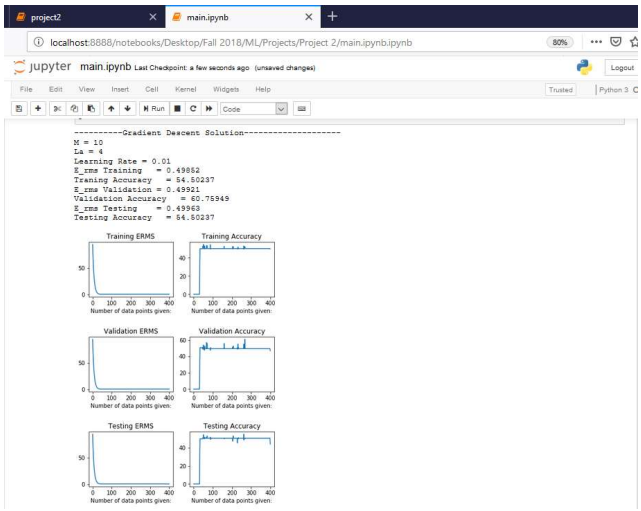
1. For Human Observed Data :
A1. For Linear Regression using Feature Concatenation Setting :

i) Parameter Set 1 :

a) For Closed form : M= 10, Lambda = 10

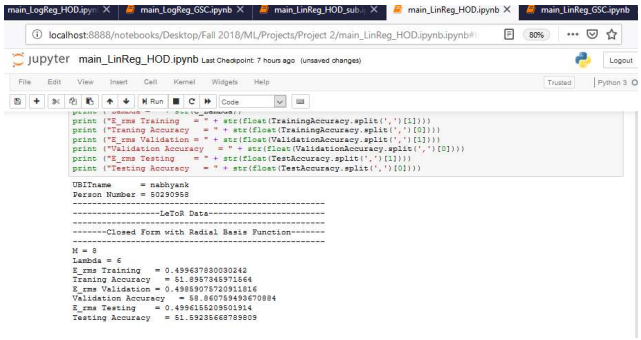


b) For Gradient Descent : M= 10, La= 4, Learning Rate = 0.01

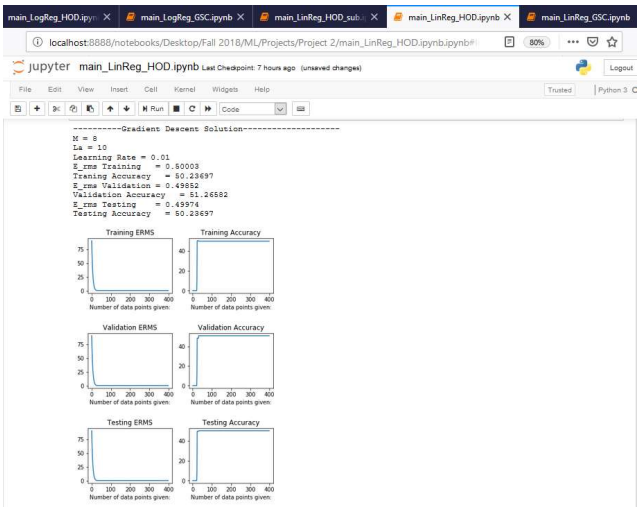


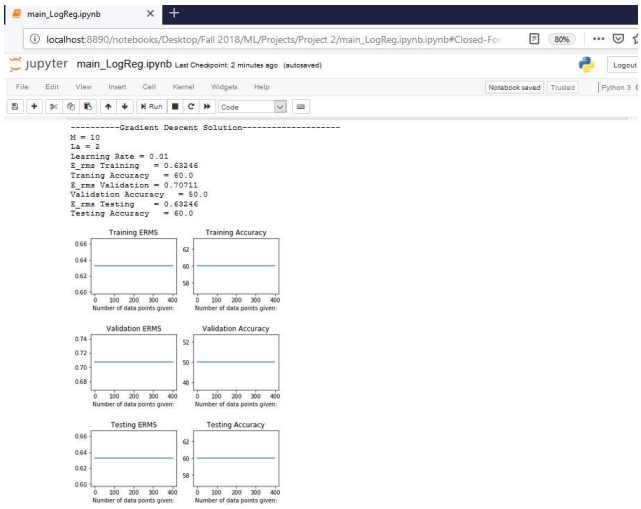
ii) Parameter Set 2 :

a) For Closed form : M= 8, Lambda = 6



b) For Gradient Descent : M= 8, La= 10, Learning Rate = 0.01

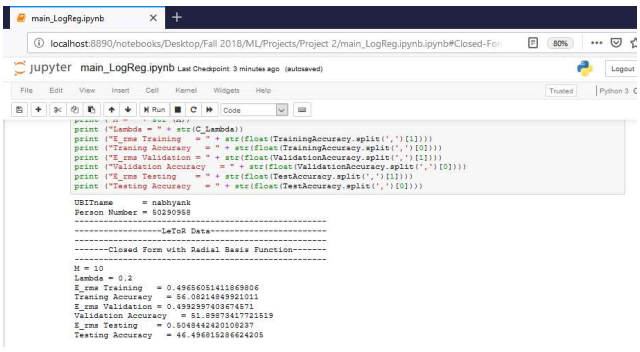




B1. For Logistic Regression using Feature Concatenation Setting :

i) Parameter Set 1 :

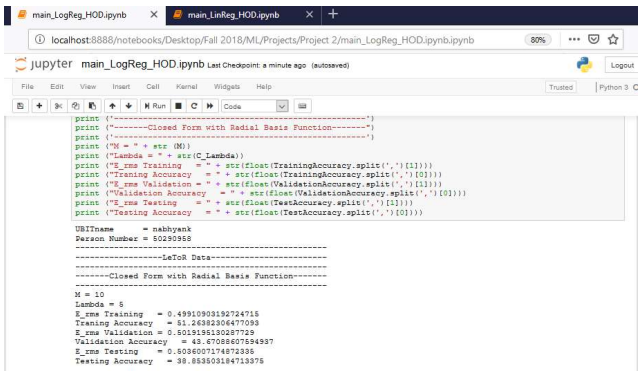
a) For Closed form : M= 10, Lambda = 0.2



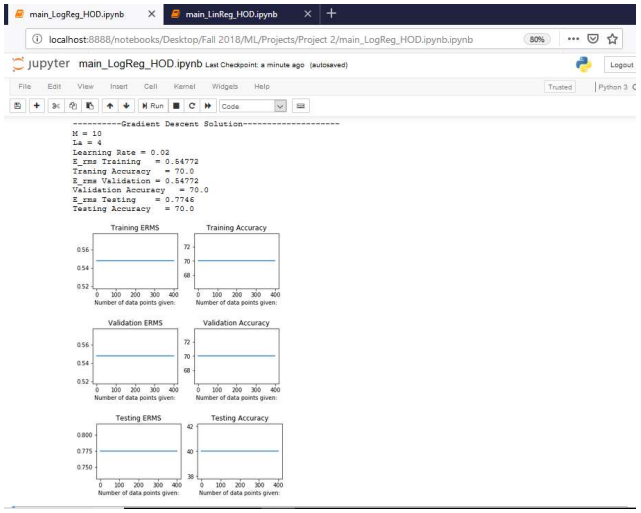
b) For Gradient Descent : M= 10, La= 2, Learning Rate = 0.01

ii) Parameter Set 2 :

a) For Closed form : M= 10, Lambda = 5



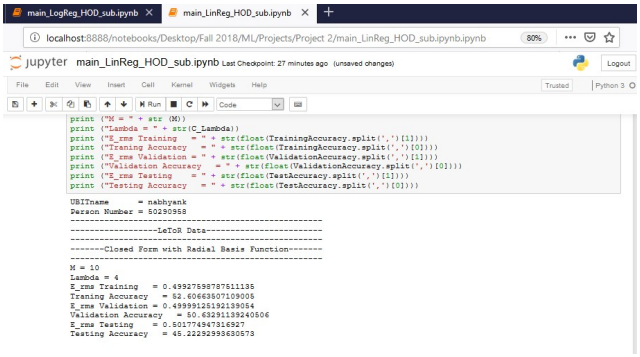
b) For Gradient Descent : M= 10, La= 4, Learning Rate = 0.02



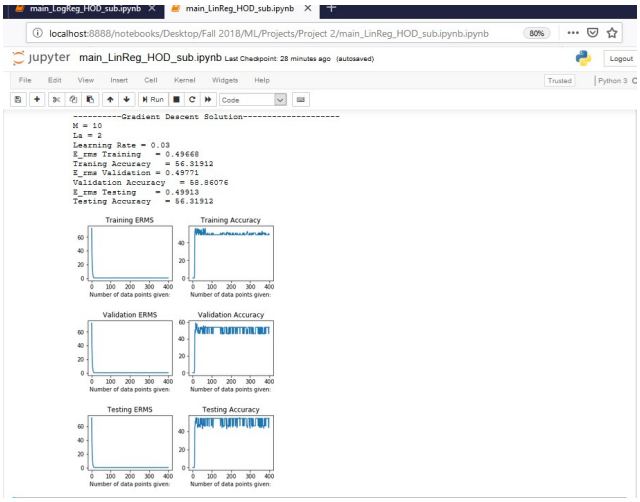
A2. For Linear Regression using Feature Subtraction Setting :

i) Parameter Set 1 :

a) For Closed form : M= 10, Lambda = 4

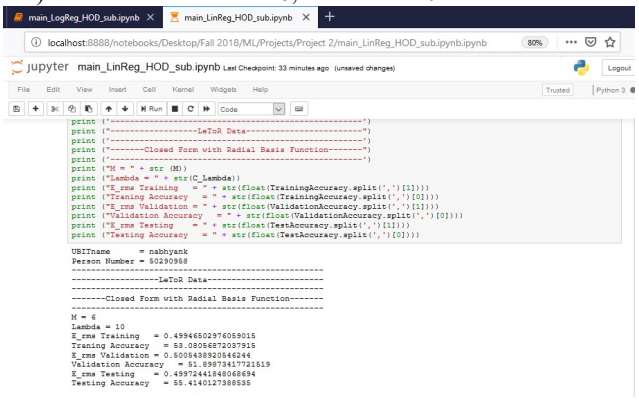


b) For Gradient Descent : M= 10, La= 2, Learning Rate = 0.03

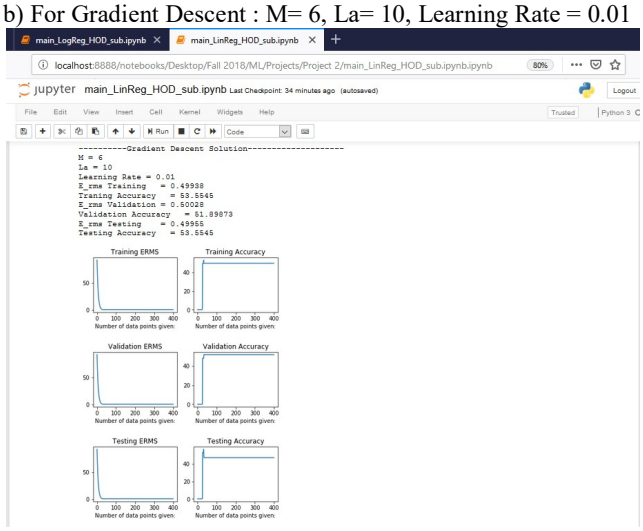


ii) Parameter Set 2 :

a) For Closed form : M= 6, Lambda = 10



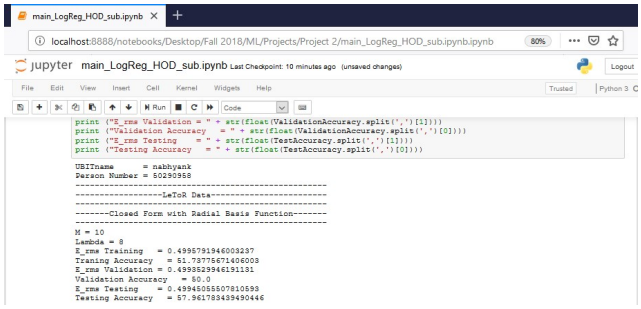
b) For Gradient Descent : M= 6, La= 10, Learning Rate = 0.01



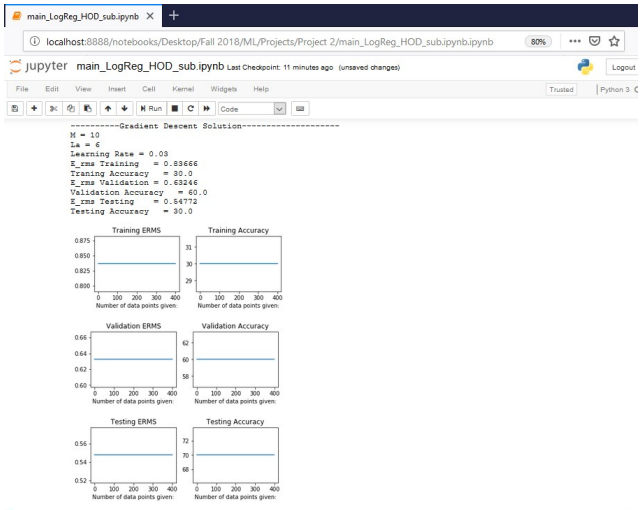
B2. For Logistic Regression using Feature Subtraction Setting :

i) Parameter Set 1 :

a) For Closed form : M= 10, Lambda = 8

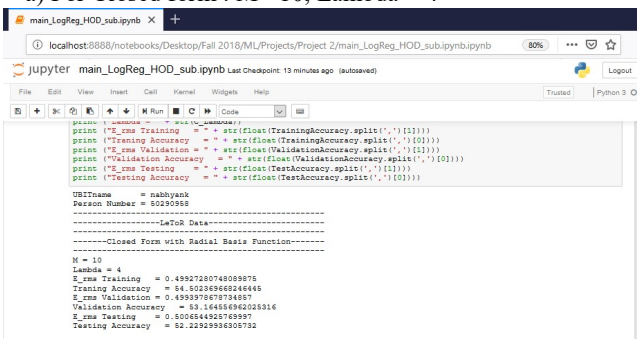


b) For Gradient Descent : M= 10, La= 6, Learning Rate = 0.03

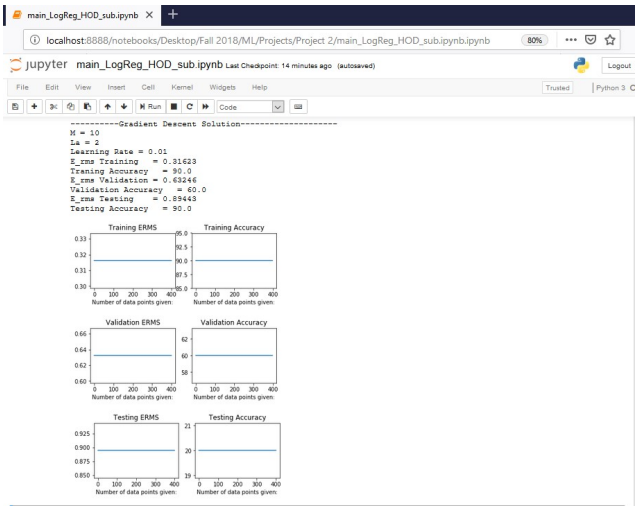


ii) Parameter Set 2 :

a) For Closed form : M= 10, Lambda = 4



b) For Gradient Descent : M= 10, La= 2, Learning Rate = 0.01



XI. WORKING AND GRAPH ANALYSIS

A. Concatenation for Human Observed Features :

The pre-processing starts with the reading of the .csv files given. After reading the .csv files for same_pairs, diffn_pair, HumanObserved-Features-Data; all elements in the same_pairs.csv file are selected and equal number of pairs are to be selected from diffn _pairs.csv file. Now, the 791x3 lists of same and different pairs are appended one below the other and then shuffle the data randomly. We then compare the same and different pair img_id_A and img_id_B with the img_id in the HumanObserved-Features-Data.csv file and find the features corresponding to the matching images. The Data is then merged with only the 18 features of the two images which are to be compared, Linear regression is performed w.r.t. the target values selected from the shuffled list of same and different pairs. For Logistic Regression the Sigmoid function is used to map the values of probabilities.

B. Subtraction for Human Observed Features:

The pre-processing starts with the reading of the .csv files given. After reading the .csv files for same_pairs, diffn_pair, HumanObserved-Features-Data; all elements in the same_pairs.csv file are selected and equal number of pairs are to be selected from diffn _pairs.csv file. Now, the 791x3 lists of same and different pairs are appended one below the other and then shuffle the data randomly. We then compare the same and different pair img_id_A and img_id_B with the img_id in the HumanObserved-Features-Data.csv file and find the features corresponding to the matching images. The Data is then subtracted with only the 9 features of the difference between two images which are to be compared, Linear regression is performed w.r.t. the target values selected from the shuffled list of same and different pairs. For Logistic Regression the Sigmoid function is used to map the values of probabilities.

If the images which are being compared have same features which are concatenated then it can be inferred that the Handwriting is of the same writer, else the Handwriting in the images belong to two different writers. In case of subtraction, if the difference between the features of two images is 0 then it can be inferred that the Handwriting is of the same writer, else the Handwriting in the images belong to two different writers.

The Graph analysis for Concatenation and Subtraction for Linear Regression show that it has a Training Validation and Testing Accuracies – 54.05 ; 60.75 ; 54.05 at maximum respectively. Accuracy can depend on several factors, especially on the given data, say if the given data is not as accurate as it should be then there can be a huge loss for Accuracy of the detection. Also if the given target values are not as accurate then the predicted values by the regression are affected on a greater scale.

The Graph analysis for Concatenation for Linear Regression show that it has a Training Validation and Testing Accuracies – 70 ; 70 ;70 at maximum respectively. For Logistic Regression there has to be an activation function, in this case the sigmoid function is used.

$$S(z) = 1 / 1+e^{-z}$$

The Logistic Regression based approach produces more accuracy compared to Linear Regression.

C. Concatenation for Human Observed Features :

The pre-processing starts with the reading of the .csv files given. After reading the .csv files for same_pairs, diffn_pair, GSC-Features; all elements in the same_pairs.csv file are selected and equal number of pairs are to be selected from diffn _pairs.csv file. Now, all elementsx512 lists of same and different pairs are appended one below the other and then shuffle the data randomly. We then compare the same and different pair img_id_A and img_id_B with the img_id in the GSC-Features.csv file and find the features corresponding to the matching images. The Data is then merged with only the

1024 features of the two images which are to be compared, Linear regression is performed w.r.t. the target values selected from the shuffled list of same and different pairs. For Logistic Regression the Sigmoid function is used to map the values of probabilities.

B. Subtraction for GSC Observed Features:
The pre-processing starts with the reading of the .csv files given. After reading the .csv files for same_pairs, diffn_pair, GSC-Features; all elements in the same_pairs.csv file are selected and equal number of pairs are to be selected from diffn _pairs.csv file. Now, all elementsx512 lists of same and different pairs are appended one below the other and then shuffle the data randomly. We then compare the same and different pair img_id_A and img_id_B with the img_id in the GSC-Features.csv file and find the features corresponding to the matching images. The Data is then subtracted with only the 512 features of the difference between two images which are to be compared, Linear regression is performed w.r.t. the target values selected from the shuffled list of same and different pairs. For Logistic Regression the Sigmoid function is used to map the values of probabilities.

If the images which are being compared have same features which are concatenated then it can be inferred that the Handwriting is of the same writer, else the Handwriting in the images belong to two different writers. In case of subtraction, if the difference between the features of two images is 0 then it can be inferred that the Handwriting is of the same writer, else the Handwriting in the images belong to two different writers.

Also since the data from the GSC Features is in large amounts compared to the Human Observed Features the pre-processing of the data takes more time and the regressions take more amount of time to process and produce an output

XII. Problem Discussion And Solutions

The Regression models with two types of Datasets are examined and studied. The Human Observed Features Dataset was much easier to pre-process w.r.t concatenation and subtraction than that of the GSC Features Dataset. Also the The Solution to the problem of time required by the regression models for large data can be improved by adding noise to the Big-Sigma inverse calculation process, still the comparison between the features is of large quantity so exponential decrease in the time cost is not expected.
Also other types of activation functions such as Re-LU, tanh etc might help in increasing the accuracy of the Logistic Regression model. Also change in the basis functions can help with more ease in processing of the regression.

GSC Dataset is sizably large in comparison to the Human Observed Dataset and takes much more time to process, train and produce outputs on the regression model.

XII. CONCLUSION

The Accuracy of the Linear & Logistic Regression models with the concatenation and Subtraction is between 50%-70%. Thus the Handwriting Analysis using the given datasets of Human Observed Features and GSC Observed Features is studied and the models were trained, tested and validated using Linear and Logistic Regressions.

References

1. https://www.microsoft.com/en-us/research/project/letor-learning-rank-information-retrieval/
2. https://pythonmachinelearning.pro/using-neural-networks-for-regression-radial-basis-function-networks/
3. https://machinelearningmastery.com/linear-regression-for-machine-learning/
4. https://en.wikipedia.org/wiki/Stochastic_gradient_descent
5. https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a
7. https://en.wikipedia.org/wiki/Training,_test,_and_validation_sets
8. https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/
9. https://books.google.com/books?id=yn6DN5hAPywC&pg=PA372&lpg=PA372&dq=Gradient+Structural+Concavity&source=bl&ots=QGeo-HdJG1&sig=3PE9Bjbr41zVJP1bYW-y4G2atDc&hl=en&sa=X&ved=2ahUKEwi518n-5LLeAhXylOAKHaPPB3EQ6AEwAnoECAcQAQ#v=onepage&q=Gradient%20Structural%20Concavity&f=false
10. https://onlinecourses.science.psu.edu/stat504/node/149/