Project-2.0 Report

Name-Shubham Khurana

UBPIN-50288698

Problem Overview-:

The problem that we are trying to solve here is to find the similarity between the samples of the known and the questioned writer, using the linear and logistic regression. Though the similarity can be decided using 0 or 1, but we would be using the real values computed out of linear regression, and for logistic regression it would be a 2-class classification problem.

DataSet-: The data set available is an AND dataset, which means that every writer was asked to write 3 manuscripts, image of AND was extracted from each of these manuscripts. Every image has got an image id , for example 1121a_num1, in which 1121 is the writer Id, a refers to the page and num1 refers to the sample number on that page. From these extracted images features have been extracted, thus leading to the formation of the 2 data sets-:

1)**Human Observed DataSet**-: This dataset is based upon the features extracted manually from the images and for each sample extracted we have got 9 features.

2)**GSC Dataset**-: This dataset has been extracted from the images using a GSC algorithm which has got 512 features.

Why Scheme is required-:

Our one data sample is actually the comparison between two image samples, and with that input for linear regression (black box) is always a vector, so to get resulting Vector, which would be acting as features for our data sample, I would be using two schemes on each data set-:

1)**Subtraction Scheme**-In this for comparing two image samples, the samples of GSC dataset from the human dataset would be subtracted, which results in the same number of features.

2)**Concatenation Scheme**-In this scheme instead of subtracting, the features of the GSC dataset would be getting concatenated with the features of the Human Observed data set. Thus total features would be equal to double the features for an image.

Preprocessing Dataset-: There are 3 csv files for each data set:

1)File having the id's of the image samples of the same writer and their targets.

2)File having the id's of the image samples of the different writers and their targets.

3)Image_id and their respective features.

Step 1-To form the training dataset, I have created a dictionary from the data of file 3, having image id's as the key and list of features as the value.

Step 2-: Forming a 2-d array depending upon the scheme being followed, fetching the list of features for the two image ids, which are being compared in file 1 or file 2, from the dictionary and placing in the row of the data matrix after applying the chosen scheme along with their respective targets.

Step 3-: After step-2 I would be having 2 matrices one available from the same writer data and another available from the different writer data. Thus to form the final data matrix I have merged the matrices of same and different writers , taking samples in equal quantity from both the sides and then randomizing the final matrix.

Step-4-: Splitting the targets(last column) and the features into two separate matrices. After the split data can be partitioned into training, validation and testing data sets.

Linear Regression:

Setup-:

1) Creation of the design matrix using gaussian radial basis function from the data matrix available for training, validation and test.
2) Applying linear regression to predict weights, using SGD for updating weights.

Value of Hyper Parameters:

Learning Rate =0.01

Number of iterations=400

Number of Clusters=10

Lambda (Regularizer)=2

Results-:

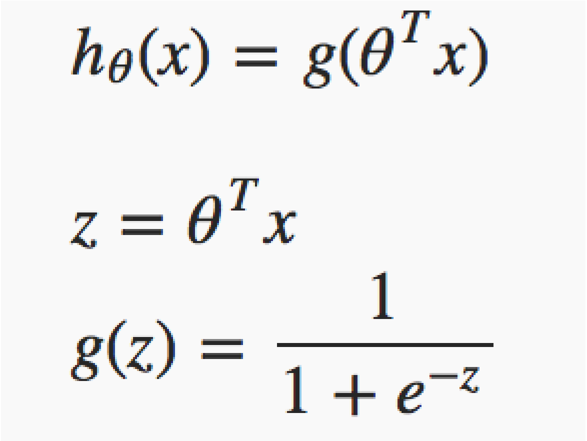
DataSet	Scheme	Erms Training	Erms Validation	Erms Testing
Human Observed	Subtraction	0.49946	0.50032	0.49849
Human Observed	Concatenation	0.4997	0.4992	0.499
GSC Data	Subtraction	0.56757	0.56385	0.55947
GSC Data	Concatenation	0.6867	0.67341	0.68716

Observations-:

- The Erms for the above hyperparameters actually remains same in the case of human observed data set but in the case of the Gsc it increases, thus impacting the accuracy.

Logistic Regression-:

A function takes inputs and returns outputs. To generate probabilities, logistic regression uses a function that gives outputs between 0 and 1 for all values of X. There are many functions that meet this description, but the used in this case is the *logistic function*. From here we will refer to it as *sigmoid.*



$$h_\theta(x) = g(\theta^T x)$$

$$z = \theta^T x$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

References- https://medium.com/@martinpella/logistic-regression-from-scratch-in-python-124c5636b8ac

Setup-:

1)Using the data matrix as the input, as sigmoid is itself a non linear function so won't be needing the basis fucntions.

2)Computing the logistic regression with the randomly initialized weights using training data set , using gradient descent for updating the weights.
3)Evaluating on the basis of the accuracy, as it is a logistic regression. Though the output of the sigmoid functions would be between 0 and 1, but we can round off the value to make it in a discreet form.

Values of Hypreparameters-:

Learning Rate =0.01

Number of iterations=400

Results-:

DataSet	**Scheme**	**Training Accuracy(percentage)**	**Validation Accuracy(percentage)**	**Testing Accuracy(percentage)**
Human Observed	Subtraction	**56.63**	54.71	54.14
Human Observed	Concatenation	**56.31**	57.86	59.87
GSC Data	Subtraction	**58.06**	54.3	57.4
GSC Data	Concatenation	**62.125**	59.9	58.3

Conclusion-:

1) For linear regression on the human data, Erms remain consistent for both the schemes but on the GSC data Erms increases for concatenation.
2) For logistic regression the human observed data shows same kind of accuracy in both the schemes but in case of the gsc accuracy increases in concatenation.