Identifying similarity between CEDAR 'AND' Images using Regression

kiran Prabhakar

November 2018

Abstract

The purpose of this project is to solve the handwriting comparison task by performing the linear regression, logistic regression and classification using neural networks.

1 Introduction

The CEDAR data set, used for training consists of set of input features for each hand-written 'AND'sample. The features are obtained from two different sources:

1. Human Observed features: Features entered by human document examiners manually. The number of features identified is 9
2. GSC features: Features extracted using Gradient Structural Concavity (GSC) algorithm. The number of features identified by this algorithm is 512.

2 Definitions

2.1 Logistic Regression

Logistic Regression is a predictive analysis model, where the target is categorical. The output of the model is the probability value for a particular category. The output of the model is the sigmodial output of $\theta^T X$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T X}} = p(y = 1|x, \theta)$$

$$Cost(h_\theta(x), y) = -ylog(h_\theta(x)) - (1 - y)log(1 - h_\theta(x))$$

After the minimizing the above cost function with respect to θ the gradient function is obtained.

$$\theta_j = \theta_j - \frac{1}{N} \sum_{i=1}^{N} \Big\{ h_\theta(x_i) - y_i \Big\} x_i^j$$

2.2 Vectorization

Vectorization is a performance optimization technique for updating the weights in gradient descent algorithm. The technique leverages the matrix operations provided by linear algebra libraries and avoids using of multiple for loops. In the current project, vectorization technique is used to update the weights in case of linear and logistic regression. The vecotrized equation for gradient descent is as below.

$$\theta = \theta - \frac{1}{n}\mathbf{X}^T(\theta\mathbf{X} - y) + \frac{\lambda}{n}\theta$$

3 Data Pre-processing

The data set is partitioned using three different approaches.

- Unseen Writer Partitioning: Same writer will not be present in both testing and training data.
- Shuffled Writer Partitioning: The entire data set is shuffled before considering the data points.
- Seen Writer Partitioning: In this we take 80% of data for training and validation, the remaining 20% for testing.

In the current project best results are obtained by using seen writer partitioning approach.

The given data set consists of two kind of data pairs, same pairs and different pairs, for both human observed and GSC data set.

- Same pairs: For both human and GSC data set it has target as 1.
- Different pairs: For both human and GSC data set it has target as 0.

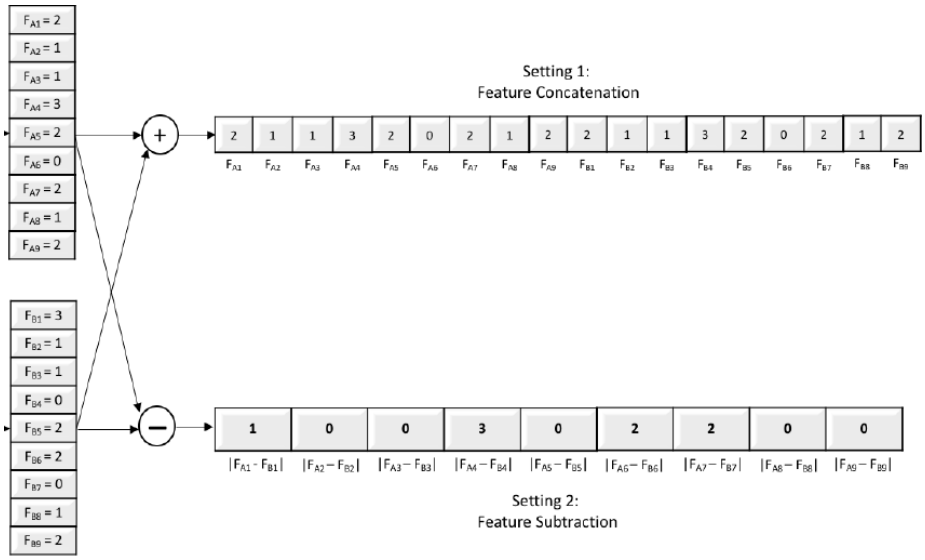
3.1 Concatenation

The Concatenation operation performed on the same pairs or different pairs on either data set, will append the features of one image to the another image and will produce the concatenated output as shown in the below image for human observed data set.

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

3.2 Subtraction

The subtraction operation is performed on the both same pairs and different pairs data set, where the feature value of each image is subtracted from the corresponding value of other image in the pair. The absolute difference is considered for the subtracted feature set.



After performing Subtraction,

- The size of each pair in human observed data is 9 bits.
- The size of each pair in GSC observed data is 512 bits.

As the GSC data set is quite heavy, 3000 data points each from same and different pairs are taken to perform regression.

The algorit

4 Linear Regression

Linear Regression is a predictive modeling technique in which a model is derived from existing input and output values. Using the model, the outcomes of feature inputs can be predicted. The number of features on which the input data is relied will play a key role to in constructing the model.

Four linear regressions is performed on the given data set. Two for human observed data (same, different) and two for GSC (same, different)

4.1 Human observed data

Linear Regression is performed on both concatenated and subtracted human observed data. The linear regression is performed using the entire data set for 1000 epochs and the corresponding accuracy's are measured

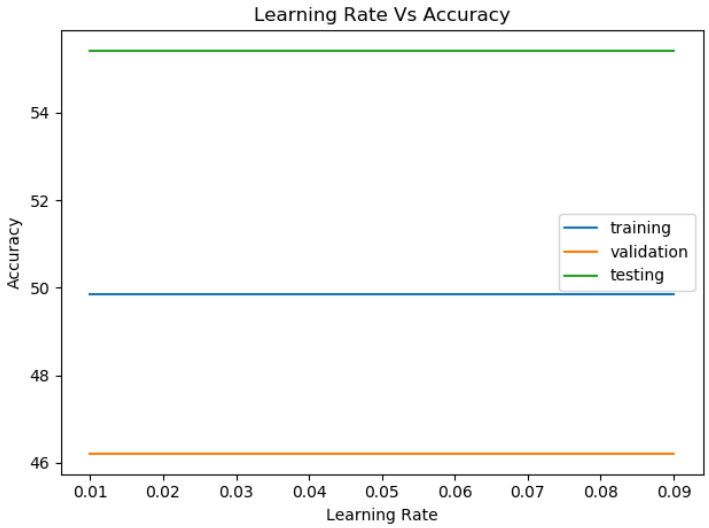
Hyper parameters for Concatenated human linear regression:

Number of clusters	Learning Rate	Regularizer	Accuracy
10	0.01	2	55.41

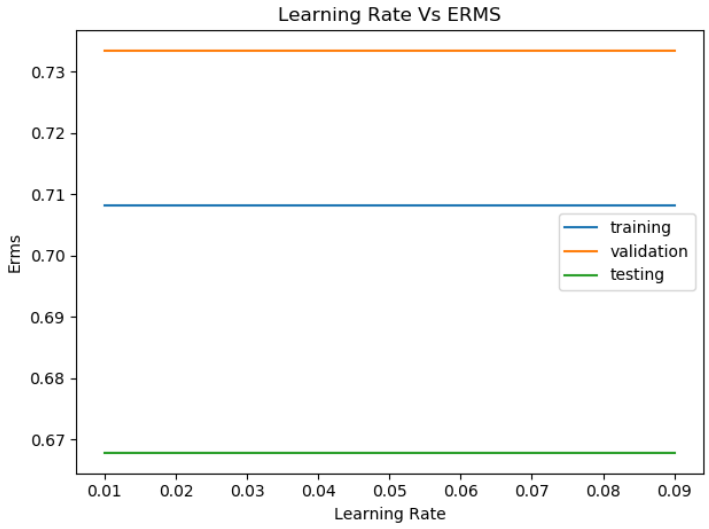
Hyper parameters for Subtracted human linear regression:

Number of clusters	Learning Rate	Regularizer	Accuracy
10	0.01	2	56.05

The accuracy remains same even after changing the hyper parameters. So, the linear regression on concatenated or subtracted human data gives results with less accuracy. Below graph indicates the accuracy and error for various values of learning rates.



The ERMS graph for various values of learning rate is as below.



4.2 GSC data

By performing linear regression on concatenated and subtracted GSC data for 1000 epochs, below are the observation results obtained.

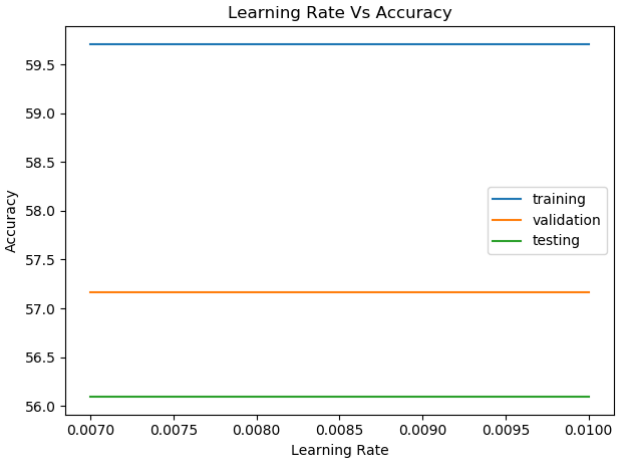
Hyper parameters for Concatenated GSC linear regression:

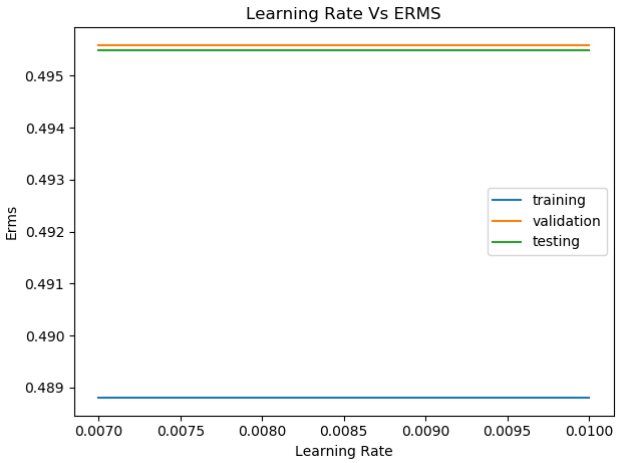
Number of clusters	Learning Rate	Regularizer	Accuracy
10	0.01	2	59.76

Hyper parameters for Subtracted GSC linear regression:

Number of clusters	Learning Rate	Regularizer	Accuracy
10	0.01	2	54.75

The below graph is drawn for accuracy and error for GSC concatenated data with multiple values of learning rate.





5 Logistic Regression

As the current project is focused on classification of the images, choosing the logistic regression for performing classification would be one of the better choices. The sigmodial function used in logistic regression gives the probability for classification, as explained in section 2.1, which is useful for comparing with the actual target value.

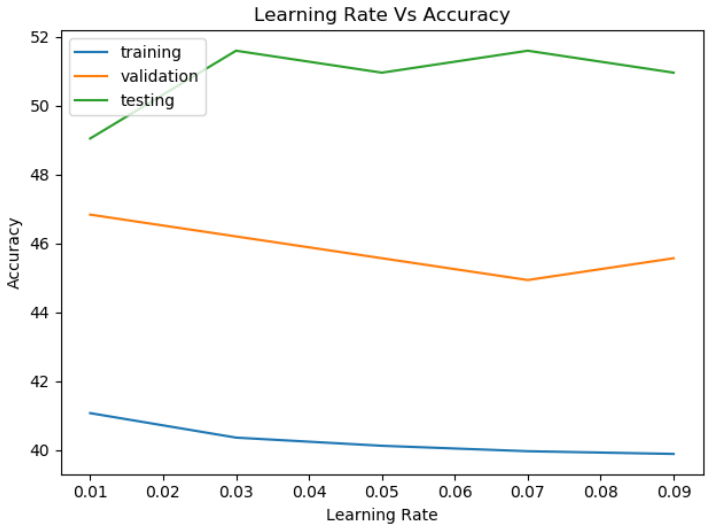
After applying the trained model on the testing data set and as the outputs are in the range [0,1], we assign value '1' if the output is ≥ 0.5 and 0 if output < 0.5

5.1 Human observed data

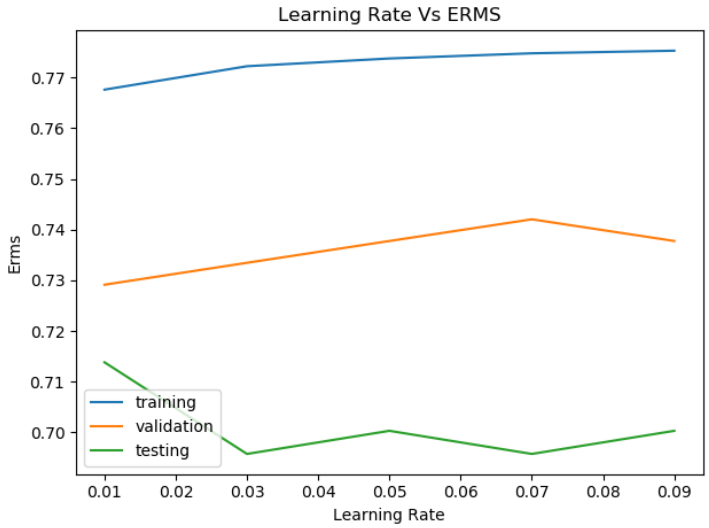
Logistic regression is performed on both concatenated and subtracted human data set and the results are obtained as below.

Hyper parameters for Human concatenated data

Learning Rate	Regularizer	Accuracy
0.01	2	51.41

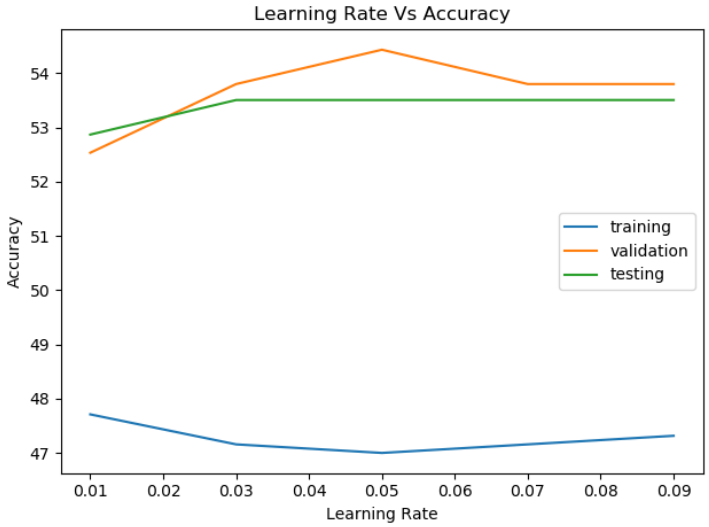


The graph for ERMS is as below

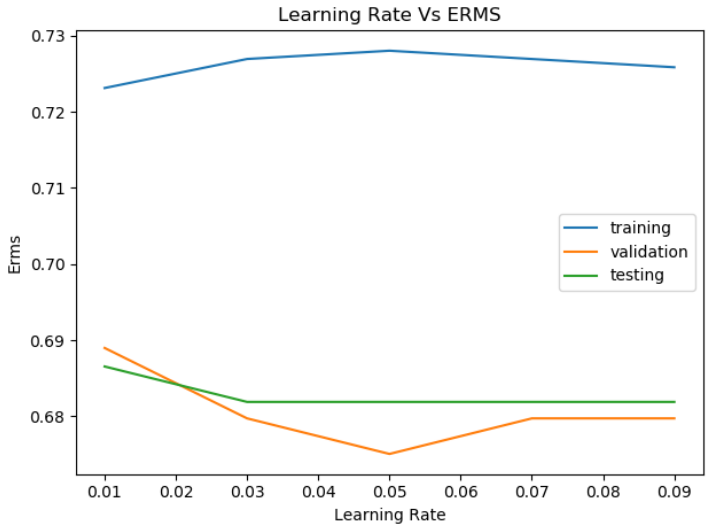


Hyper parameters for Human Subtracted data

Learning Rate	Regularizer	Accuracy
0.01	2	53.55



The graph for ERMS is as below

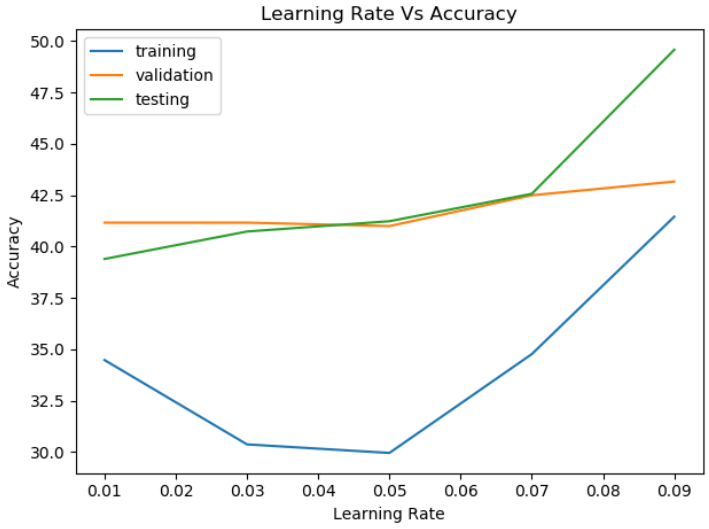


5.2 GSC data

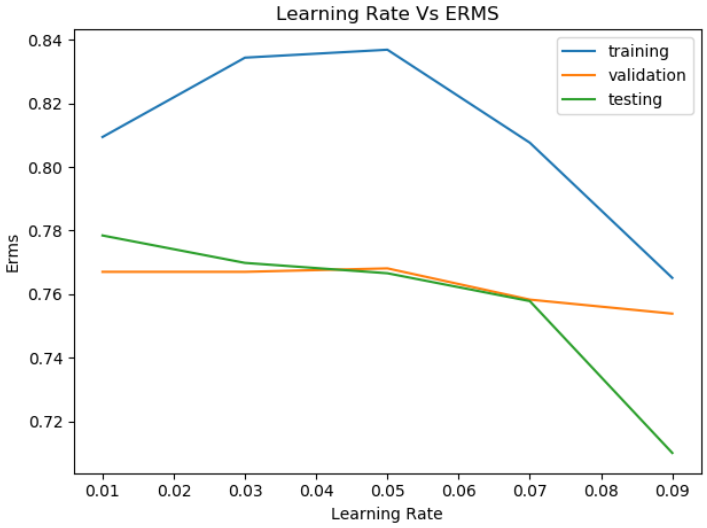
Logistic regression is performed on Concatenated and Subtracted GSC data, which has 1024 and 512 features respectively.

Hyper parameters for GSC concatenated data

Learning Rate	Regularizer	Accuracy
0.01	2	50.01



The graph for ERMS is as below



Hyper parameters for GSC Subtracted data

6 Neural Network

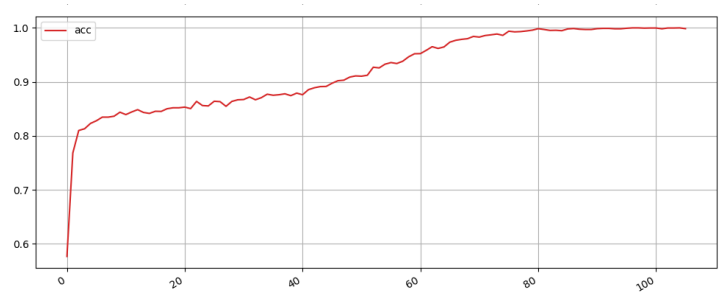
As the approaches of linear regression and logistic regressions didn't provide adequate output, neural network is used to perform the classification operation. The accuracy of the model has considerably increased after following the approach of neural networks. In this method, the number of inputs to the network depends on the number of features present in the training data.

6.1 Human observed data

As the concatenated human observed data has 18 features, we use 18 inputs for the network to perform the classification. The network is trained with the below setting of hyper parameters which produces the below graph.

Hyper parameters and accuracy for concatenated Human observed data

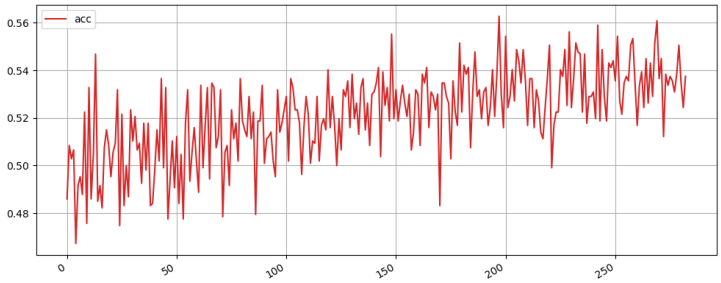
No. of Hidden layers	optimizer	Activation	Inputs	Testing Accuracy
2	RmsProp	Sigmoid	18	90.17



By using the same setting for subtracted human observed data, below are results obtained. The graph shows the training accuracy of the model.

Hyper parameters and accuracy for subtracted Human observed data

No. of Hidden layers	optimizer	Activation	Inputs	Testing Accuracy
2	RmsProp	Sigmoid	9	94.17

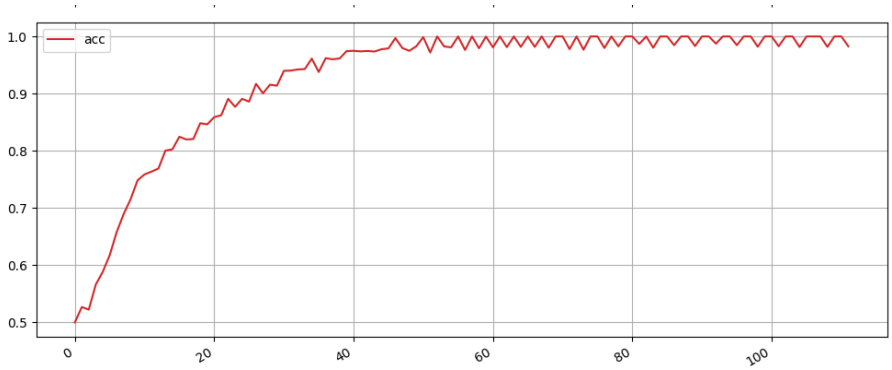


6.2 GSC data

The concatenated GSC data has 1024 features and the subtracted GSC has 512 features. The model is trained with the below hyper parameters and the following accuracy and training accuracy graph is obtained.

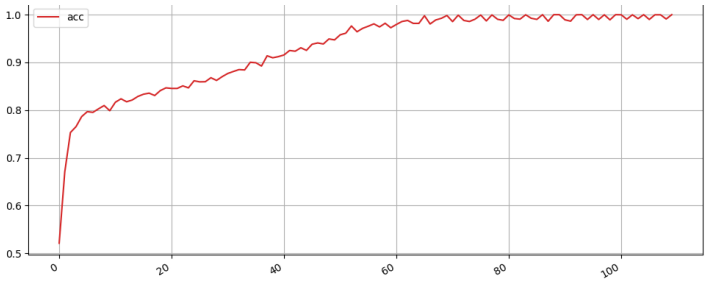
Hyper parameters and training accuracy graph for concatenated GSC observed data

No. of Hidden layers	optimizer	Activation	Inputs	Testing Accuracy
2	RmsProp	Sigmoid	1024	60.21



Hyper parameters and training accuracy graph for subtracted GSC observed data

No. of Hidden layers	optimizer	Activation	Inputs	Testing Accuracy
2	RmsProp	Sigmoid	1024	58.63



7 Conclusion

From the above observations it is evident that, the high accuracy for the classification is obtained by using neural network technique with subtracted human observed data.

8 References

- https://keras.io
- https://www.tensorflow.org/tutorials
- https://medium.com/data-science-group-iitr
- https://towardsdatascience.com