Machine Learning Project 2

Dhanashree Solanke
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
ddsolank@buffalo.edu

Abstract

We take up the problem of identifying two handwriting images and analyzing whether they are from the same author or not. We formulate this problem using linear regression and logistic regression. We will test our model on two datasets, human observed datasets and gsc featured datasets.

1 Datasets

We take up two datasets where we have images of word "AND". In Human dataset, there are 9 features which are extracted manually. And in the GSC Dataset, there are 512 features per image extracted automatically. Now, we are given pair of images with target values as 0 or 1 based on whether they are from same author or not as training data. Basically, for each given pair of images we fetch the features and come up with our dataset, such that the final dataset would contain two images and their features first concatenated and then subtracted. Thus, we come up with the model for Linear and Logistic Regression based on these 4 datasets.

2 Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is an explanatory variable, and the other is a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

Multivariate Linear Regression

Multivariate is like the basic/simple linear regression but with multiple independent variables contributing to the dependent variable and hence multiple coefficients to determine.

$$Y_i=\alpha+\beta_1 x_i^{(1)}+\beta_2 x_i^{(2)}+\ldots.+\beta_n x_i^{(n)}$$

Y_i is the estimate of i^{th} component of dependent variable y, where we have **n** independent variables and x_{ij} denotes the ith component of the j^{th} independent variable/feature.

2.1 Model

For Linear Regression we used the gradient descent solution to find the parameters for our equation. We divide up our dataset into three parts of 80:10:10 ratio forming training, validation and testing data respectively. We start our gradient descent solution by initializing the weights to zero. Then for iteration we update the weights until we achieve minimum error. We measure accuracy in Root mean Square error here.

3 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. In logistic regression, the target variable is binary, i.e. it only contains data coded as 1 (TRUE) or 0 (FALSE).

The logistic function is defined as

transformed = 1 / (1 + e^-x)

Where e is the numerical constant Euler's number and x is an input we plug into the function.

The goal is to find the best fitting model to describe the relationship between the outcome variable and a set of independent variables. Logistic regression generates the coefficients of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

Logistic regression equation

$$logit(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

Odds=p/(1-p)

and

Logit(p)=ln(p/(1-p))

Estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

3.1 Model

Our model for Logistic regression is similar to Linear Regression, just the technique to update the weights will differ. We start with initializing weights to 0. And each iteration of gradient descent we update the weights using Sigmoidal Function. We measure our accuracy using the Cross-Entropy Function here.

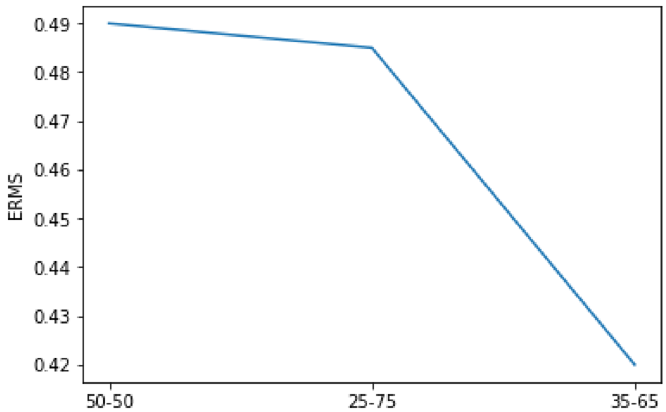
4 Changing Hyper Parameters for Linear and Logistic Regression

- I have tried taking different number of same writer and different writer pairs thus

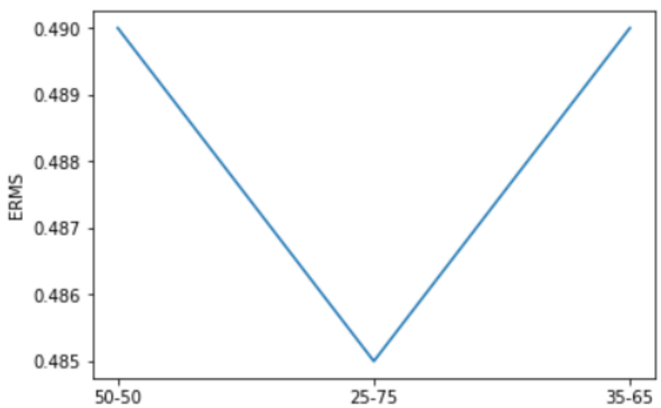
varying the data set sizes.
The pairs were in following combinations
50% - 50% , 35% - 65%, 25% - 75%

Linear Regression

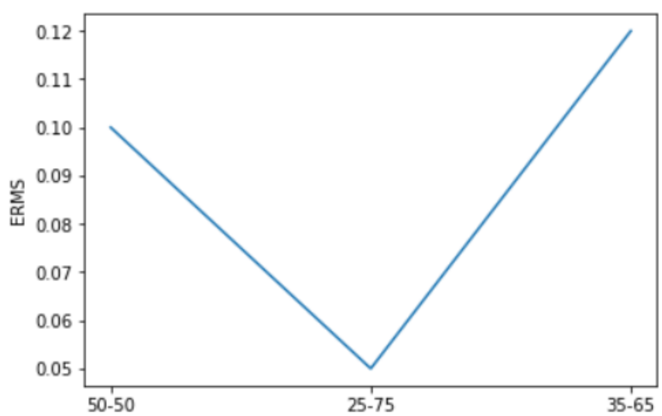
- Human Observed – Concatenation Graph



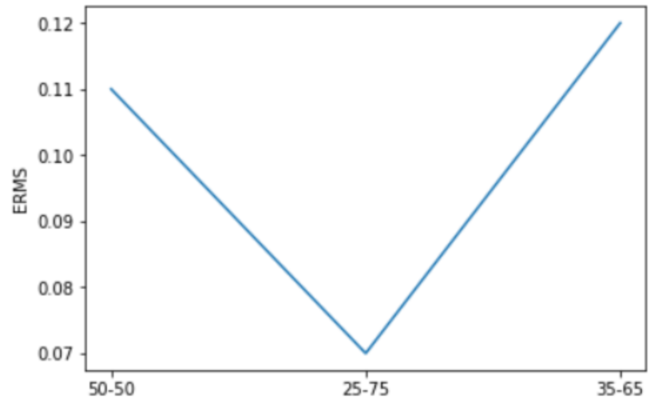
- Human Observed – Subtraction Graph



- GSC Observed - Concatenation

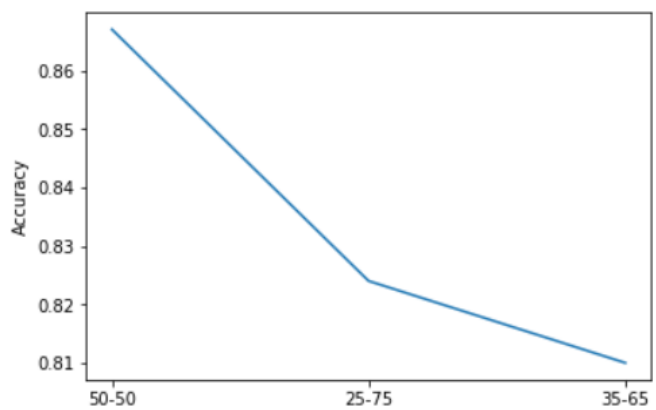


- GSC Observed – Subtraction

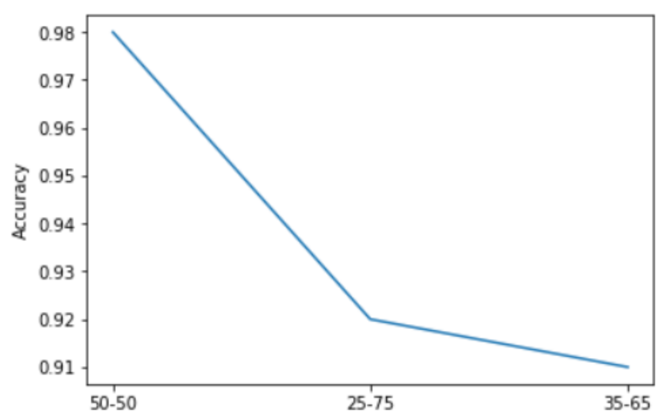


Logistic Regression

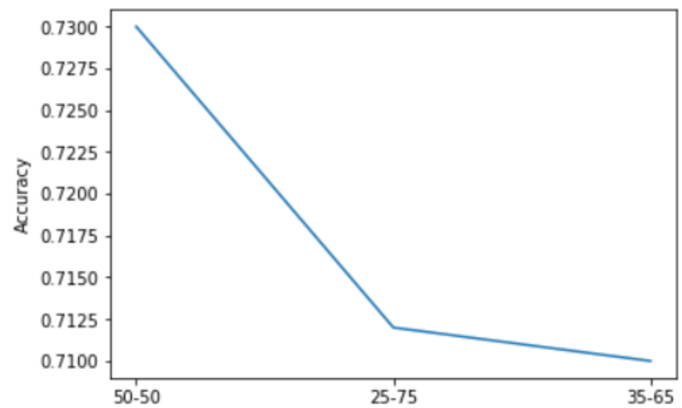
- Human Observed – Concatenation Graph



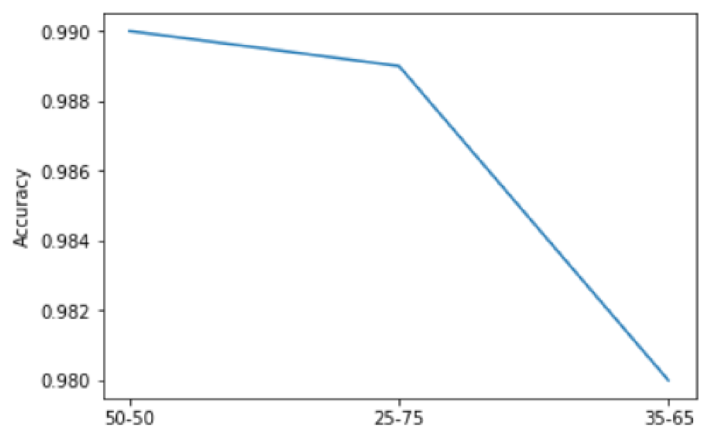
- Human Observed – Subtraction Graph



- GSC Observed - Concatenation



- GSC Observed – Subtraction



Similarly, I have tried changing learning rates, number of iterations in gradient descent, etc. Finally, the best possible erms and accuracy values are obtained by following parameters in both the models.

M = 10
Learning Rate = 0.01
Lambda = 2

Final accuracy we get are as follows:
Linear Regression:
1. Human Observed (Concatenation) – 0.495
2. Human Observed (Subtraction) – 0.489
3. GSC Observed (Concatenation) – 0.1
4. GSC Observed (Subtraction) – 0.24

Logistic Regression
1. Human Observed (Concatenation) – 0.867
2. Human Observed (Subtraction) – 0.985
3. GSC Observed (Concatenation) – 0.738
4. GSC Observed (Subtraction) – 0.99s

5 Neural Networks

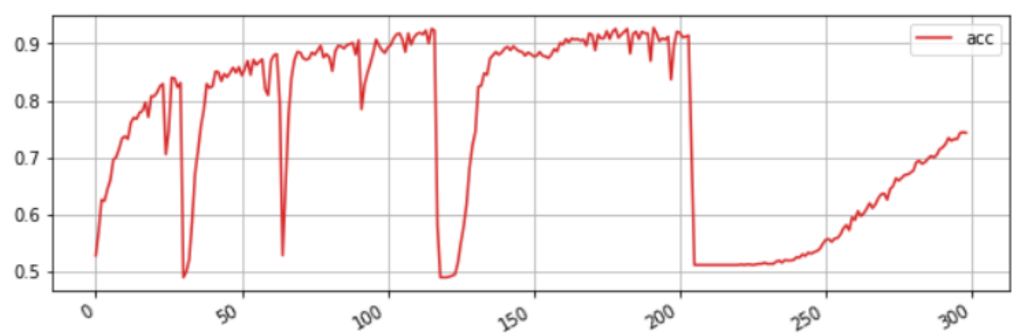
Neural nets are a means of doing machine learning, in which a computer learns to perform some task by analyzing training examples. Usually, the examples have been hand-labeled in advance.

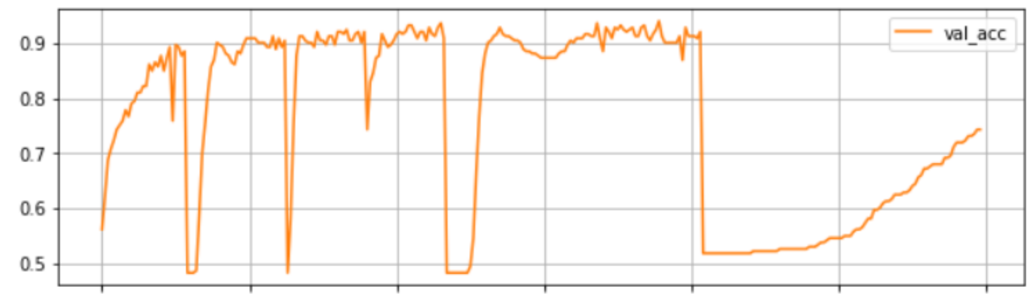
Modeled loosely on the human brain, a neural net consists of thousands or even millions of simple processing nodes that are densely interconnected. Neural nets are organized into layers of nodes, and they're "feed-forward," meaning that data moves through them in only one direction. An individual node might be connected to several nodes in the layer beneath it, from which it receives data, and several nodes in the layer above it, to which it sends data.

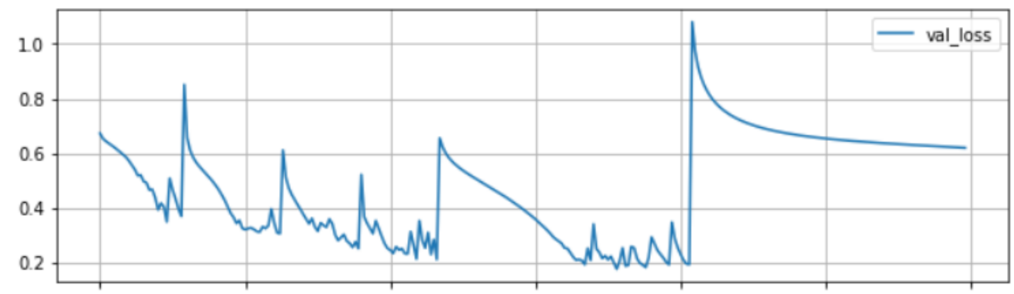
5.1 Model

We develop our model using the Keras implementation of Neural Networks. We use SGD as our optimizer, accuracy as our measure and categorical cross entropy as our loss function. The best accuracy is obtained when we have below hyper-parameters

drop_out = 0.2
first_dense_layer_nodes = 256
second_dense_layer_nodes = 2
validation_data_split = 0.2
num_epochs = 1000
model_batch_size = 128
tb_batch_size = 32
early_patience = 100







6 Conclusion

Thus, we have developed a linear regression, logistic regression and a neural network with best possible accuracies for our handwriting recognition problem.