Project 2

Tanmay Singh
tanmaypr@buffalo.edu

Abstract

In the following report we analyse the accuracy and error on performing linear regression,logistic regression and using neural network on two pair of datasets.

1.Introduction

In project 2 we have been provided with two datasets human observed dataset and GSC datasets. The datasets together provide information on whether the word is written by the same person or not. The human observed dataset consist of 9 features whereas the GSC dataset consist of 512 features.

2.Processing Datasets

Initially we have three different csv files same pair, different pair and features. The same pair csv file consist of same pair of images where the word is written by the same writters,similarly different pair csv file consist of images where the word is written by the different writters.The Feature csv consist of features for each image which is 9 in case of human observed datasets and 512 in case of GSC datasets. The target value when the images belong to the same writter is 1 otherwise it is 0.

3.Hyperparameters

The linear regression model has the following hyperparameters
-Number of clusters
-Regularization(Lambda)
-Learning Rate

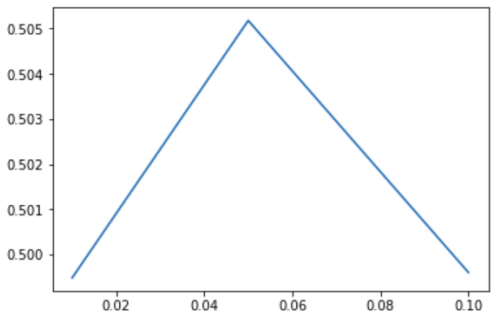
Neural Networks
-binary cross-entropy:-we are using binary cross entropy because the final output has only two possibilities either 1 or 0 I.e either the writer is same or different.
-we are taking the number of input nodes as equal to the number of features therefore for human observed the number of nodes in the first hidden layer is 9 and for GSC it is 512
-relu activation is used between the first and the second hidden layer
-sigmoid activation is used between the second hidden layer and the last layer so that the final output is either 0 or 1.

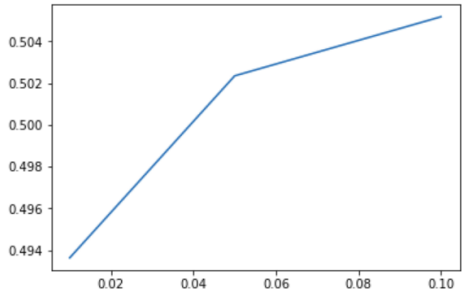
4.Linear Regression Model

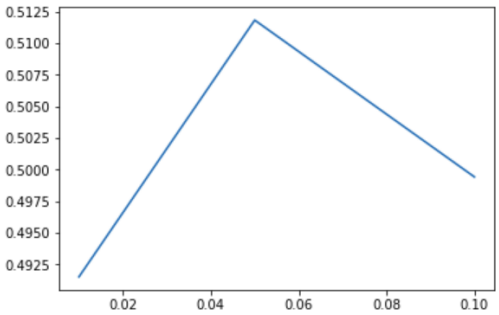
we calculate the efficiency based on the error E_rms values

1)Human observed concatanated

E_rms on Y axis vs Learning Rate on X axis for training ,validation and testing:-

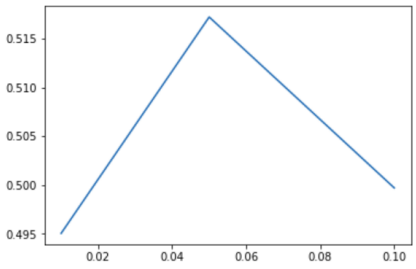


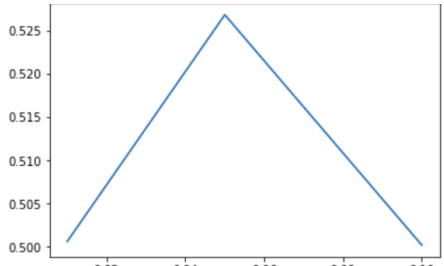


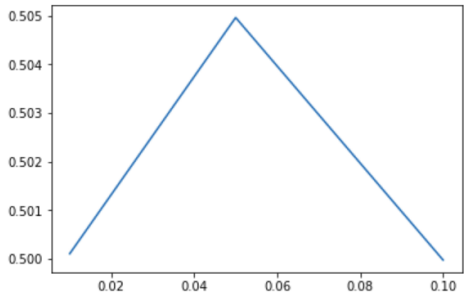


2)Human observed subtracted

E_rms on Y axis vs Learning Rate on X axis for training ,validation and testing:-

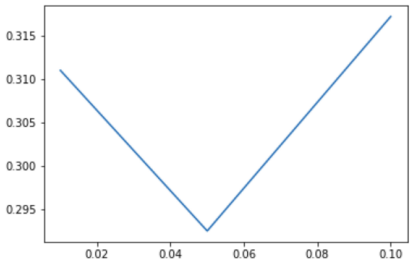


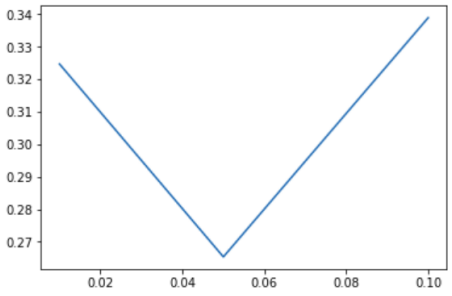


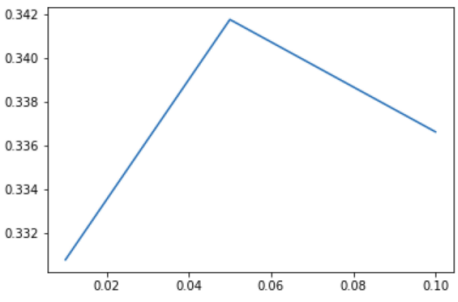


3)GSC concatanated

E_rms on Y axis vs Learning Rate on X axis for training ,validation and testing:-

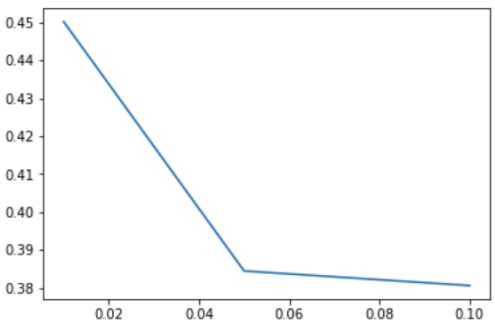


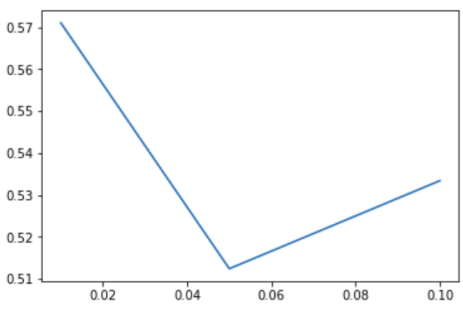


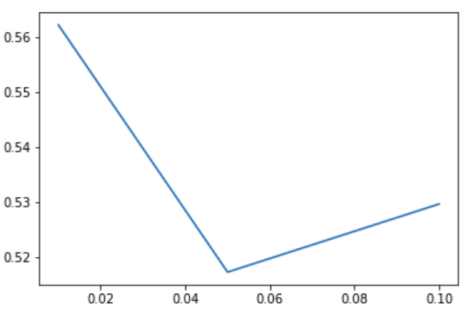


4)GSC subtracted

E_rms on Y axis vs Learning Rate on X axis for training ,validation and testing:-

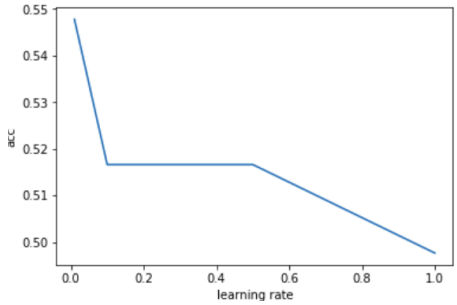


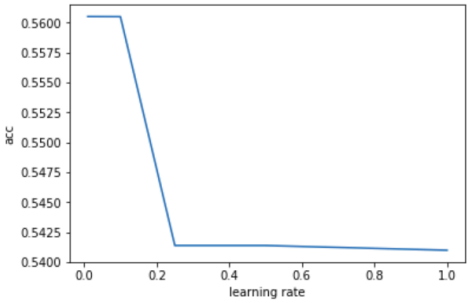


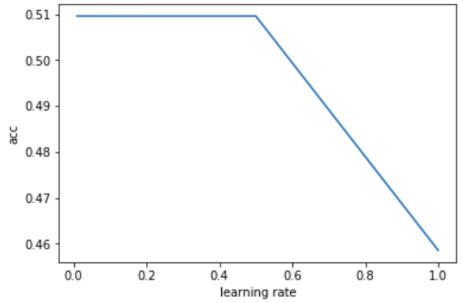


4.Logestic Regression Model

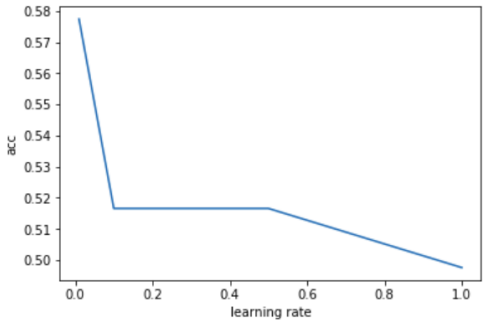
1)Human observed concatanated

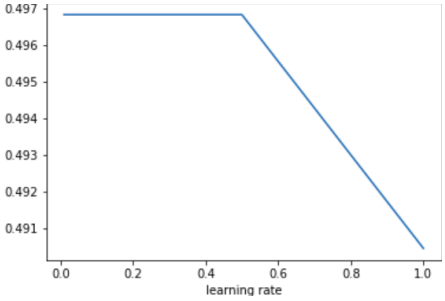


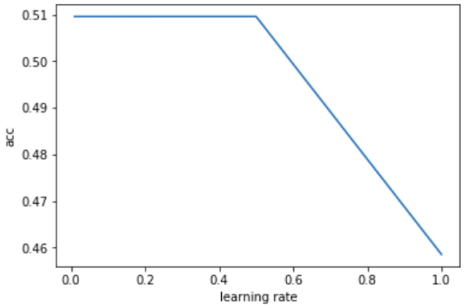




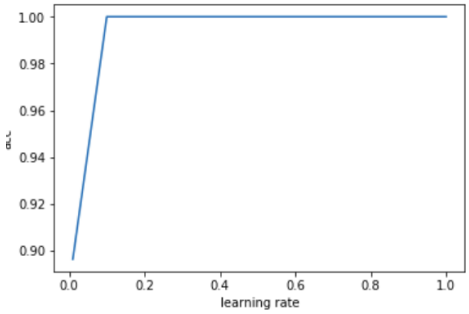
2)Human observed subtracted

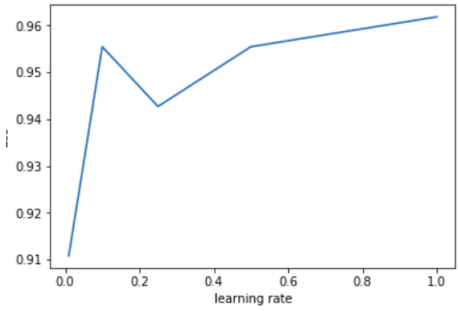


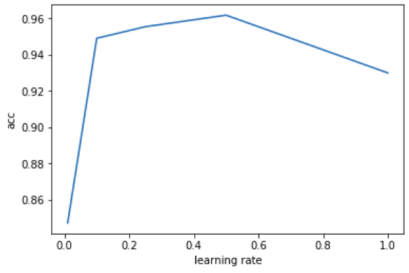




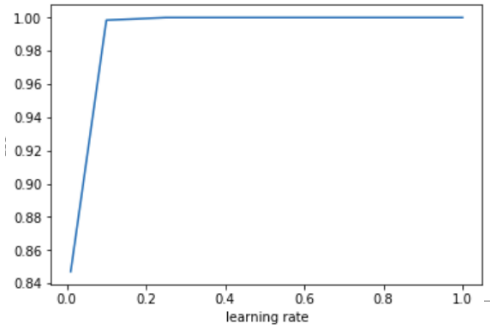
3)GSC observed concatanated

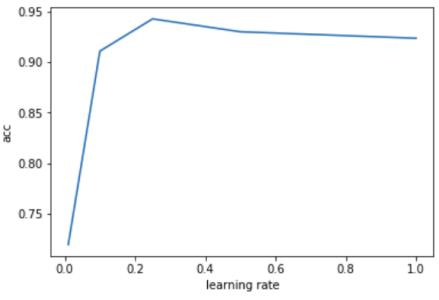


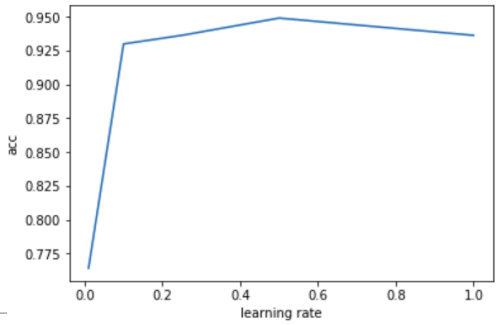




4)GSC observed subtracted

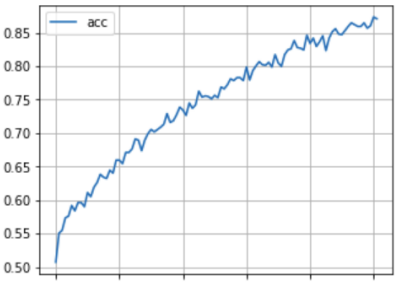


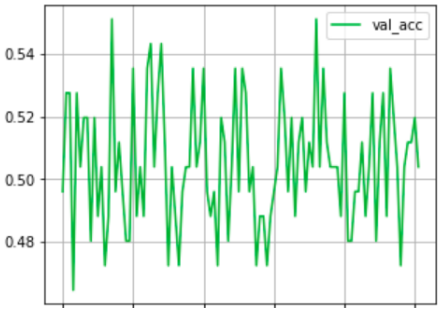




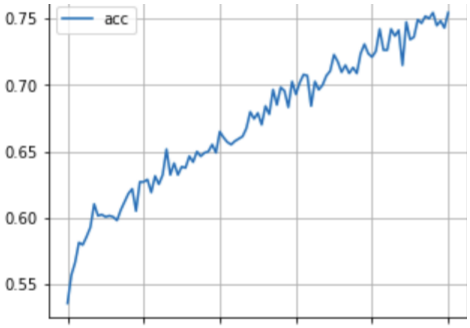
5.Neural Network using keras

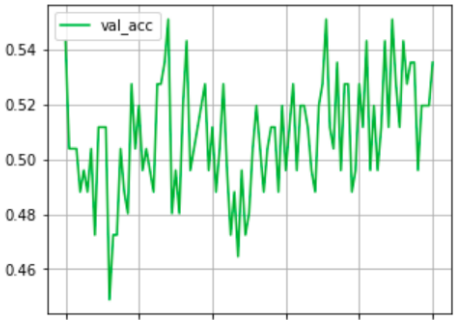
1)Human observed concatanated: acc vs number of epoch



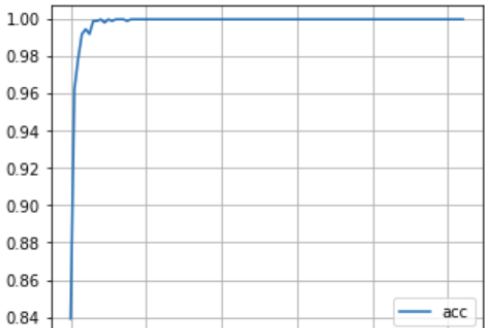


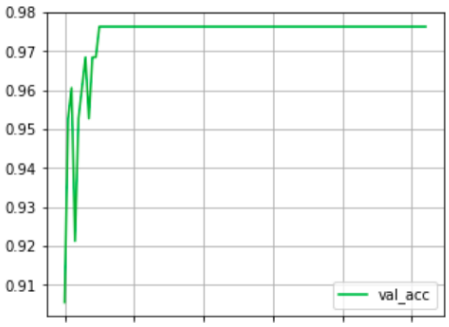
2)Human observed subtracted: acc vs number of epoch



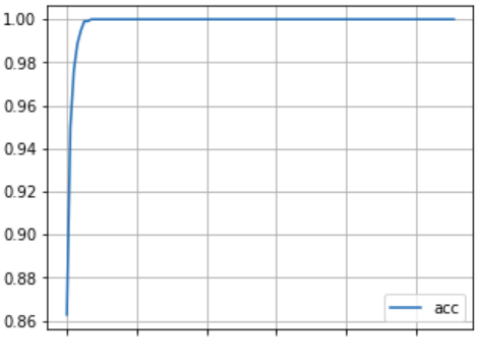


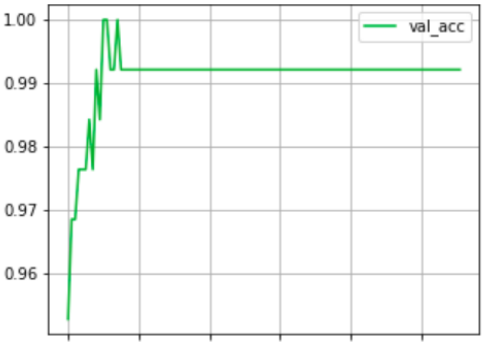
3)GSC concatanated: acc vs number of epoch





4)GSC subtracted :acc vs number of epoch





6.ERROR

In linear regression we using E_rms,Root mean square to determine the difference between the target value and the value that is generated by our model. We calculate the root mean square for training,testing and validation to determine the changes that needs to be implemented to the weight.

In logistic regression we use log loss or cross entropy loss to calculate the error this is because the rms will give many local minima leading to non convex gradient descent but log loss will provide us with a convex gradient descent making it easier to find the local minima.

Refrences

1 https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html
2 project 1.1
3 project 1.2