Handwriting Comparison

Gagan Suneja
School of Management
University at Buffalo
Buffalo, NY 14260
gagansun@buffalo.edu

Abstract

The report presents implementation of machine learning model that compares handwriting of two samples and predicts if they have been written by the same person or not. The model also compares different approaches to solve the problem, namely- Stochastic Gradient Descent, Logistic Regression, using two operations- concatenation of features and subtraction of features

1 Introduction

Stochastic Gradient Descent

The stochastic Gradient Descent algorithm traverses the data points, each at once, and simultaneously updates the weights thereby reducing the E_{RMS} with each iteration at a learning rate ɳ, until a minimum is reached upon which the algorithm converges. For each of the iteration, weights are updated as follows:

$$w^{\tau+1} = w^{\tau} + \Delta w^{\tau},\ where\ \Delta w^{\tau} = -\eta^{\tau}\nabla E$$

The learning rate should be optimally chosen as a low learning result in slow convergence of the algorithm and a high learning rate will result in overshooting of the model which leads to a high increase in Error value. The algorithm is implemented with below relationship-

$\nabla E = \nabla E_W + \lambda \nabla E_W$

Logistic Regression

Logistic regression is a regression analysis algorithm used when the dependent variable is dichotomous. The predicted output can be either of the two values (0 or 1). Here, usually, sigmoid function is used to achieve the dichotomous functionality. A sigmoid function is given as

$$S(x) = \frac{1}{1 + e^{-x}}$$

Dataset

The dataset comprises feature data of "AND" images extracted from CEDAR Letter dataset. The feature data used in the project is of two types, namely- Human Observed Dataset and GSC Dataset. The Human Observed dataset comprise of 9 features and the GSC dataset comprise of a total of 512 features. For each of the dataset, there are three types of data files-

1. File containing features of each of the image ids.
2. File with matching pair of image ids (same_pairs.csv)
3. File with not matching pair of image ids (diffn_pairs.csv).

2 Dataset Preparation

The dataset has been prepared by merging the features in feature data file with the same pairs and different pair files. There are two types of merge operations performed on the dataset- concatenation (which involved concatenation of the number of features of

matching and non-matching image pairs resulting in 18 features) and subtraction (which involved difference in the features of matching and non-matching image pairs resulting in 9 feature columns) for Human dataset. Similarly, for GSC data, there are 1024 features for concatenation and 512 features for subtraction. After this, equal number of samples (791 for Human dataset and 5000 for GSC dataset) have been taken from the same pairs and different pair datasets. Post the sample creation, the dataset is shuffled to minimize the probability for biased results.

3 Analysis

Human Observed Dataset

Below are the findings for Stochastic Gradient Descent for Human Data

Table 1: SGD for Concatenation Operation

Learning Rate	ERMS Testing
0.1	0.6639971
0.05	0.499644846
0.01	0.507550654
0.005	0.507469217
0.001	0.586446862
0.0005	1.199041008

Table 2: SGD for Subtract Operation

Eta	ERMS Testing
0.01	0.519052
0.0001	0.499567
0.0005	0.498927
0.001	0.498893
0.0011	0.498925
0.0012	0.498963
0.0013	0.499004
0.0014	0.499047
0.0015	0.499089
0.0016	0.499131
0.0017	0.499172
0.0018	0.49921
0.0019	0.499247
0.002	0.499281
0.003	0.49959
0.004	0.500193
0.005	0.501475
0.01	0.519052

Table 3: Logistic Regression for Concatenation operation

Learning Rate	ERMS Testing
0.001	0.931064041
0.002	0.500042686
0.0025	0.49862551
0.003	0.500391393
0.005	0.509926738
0.007	0.522458951
0.01	0.543029566
0.05	0.624863778
0.1	0.62138214

Table 4: Logistic regression for Subtract Operation

ETA	ERMS Testing
0.001	0.497640062
0.002	0.497069168
0.003	0.496903474
0.004	0.496812289
0.005	0.496694911
0.006	0.496538358
0.007	0.496358264
0.01	0.495828078
0.02	0.494906129

0.2	3.04E+134
0.3	Overflow
0.5	Overflow

0.03	0.494735023
0.035	0.494811043
0.04	0.494979071
0.045	0.495242011
0.05	0.495610717
0.06	0.496733347
0.07	0.498494536

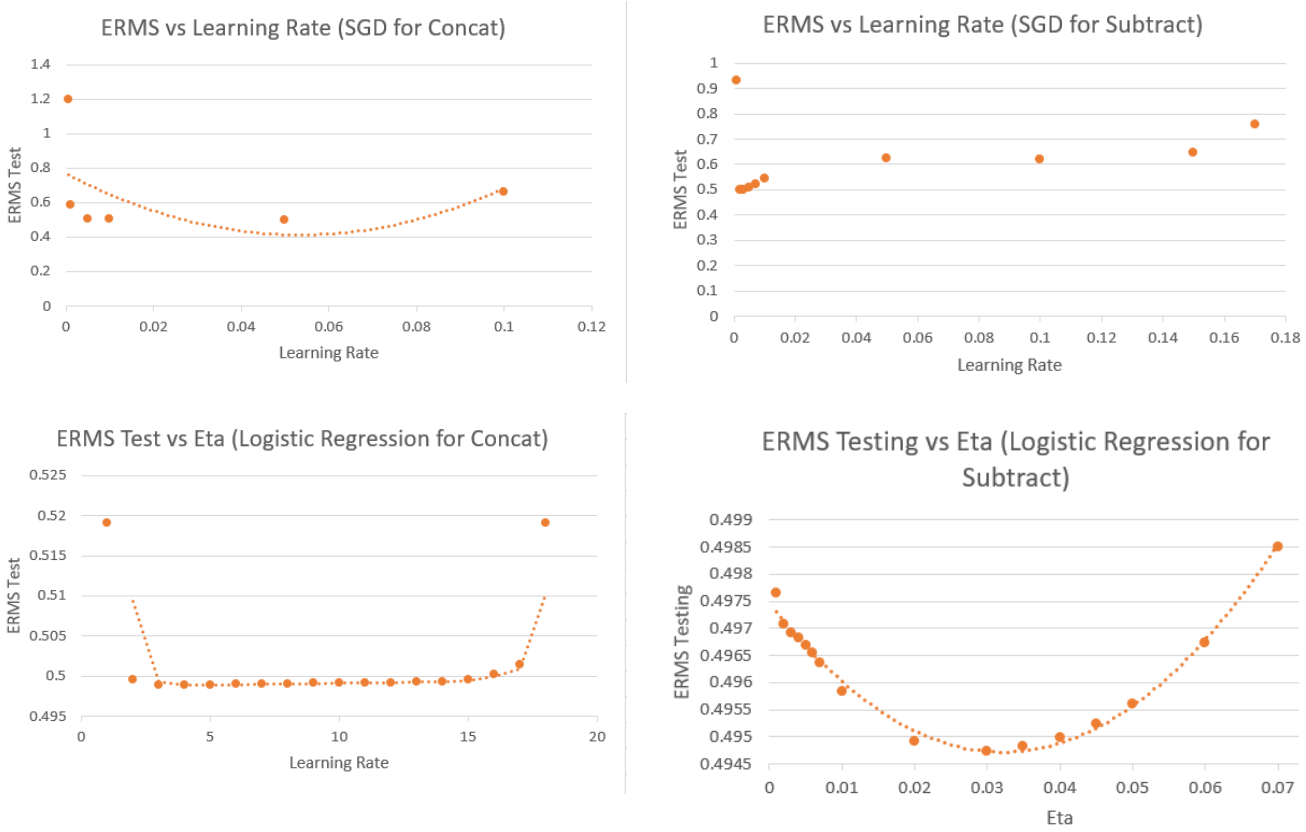

Fig 1

GSC Dataset

Below are the findings for the GSC dataset

Table 5: SGD for Concatenation Operation

Learning Rate	ERMS Testing
0.0002	1.68155698
0.0008	1.196075282
0.0009	0.685371021
0.00098	0.536246233
0.00099	0.529337341
0.001	0.524346559
0.0011	0.538146618

Table 6: SGD for Subtraction Operation

Learning Rate	ERMS Testing
0.0007	0.740237607
0.0008	0.590165222
0.0009	0.579470366
0.001	0.593179928
0.0011	0.604908202
0.0015	0.619293433

0.002	0.690928468
0.005	0.693148735
0.01	1.27944494

Table 7: Logistic Regression for Concatenation Operation

Eta	ERMS Testing
0.0001	0.495466913
0.001	0.478467385
0.005	0.444807743
0.006	0.439610268
0.007	0.435791676
0.008	0.433758142
0.009	0.433744092
0.01	0.435778517
0.05	0.570029247
0.1	0.9

Table 8: Logistic Regression for Subtraction Operation

Eta	E_RMS Testing
0.0001	0.497732498
0.0005	0.492231646
0.001	0.487059535
0.005	0.467170193
0.01	0.453239837
0.02	0.450740227
0.03	0.465816811
0.04	0.481423732
0.05	0.494285705
0.06	0.505193162

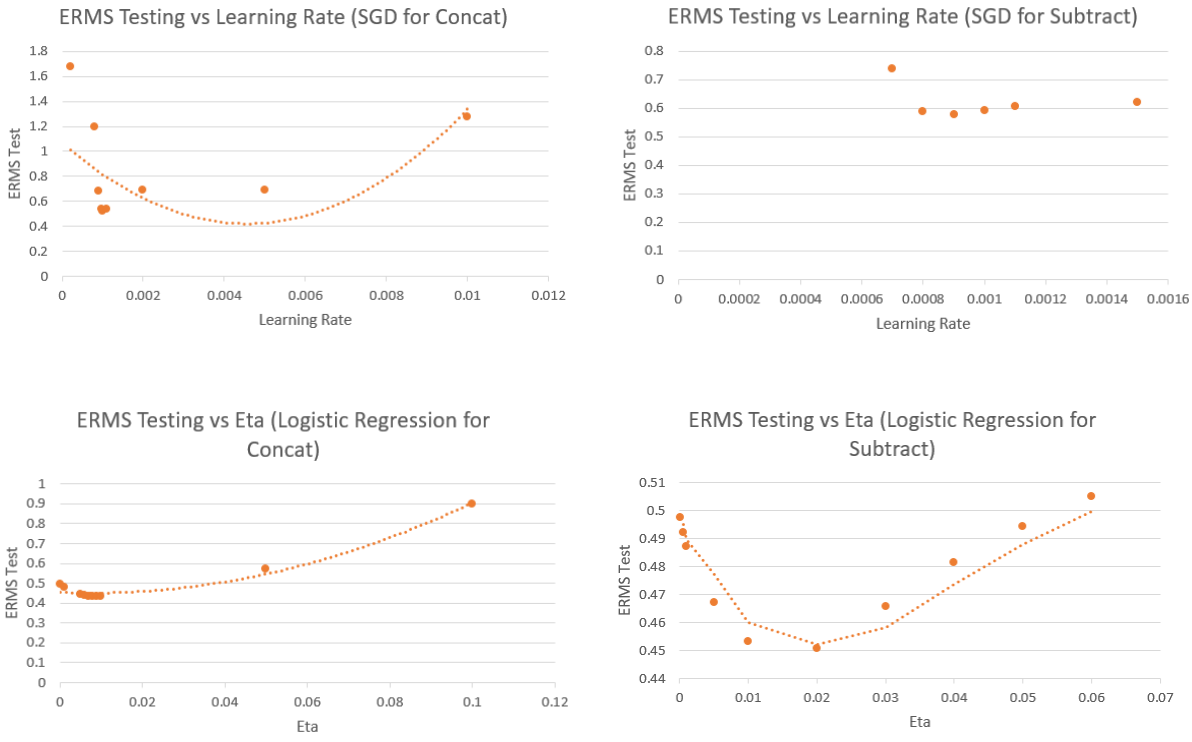

Fig 2

Other Observations

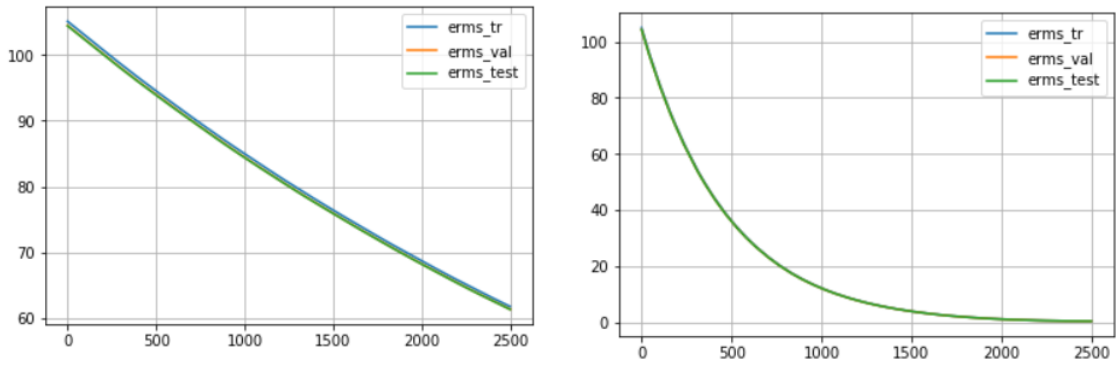

Fig3

4 Inference and Conclusion

From Table1-2 and Fig 1, it is evident that for human dataset, SGD for Concat performs better than for Subtract operation (ERMS in Concat is less than Subtract operation)

From Table 3-4 and Fig 1, it is evident that for human dataset, Logistic regression for subtract performs better than Concat operation (ERMS in Subtract is less than Concat operation)

From Table5-6 and Fig 2, it is evident that for GSC dataset, SGD for Concat performs better than for Subtract operation (ERMS in Concat is less than Subtract operation)

From Table 7-8 and Fig 2, it is evident that for human dataset, Logistic regression for concat performs better than subtract operation (ERMS in concat is less than subtract operation)

From Fig 3, it is clear that as we increase the value of learning rate in SGD, the model starts converging early.