CSE 574 - Project 2: CEDAR Letter dataset

Baskar Adyar Krishnan
Person 50291475
UBIT: BASKARAD

Abstract

The project requires to apply machine learning to solve the handwriting comparison task in forensics to help the Bureau and police departments to solve criminal cases dealing with evidence provided by handwritten documents such as wills and ransom notes. The task is to find similarity between the handwritten samples of the known and the questioned writer. We are going to use Linear Regression, Logistic regression and neural network model to identify if the the writer of both the images are same or not.

1 Dataset

We have a CEDAR dataset for this project, each instance in the CEDAR "AND" training data consists of set of input features for each hand-written "AND" sample. Image snippets of the word "AND" were extracted from each of the manuscript using transcript-mapping function of CEDAR-FOX.

The features are obtained from two different sources: There are two data sources

1. **Human Observed features**: Features entered by human document examiners manually. The Human Observed dataset shows only the cursive samples in the data set, where for each image the features are entered by the human document examiner. There are total of 18 features for a pair of handwritten "AND" sample (9 features for each sample).The entire dataset consists of 791 same writer pairs and 293,032 different writer pairs(rows)
2. **GSC features**: Features extracted using Gradient Structural Concavity (GSC) algorithm. Gradient Structural Concavity algorithm generates 512 sized feature vector for an input handwritten "AND"image. The entire data set consists of 71,531 same writer pairs and 762,557 different writer pairs(rows).

2 Preprocessing of Data

Firstly, the data from the CSV files are extracted for both Human Observed features and GSC features. We are provided with two files in each source namely, same_pair and different_pair which denotes the word "AND" is from the same writer or the different writer. If the writers are same then the target values is one. If the writers are different target values is zero.

In order to do form a dataset to feed to a Machine learning model we need to process the features of the input. We are combining the features by two methods:

1. Concatenation: The features of the two images to be compared are concatenated adjacent to each other to form a single row.
2. Subtraction: The features of the two images to be compared are subtracted with their respective feature. The feature1 of the first image is subtracted with the feature1 of the second image. Similarly it is done for all the features of the image.

2.1 Human Observed features:

Each image will have a total of 9 features associated with it. With Human Observed features we would get two set of datasets based on the above mentioned two methods:

1. Feature Concatenation - It will have 18 features
2. Feature subtraction - It will have 9 features

2.2 GSC features:

Each image will have a total of 512 features associated with it. With Human Observed features we would get two set of datasets based on the above mentioned two methods:

1. Feature Concatenation - It will have 1024 features
2. Feature subtraction - It will have 512 features

The raw input dataset and target values are divided into 3 parts for all the four datasets mentioned above:

1. Training Data and target which we are going to use for training (80 percent)
2. Validation Data and target: (10 percent) is used for validating the model during training to check how good the model predicts the unseen inputs.
3. Testing Data and target is used to check the accuracy of the model after training is over. (10 percent)

3 Training using Linear Regression

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available

3.1 Stochastic Gradient Descent Solution

Gradient Descent is the process of minimizing error function by descending down the gradients. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.The stochastic gradient descent algorithm first takes a random initial value $w^{(0)}$. Then it updates the value of w adding delta(w).

η is the learning rate, deciding how big each update step would be while descending the gradient and λ is the regularization term.

The target values can be represented by below equation:

$$w^T \phi(X) = w_1\phi(x) + w_2\phi(x)...w_M\phi(x)$$

Like the process we followed for closed form, once the weights for which the ERMS is low is determined using gradient descent algorithm, the training is complete. Using the obtained weights the model is tested using testing dataset and target.

Let us discuss the results we got for various datasets using Linear Regression:

3.2 Human Observed Dataset - Feature Concatenation

3.2.1 Parameters:

M = 10
Lambda = 1
Learning Rate = 0.01

3.2.2 Results:

E_rms Training = 0.49772
E_rms Validation = 0.49426
E_rms Testing = 0.49933
Accuracy Training = 55.52923
Accuracy Validation = 60.75949
Accuracy Testing = 53.50318

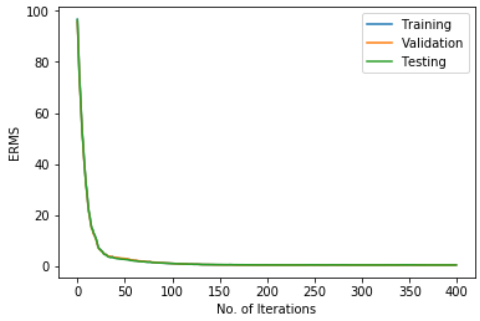

Figure 1: ERMS

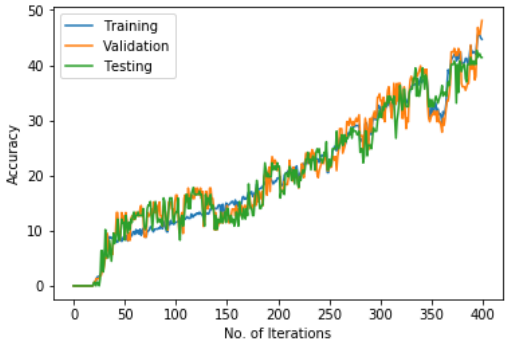

Figure 2: Accuracy

3.3 Human Observed Dataset - Feature Subtraction

3.3.1 Parameters:

M = 10
Lambda = 1
Learning Rate =0.01

3.3.2 Results:

E_rms Training = 0.4703
E_rms Validation = 0.44755
E_rms Testing = 0.48425
Accuracy Training = 63.82306
Accuracy Validation = 61.25316
Accuracy Testing = 63.96815

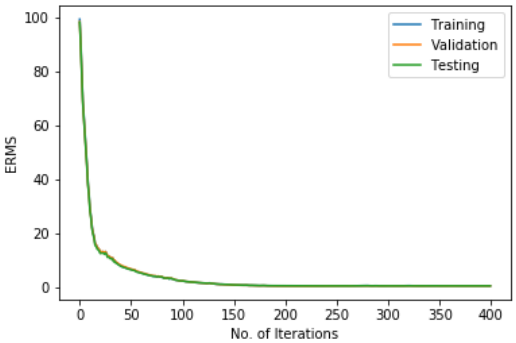

Figure 3: ERMS

Figure 4: Accuracy

3.4 GSC Dataset - feature Concatenation

The entire data set is taken to run the model as it would take more time and complexity only a sample of the dataset which is shuffled is taken to train the model. The dataset is shuffled in because we do not want our dataset to be skewed.

3.4.1 Parameters:

M = 10
Lambda = 1
Learning Rate =0.01

3.4.2 Results:

E_rms Training = 0.6721
E_rms Validation = 0.70169
E_rms Testing = 0.65188
Accuracy Training = 53.375
Accuracy Validation = 51.17057
Accuracy Testing = 55.51839

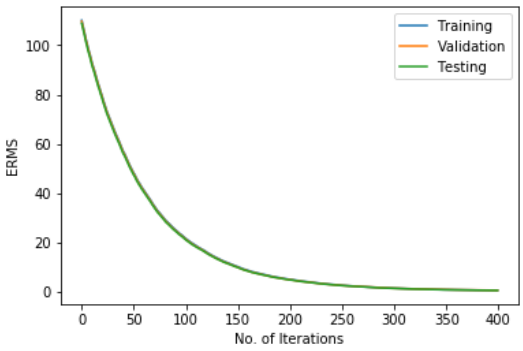

Figure 5: erms

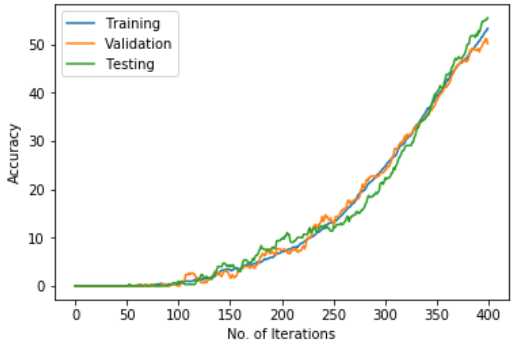

Figure 6: Accuracy

3.5 GSC Dataset - feature Subtraction

3.5.1 Parameters:

M = 10
Lambda = 1
Learning Rate =0.01

3.5.2 Results:

E_rms Training = 0.37726
E_rms Validation = 0.39172
E_rms Testing = 0.3975
Accuracy Training = 80.75
Accuracy Validation = 79.26421
Accuracy Testing = 76.92308

4 Logistic regression

Logistic Regression is used when the output has many classes. The logistic regression hypothesis is defined as

$$h_w(x) = g(w^t x)$$

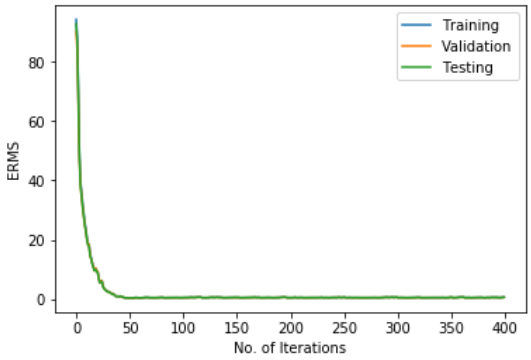

Figure 7: ERMS

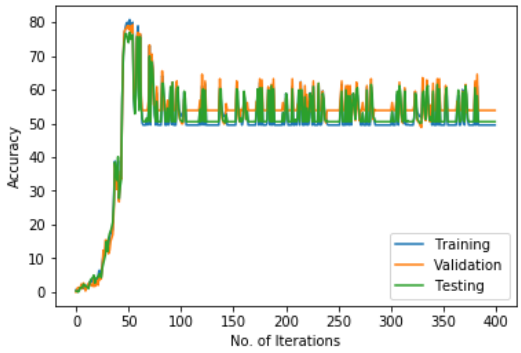

Figure 8: Accuracy

where g is the sigmoid function given as

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

This is used to update the weights as follows

$$w = w - \lambda(\delta w)$$

4.1 Human Observed Dataset - Feature Concatenation

Acc Training = 57.10900473933649
Acc Validation = 58.860759493670884
Acc Testing = 52.86624203821656

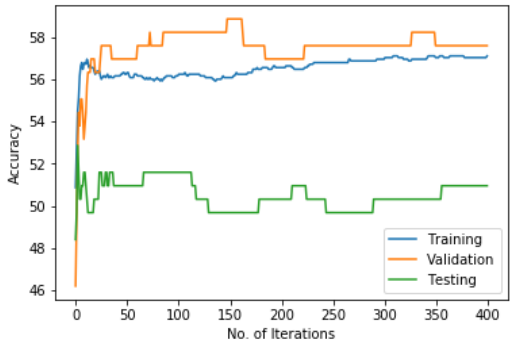

Figure 9: Feature Concatenation - Accuracy

4.2 Human Observed Dataset - Feature Subtraction

Acc Training = 98.9210
Acc Validation = 99.1484
Acc Testing = 99.6445

4.3 GSC Dataset - feature Concatenation

Acc Training = 58.25
Acc Validation = 54.180602006688964
Acc Testing = 54.180602006688964

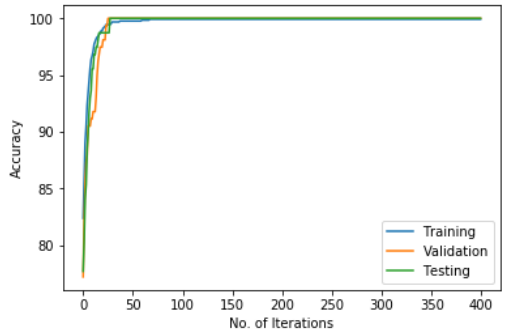

Figure 10: Feature Subtraction - Accuracy

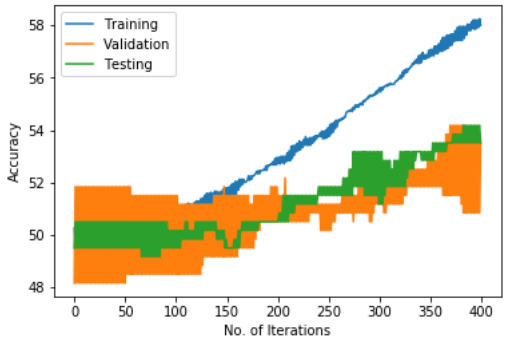

Figure 11: Feature Concatenation - Accuracy

4.4 GSC Dataset - feature Subtraction

Acc Training = 100.0 Acc Validation = 99.33110367892976 Acc Testing = 99.66555183946488

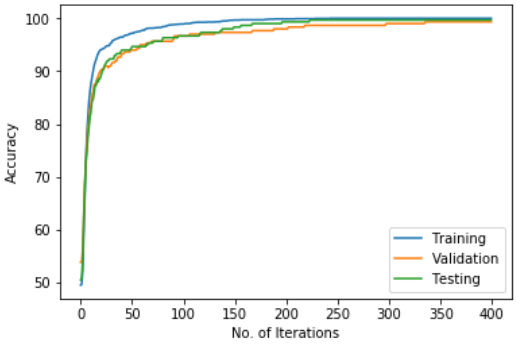

Figure 12: Feature Subtraction - Accuracy

5 Neural Network

5.1 Human Observed Dataset with feature Subtraction

Hidden layers = 128. Drop out = 0.1. Testing Accuracy = 86.78848493. This is better than linear reggression and logistic regression

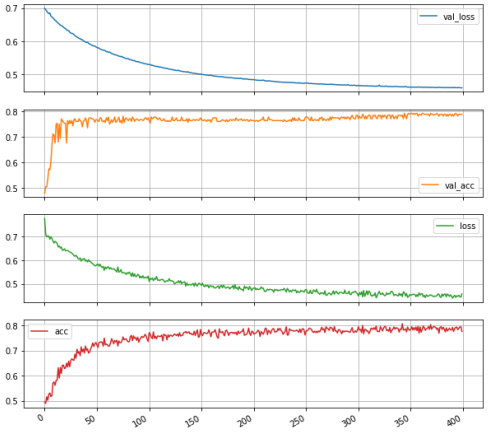

Figure 13:

5.2 Human Observed Dataset with feature Concatenation

Testing Accuracy = 91.0828025477707

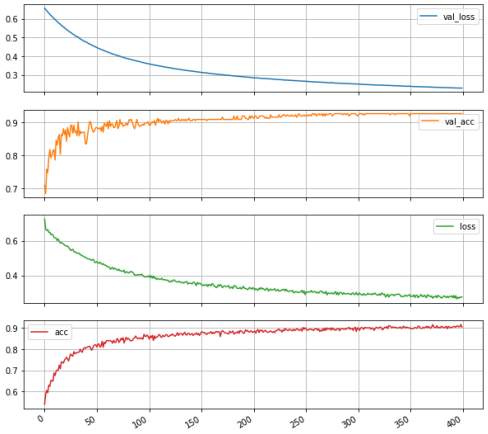

Figure 14:

5.3 GSC Observed Dataset with feature Subtraction

Testing Accuracy = 92.65871921922

5.4 GSC Dataset with feature Concatenation

Testing Accuracy= 97.65184886

6 Inference

1. Which Dataset is better - Subtracted or Concatenated? From all the above results we can conclude that subtraction of features is better than concatenated dataset for both Human observed and GSC features irrespective of the model.
2. Which Model is better? - From all the above graphs its can be said that out of linear regression , logistic regression and neural network model , neural network model is the best for this data set.

3. Which type of dataset is better - Human observed or GSC features? From the accuracy and ERMS results we got from the above observations, we can say that GSC features gives the better results. This is because the there are more number of features in GSC than the Human observed features. As there are more features, models gets to learn more and gives better prediction.
4. Human observed data set is of lesser size than GSC data set Due to this the performance of the model is better for GSC. Hence can me told that larger the data samples better is the learning and better performance

7 Conclusion

The identification of weather the writer of two "and" images are same or not is done using machine learning models - linear regression , logistic regression and neural networks. Differ data sources - Human observed and GSC features are analyzed and compared for the given dataset.