# Project 2: Handwriting Comparison

**Rohan Gupta**
Department of Computer Science
University at Buffalo, SUNY
rgupta24@buffalo.edu
50290793

## 1 Overview

In this project we will apply machine learning to solve the handwriting comparison task in forensics. Our task is to find the similarity between the handwriting samples of the known and the questioned writers by using three approaches: **linear regression**, **logistic regression** and **neural networks**.

For this project, we have used the **CEDAR "AND" training dataset** which contains set of input features for each handwritten "AND" sample. These features have been obtained from two sources:
1. **Human Observed Features** - These features have been entered by human document examiners manually. There are 9 features for each sample.
2. **GSC** - These features have been extracted using Gradient Structural Concavity algorithm. There are 512 features for each sample.

## 2 Solutions (Model Training Approaches)

### 2.1 Linear Regression

Linear Regression is a statistical method to modelling the relationship between two or more continuous (quantitative) variables: one variable y is a response (dependent) and the other variables x are explanatory variables (independent). When there are more than one explanatory variables, the process is called multiple linear regression. Linear Regression is a supervised learning task. It is extensively used for prediction and error reduction.

In this solution, we treat the handwriting comparison problem as a Linear Regression problem where we map an input variable *x* to a real-valued scalar target *y(x, w)*.

The **Genesis Equation for Linear Regression** has the following form:

$$\hat{y} = w^T \phi(x) \tag{1}$$

where $w = (w_0, w_1, w_2 ... w_{m-1})$ is the weight vector which will be computed using the training samples and $\phi = (\phi_0, \phi_1, \phi_2 ... \phi_{m-1})^T$ is a vector of M basis functions.

The **Loss/Error function for Linear Regression is Least Squared Error** which has the following form:

$$L = \frac{1}{2}(y - \hat{y})^2 \tag{2}$$

### 2.2 Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable $y$ is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The **Genesis Equation for Logistic Regression** has the following form:

$$\hat{y} = \sigma(w^T x) = \sigma(w_1 x_1 + w_2 x_2 + ... + w_n x_n + b) \tag{3}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

where $w = (w_0, w_1, w_2 ... w_{n-1})$ is the weight vector which will be computed using the training samples, $x = (x_0, x_1, x_2 ... x_{n-1})^T$ is the feature vector and $b$ is the bias term.

The **Loss/Error function for Logistic Regression is Cross Entropy** which has the following form:

$$L = -(y * log(\hat{y}) + (1 - y) * log(1 - \hat{y})) \tag{5}$$

### 2.3 Neural Network

Neural Network is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, processes information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve a specific problem. Neural Networks, like people, learn by examples. In a neural network, we have a series of one or more layers with their associated weights.

The **Genesis Equation for Neural Network** has the following form:

$$\hat{y} = f(w, x) \tag{6}$$

The **Loss/Error function we used for Neural Network is Sigmoid Cross Entropy with Logits**. Let $x$ = logits, $z$ = labels. The logistic loss $L$ is

$$L = z * -log(sigmoid(x)) + (1 - z) * -log(1 - sigmoid(x)) \tag{7}$$

## 3 Evaluation

Finally, we evaluate all our solutions on the test dataset using Root Mean Square (RMS) error, as defined below

$$E_{RMS} = \sqrt{2E(w^*)/N_v}$$

where $w*$ is the solution and $N_V$ is the size of the test dataset.

## 4 Results

### 4.1 Results for Linear Regression :

Human Observed Dataset with feature concatenation

| | | | | Training | Validation | Testing |
|---|---|---|---|---|---|---|
| | M | $\lambda$ | Epochs | E RMS | E RMS | E RMS |
| 1. | 7 | 0.01 | 400 | 0.4990 | 0.4961 | 0.5044 |
| 2. | 7 | 0.01 | 500 | 0.4996 | 0.4996 | 0.4983 |

Human Observed Dataset with feature subtraction

| | | | | Training | Validation | Testing |
|---|---|---|---|---|---|---|
| | M | $\lambda$ | Epochs | E RMS | E RMS | E RMS |
| 1. | 3 | 0.01 | 400 | 0.5102 | 0.5031 | 0.5098 |
| 2. | 3 | 0.01 | 500 | 0.5032 | 0.5090 | 0.5052 |

GSC Dataset with feature concatenation

| | | | | Training | Validation | Testing |
|---|---|---|---|---|---|---|
| | M | $\lambda$ | Epochs | E RMS | E RMS | E RMS |
| 1. | 100 | 0.01 | 200 | 0.4124 | 0.4143 | 0.4121 |

GSC Dataset with Feature subtraction

| | | | | Training | Validation | Testing |
|---|---|---|---|---|---|---|
| | M | $\lambda$ | Epochs | E RMS | E RMS | E RMS |
| 1. | 50 | 0.01 | 200 | 0.4021 | 0.4261 | 0.3914 |

### 4.2 Results for Logistic Regression:

Human Observed Dataset with feature concatenation

| | | | Training | Validation | Testing |
|---|---|---|---|---|---|
| | Epoch | $\eta$ | Accuracy | Accuracy | Accuracy |
| 1. | 400 | 0.01 | 65.78 | 63.88 | 64.98 |
| 2. | 400 | 0.05 | 60.96 | 55.89 | 55.78 |

Human Observed Dataset with feature subtraction

| | | | Training | Validation | Testing |
|---|---|---|---|---|---|
| | Epoch | $\eta$ | Accuracy | Accuracy | Accuracy |
| 1. | 400 | 0.01 | 69.98 | 65.67 | 61.89 |
| 2. | 400 | 0.05 | 56.78 | 51.12 | 53.78 |

GSC Dataset with feature concatenation

| | | | Training | Validation | Testing |
|---|---|---|---|---|---|
| | Epoch | $\eta$ | Accuracy | Accuracy | Accuracy |
| 1. | 400 | 0.01 | 94.78 | 85.89 | 87.89 |
| 2. | 400 | 0.05 | 85.89 | 79.56 | 80.98 |

GSC Dataset with feature subtraction

| | | | Training | Validation | Testing |
|---|---|---|---|---|---|
| | Epoch | $\eta$ | Accuracy | Accuracy | Accuracy |
| 1. | 400 | 0.01 | 95.78 | 88.56 | 91.56 |
| 2. | 400 | 0.05 | 90.67 | 85.78 | 88.56 |

### 4.3 Results for Neural Network :

Human Observed Dataset with feature concatenation

| | | | | Number of | Number of | Training | Validation | Testing |
|---|---|---|---|---|---|---|---|---|
| | Epochs | $\eta$ | Batch Size | Hidden Layers | Neurons in Layer | Accuracy | Accuracy | Accuracy |
| 1. | 4000 | 0.01 | 100 | 1 | 256 | 89.49 | 50.50 | 53.16 |
| 2. | 1500 | 0.05 | 100 | 1 | 256 | 89.41 | 58.86 | 51.27 |
| 3. | 2000 | 0.02 | 100 | 2 | 256, 128 | 87.59 | 55.69 | 48.10 |
| 4. | 2000 | 0.05 | 100 | 2 | 256, 128 | 97.47 | 55.69 | 53.8 |

Human Observed Dataset with feature subtraction

| | | | | Number of | Number of | Training | Validation | Testing |
|---|---|---|---|---|---|---|---|---|
| | Epochs | $\eta$ | Batch Size | Hidden Layers | Neurons in Layer | Accuracy | Accuracy | Accuracy |
| 1. | 7500 | 0.01 | 100 | 1 | 256 | 76.14 | 57.59 | 49.7 |
| 2. | 2000 | 0.05 | 100 | 1 | 256 | 74.88 | 56.96 | 55.06 |
| 3. | 3000 | 0.05 | 100 | 2 | 256, 128 | 85.86 | 57.58 | 57.59 |
| 4. | 5000 | 0.05 | 100 | 2 | 256, 128 | 95.65 | 47.46 | 57.51 |

GSC Dataset with feature concatenation

| | | | | Number of | Number of | Training | Validation | Testing |
|---|---|---|---|---|---|---|---|---|
| | Epochs | $\eta$ | Batch Size | Hidden Layers | Neurons in Layer | Accuracy | Accuracy | Accuracy |
| 1. | 300 | 0.01 | 100 | 1 | 256 | 99.68 | 92.91 | 93.02 |
| 2. | 800 | 0.005 | 100 | 1 | 256 | 99.99 | 93.10 | 93.01 |

GSC Dataset with feature subtraction

| | | | | Number of | Number of | Training | Validation | Testing |
|---|---|---|---|---|---|---|---|---|
| | Epochs | $\eta$ | Batch Size | Hidden Layers | Neurons in Layer | Accuracy | Accuracy | Accuracy |
| 1. | 500 | 0.01 | 100 | 1 | 256 | 99.99 | 84.25 | 83.87 |
| 2. | 800 | 0.01 | 256 | 1 | 256 | 98.73 | 83.75 | 84.60 |

## References

[1] https://en.wikipedia.org/wiki/Linear_regression

[2] https://www.statisticssolutions.com/what-is-logistic-regression/

[3] https://www.doc.ic.ac.uk/ nd/surprise_96/journal/vol4/cs11/report.html