CSE474/574: Introduction to Machine Learning (Fall 2018) UBIT: sassikes
Project 2: Handwriting Comparison Person no: 50291558

Introduction:

The primary objective of this project is to solve the handwriting comparison task using Linear regression and Logistic Regression.
The tasks that required to be done are

1. Train a regression model on Human observed and GSC features of the handwriting using gradient decent of Linear Regression.
2. Train a regression model on Human observed and GSC features of the handwriting using gradient decent of Logistic Regression.

The features are obtained from two different sources:

- Human Observed features: Features entered by human document examiners manually
- GSC features: Features extracted using Gradient Structural Concavity (GSC) algorithm.

Experimental Setup:

The following procedure has to be followed step by step,

1. ***Extract features values and Image Ids from the data***: Process the original CSV data _les into a Numpy matrix or Pandas Dataframe. Process the csv _les to derive four datasets:
   (a) Human Observed Dataset with feature concatenation
   (b) Human Observed Dataset with feature subtraction
   (c) GSC Dataset with feature concatenation
   (d) GSC Dataset with feature subtraction

2. ***Data Partitioning***: Partition your data into training, validation and testing data.

3. ***Train using Linear Regression***: Use Gradient Descent for linear regression to train the model using a group of hyperparameters on each of the 4 input datasets.

4. ***Train using Logistic Regression***: Use Gradient Descent for logistic regression to train the model using a group of hyperparameters on each of the 4 input datasets.

5. ***Tune hyper-parameters***: Validate the regression performance of your model on the validation set. Change your hyper-parameters. Try to find what values those hyper-parameters should take so as to give better performance on the validation set.

6. ***Test your machine learning scheme on the testing set***: After finishing all the above steps, fix your hyper-parameters and model parameter and test your models performance on the testing set. This shows the ultimate effectiveness of your models generalization power gained by learning.

7. Report your results

50291558
UBIT: sassikes

Linear Regression:

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

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).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

Linear Regression is made with an assumption that there's a linear relationship between X and Y.

Form of Linear Regression

Mathematically, we can write a linear relationship as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Logistic Regression:

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

Model
Output = 0 or 1
Hypothesis => Z = WX + B
hΘ(x) = sigmoid (Z)

50291558
UBIT: sassikes

Cost Function:

$$\text{Cost}(h_\Theta(x), Y(\text{actual})) = -\log(h_\Theta(x)) \text{ if } y=1$$

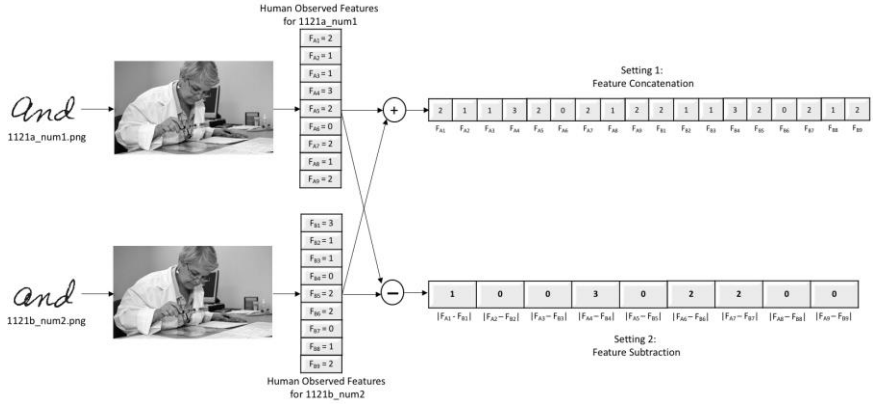
$$-\log(1-h_\Theta(x)) \text{ if } y=0$$

Feature Extraction:

Human observed data set:

You will have to build your dataset using HumanObserved-Features-Data.csv, same pairs.csv and diffn pairs.csv given.

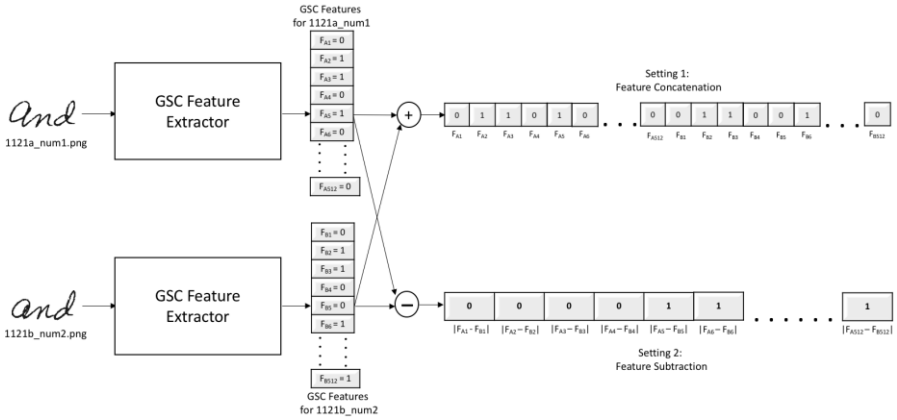
Setting 1: Feature Concatenation [18 features]
Setting 2: Feature subtraction [9 features]



GSC Dataset:

You will have to build your dataset using GSC-Features-Data.csv, same pairs.csv and diffn pairs.csv.

Setting 1: Feature Concatenation [1024 features]
Setting 2: Feature subtraction [512 features]

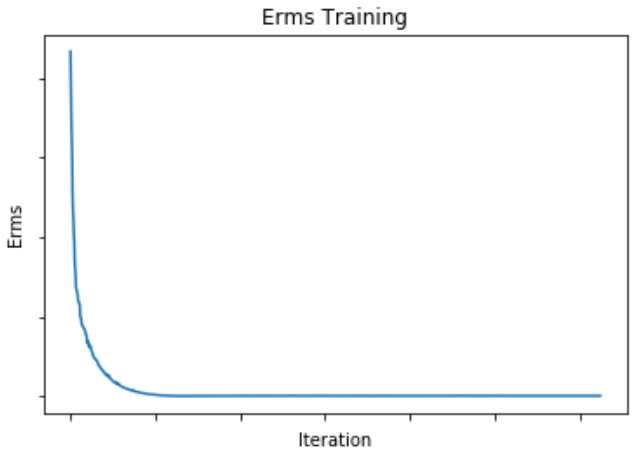


50291558
UBIT: sassikes

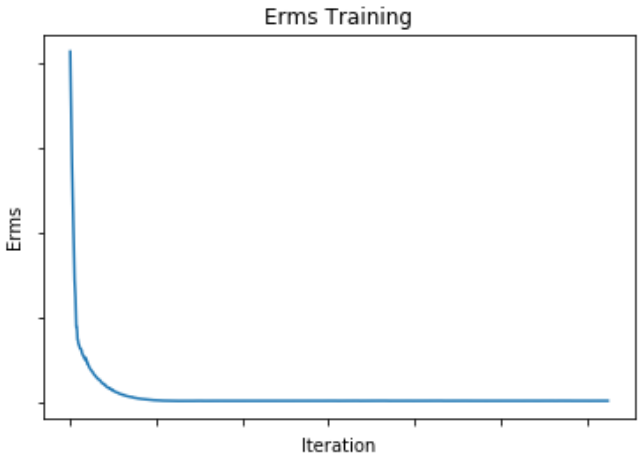
Results:

Linear Regression:

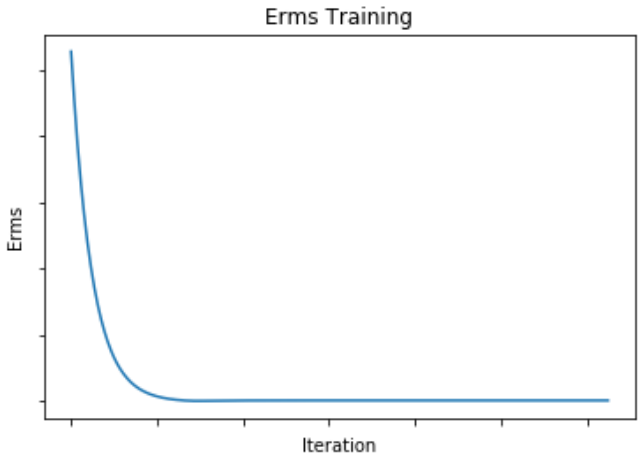
```
----------Linear Regression--------------------
----------Gradient Descent Solution human_concate_X.csv  and  human_concate_t
.csv --------------------
eta= 0.01
E_rms Training   = 0.37593
E_rms Validation = 0.37204
E_rms Testing    = 0.39347
```



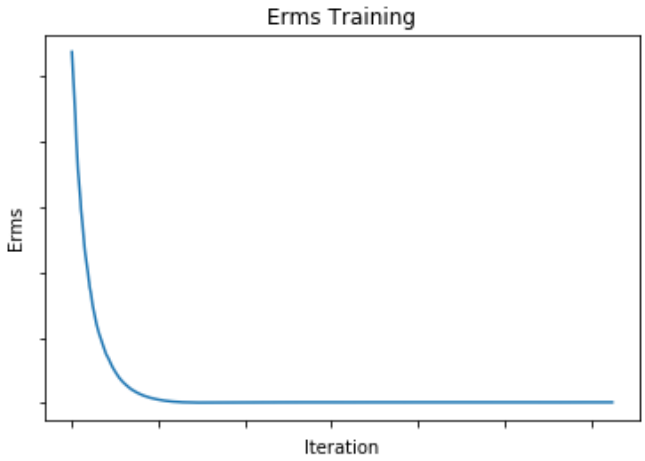

```
----------Linear Regression--------------------
----------Gradient Descent Solution human_sub_X.csv  and  human_sub_t.csv ---
-----------------

eta= 0.01
E_rms Training   = 0.43803
E_rms Validation = 0.43695
E_rms Testing    = 0.43054
```



```
----------Linear Regression--------------------
----------Gradient Descent Solution gsc_concate_X.csv  and  gsc_concate_t.csv
--------------------

eta= 0.01
E_rms Training   = 0.33971
E_rms Validation = 0.3395
E_rms Testing    = 0.33681
```

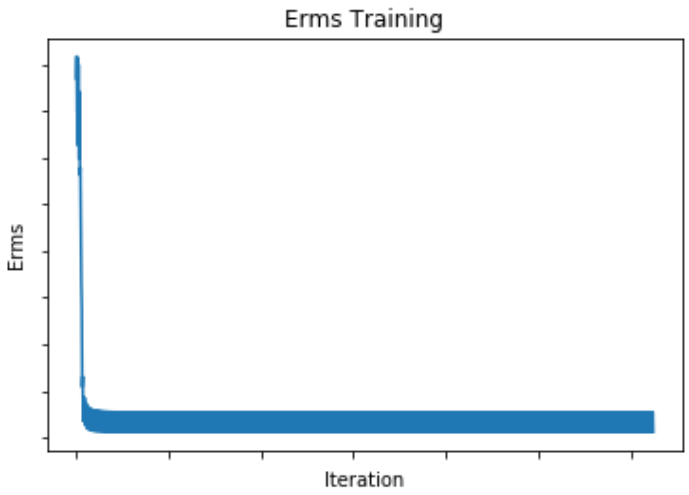



```
----------Linear Regression--------------------
----------Gradient Descent Solution gsc_sub_X.csv  and  gsc_sub_t.csv -------
-------------

eta= 0.01
E_rms Training   = 0.40016
E_rms Validation = 0.37767
E_rms Testing    = 0.38925
```

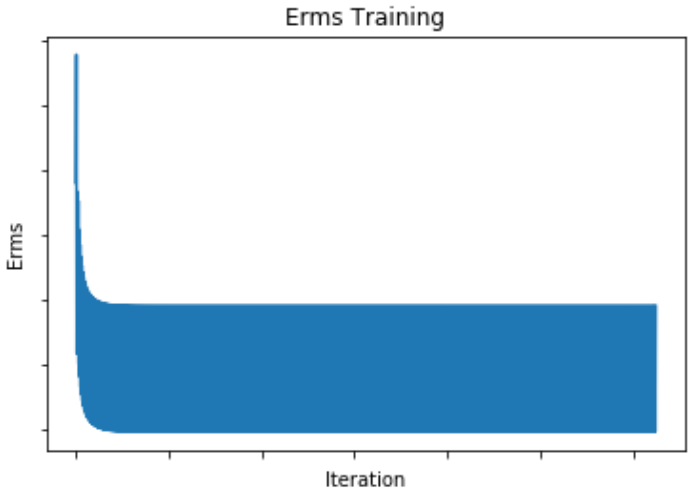


Logistic Regression:

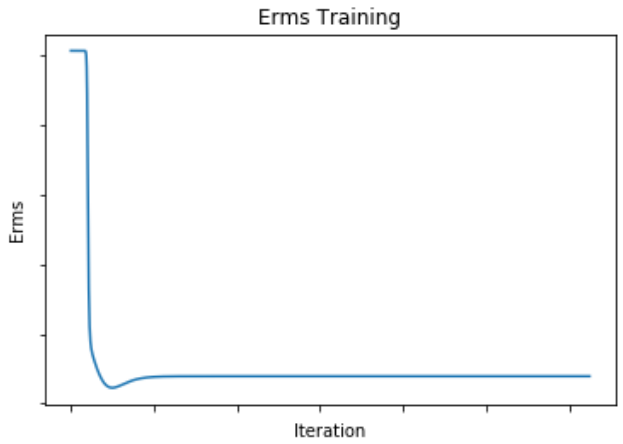
```
----------Logistic Regression--------------------
----------Gradient Descent Solution human_concate_X.csv  and  human_concate_t
.csv --------------------

eta= 0.01
E_rms Training   = 0.50309
E_rms Validation = 0.47925
E_rms Testing    = 0.51868
```



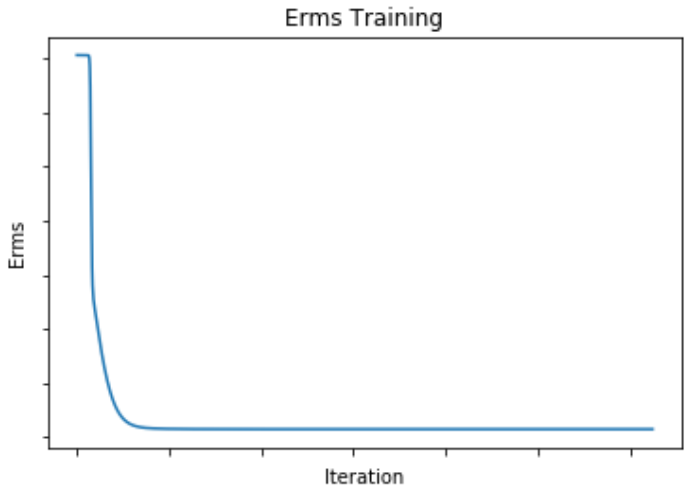

```
----------Logistic Regression--------------------
----------Gradient Descent Solution human_sub_X.csv  and  human_sub_t.csv ---
-----------------

eta= 0.01
E_rms Training   = 0.52978
E_rms Validation = 0.52724
E_rms Testing    = 0.50349
```



```
----------Logistic Regression--------------------
----------Gradient Descent Solution gsc_concate_X.csv  and  gsc_concate_t.csv
--------------------

eta= 0.01
E_rms Training   = 0.22295
E_rms Validation = 0.22944
E_rms Testing    = 0.22732
```

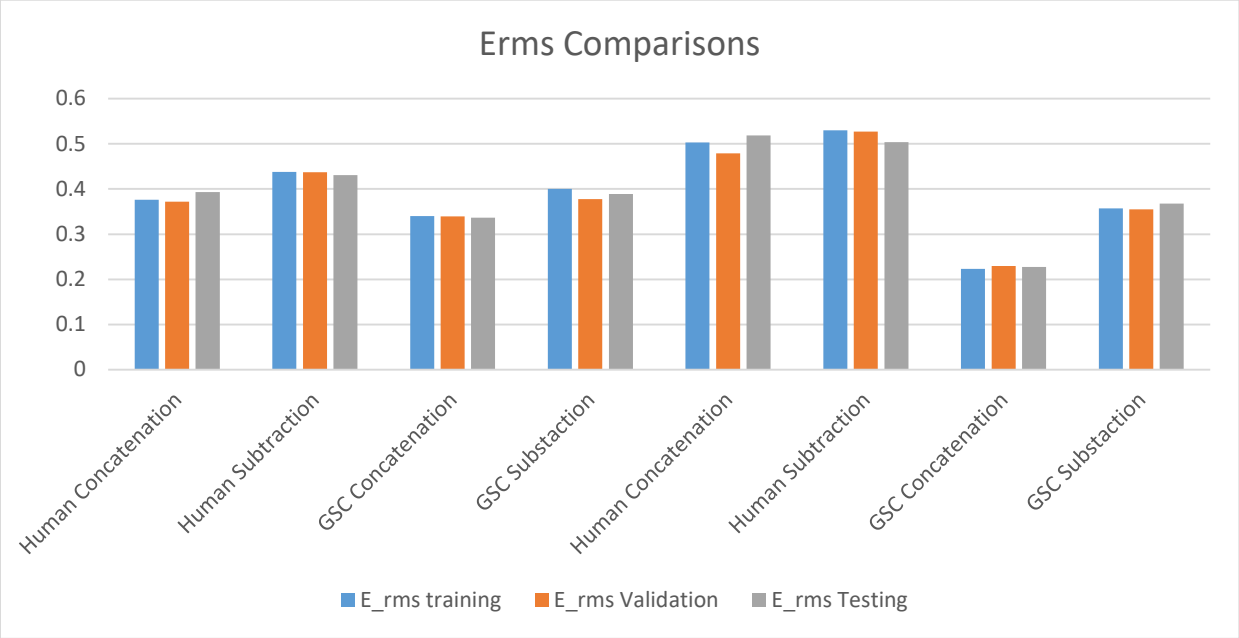



```
----------Logistic Regression--------------------
----------Gradient Descent Solution gsc_sub_X.csv  and  gsc_sub_t.csv -------
-------------

eta= 0.01
E_rms Training   = 0.35731
E_rms Validation = 0.35506
E_rms Testing    = 0.36769
```



50291558
UBIT: sassikes

Comparison:



References:

https://www.medcalc.org/
https://www.statisticssolutions.com
https://machinelearningmastery.com
https://towardsdatascience.com
https://en.wikipedia.org/