Introduction to Machine Learning

October 21, 2018

1 Overview

You have been hired by the FBI to develop predictive models for detection of crime, your task is to help the Bureau and police departments to solve criminal cases dealing with evidence provided by handwritten documents such as wills and ransom notes. You are assigned to a forensic project by the FBI. The project requires you to apply machine learning to solve the handwriting comparison task in forensics. We formulate this as a problem of linear regression where we map a set of input features x to a real-valued scalar target $y(x,w)$.

Your task is to find similarity between the handwritten samples of the known and the questioned writer by using linear regression.

Each instance in the CEDAR *"AND"* training data consists of set of input features for each handwritten *"AND"* sample. The features are obtained from two different sources:

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

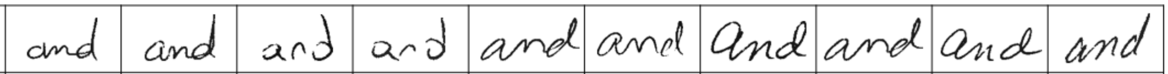
The target values are scalars that can take two values {1:same writer, 0:different writers}. Although the training target values are discrete we use linear regression to obtain real values which is more useful for finding similarity (avoids collision into only two possible values).

2 Dataset

2.1 Source of Dataset

Our dataset uses "AND" images samples extracted from CEDAR Letter dataset. Image snippets of the word "AND" were extracted from each of the manuscript using transcript-mapping function of CEDAR-FOX. Figure 1. shows examples of the "AND" image fragments.

Figure 1: Example of Dataset



Sample ID [XXXXy_numZ]	0001a_num1	0001a_num2	0002a_num1	0002a_num2	0005b_num1	0005b_num2	1121a_num1	1121a_num2	1160a_num1	1160a_num2
Writer Number [XXXX]	Writer 0001	Writer 0001	Writer 0002	Writer 0002	Writer 0005	Writer 0005	Writer 1121	Writer 1121	Writer 1160	Writer 1160
Page Number [y]	Page 1	Page 1	Page 1	Page 1	Page 2	Page 2	Page 1	Page 1	Page 1	Page 1
Sample Number [Z]	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2

2.2 Types of Datasets:

Based on feature extraction process, we have provided two datasets:

2.2.1 Human Observed Dataset

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. The description of each of the Human Observed features are given in Table 1. There are total of 18 features for a pair of handwritten "AND" sample (9 features for each sample). The dataset is named as *"HumanObserved-Features-Data"*. Dataset is available on UBLearns under the Assignments section in *"HumanObserved-Dataset.zip"*. The entire dataset consists of 791 same writer pairs and 293,032 different writer pairs(rows). You will have to build your dataset using HumanObserved-Features-Data.csv, same_pairs.csv and diffn_pairs.csv. Figure 2. shows two sample rows derived using the three csv files for human observed dataset:

Figure 2: Human Observed Dataset Example

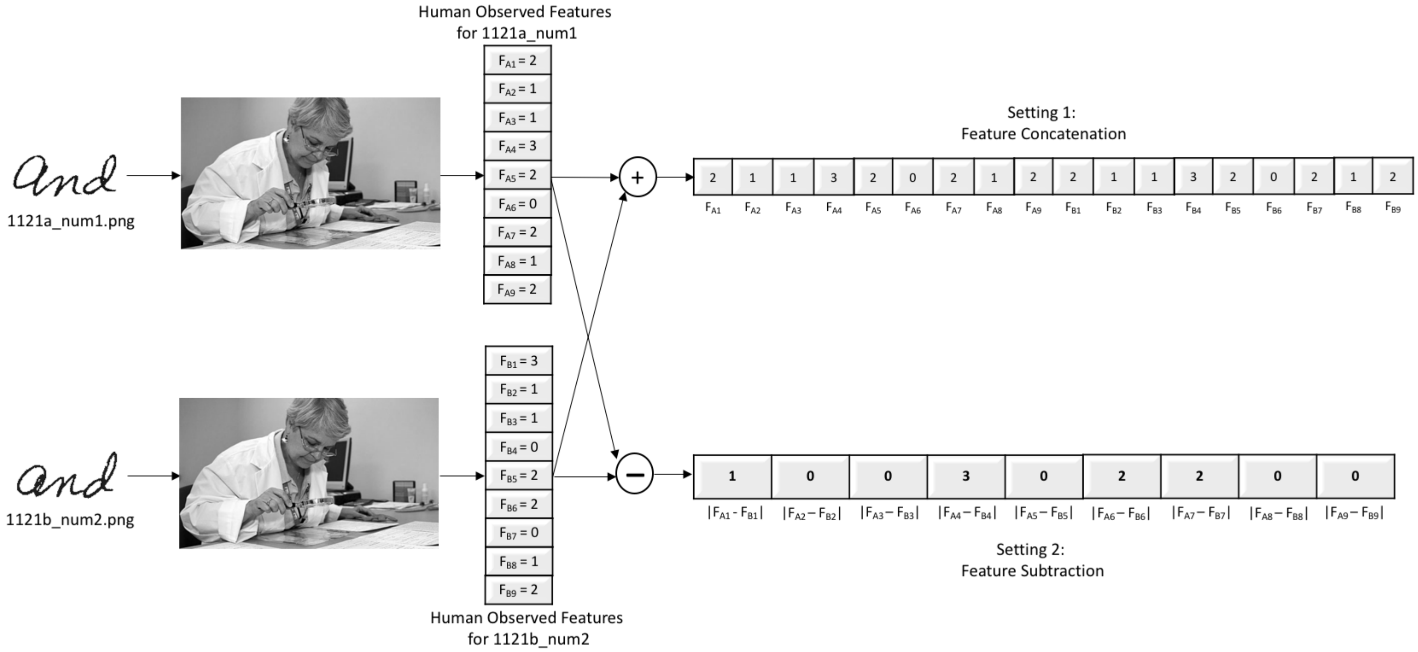
img_id_A	img_id_B	f_{A1}	f_{A2}	f_{A3}	f_{A4}	f_{A5}	f_{A6}	f_{A7}	f_{A8}	f_{A9}	f_{B1}	f_{B2}	f_{B3}	f_{B4}	f_{B5}	f_{B6}	f_{B7}	f_{B8}	f_{B9}	t
1121a_num1	1121b_num2	2	1	1	3	2	2	0	1	2	2	1	1	0	2	2	0	3	2	1
1121a_num1	1386b_num1	2	1	1	3	2	2	0	1	2	3	1	1	0	2	2	0	1	2	0

Table 1: Feature Description for Human Observed Dataset

Initial stroke of formation of *a* (x_1)	**Formation of staff of *a*** (x_2)	**Number of arches of *n*** (x_3)	**Shape of arches of *n*** (x_4)	**Location of mid-point of *n*** (x_5)	**Formation of staff of *d*** (x_6)	**Formation of initial stroke of *d*** (x_7)	**Formation of terminal stroke of *d*** (x_8)	**Symbol in place of the word *and*** (x_9)
Right of staff (0)	Tented (0)	One (0)	Pointed (0)	Above baseline (0)	Tented (0)	Overhand (0)	Curved up (0)	Formation (0)
Left of staff (1)	Retraced (1)	Two (1)	Rounded (1)	Below baseline (1)	Retraced (1)	Underhand (1)	Straight across (1)	Symbol (1)
Center of staff (2)	Looped (2)	No fixed pattern (2)	Retraced (2)	At baseline (2)	Looped (2)	Straight across (2)	Curved down (2)	None (2)
No fixed pattern (3)	No staff (3)		Combination (3)	No fixed pattern (3)	No fixed pattern (3)	No fixed pattern (3)	No obvious ending stroke (3)	
	No fixed pattern (4)		No fixed pattern (4)				No fixed pattern (4)	

Figure 3. describes the two settings under which you need to perform linear regression
Setting 1: Feature Concatenation [18 features]
Setting 2: Feature subtraction [9 features]

Figure 3: Feature Extraction for Human Observed Dataset



2.2.2 GSC Dataset using Feature Engineering

Gradient Structural Concavity algorithm generates 512 sized feature vector for an input handwritten "AND"image. The dataset is named as *"GSC-Features-Data"*. Dataset is available on UBLearns under the Assignments section in *"GSC-Dataset.zip"*. The entire dataset consists of 71,531 same writer pairs and 762,557 different writer pairs(rows). You will have to build your dataset using GSC-Features-Data.csv, same_pairs.csv and diffn_pairs.csv. Figure 4. shows two sample rows derived using the three csv files for GSC dataset:

Figure 4: GSC Dataset Example

img_id_A	img_id_B	f_{A1}	f_{A2}	f_{A3}	f_{A4}	f_{A5}	f_{A6}	. . .	f_{A512}	f_{B1}	f_{B2}	f_{B3}	f_{B4}	f_{B5}	f_{B6}	. . .	f_{B512}	t
1121a_num1	1121b_num2	0	1	1	0	1	0	. . .	0	0	1	1	0	0	1	. . .	1	1
1121a_num1	1386b_num1	0	1	1	0	1	0	. . .	0	1	1	1	0	1	0	. . .	0	0

Figure 3. describes the two settings under which you need to perform linear regression
Setting 1: Feature Concatenation [1024 features]
Setting 2: Feature subtraction [512 features]

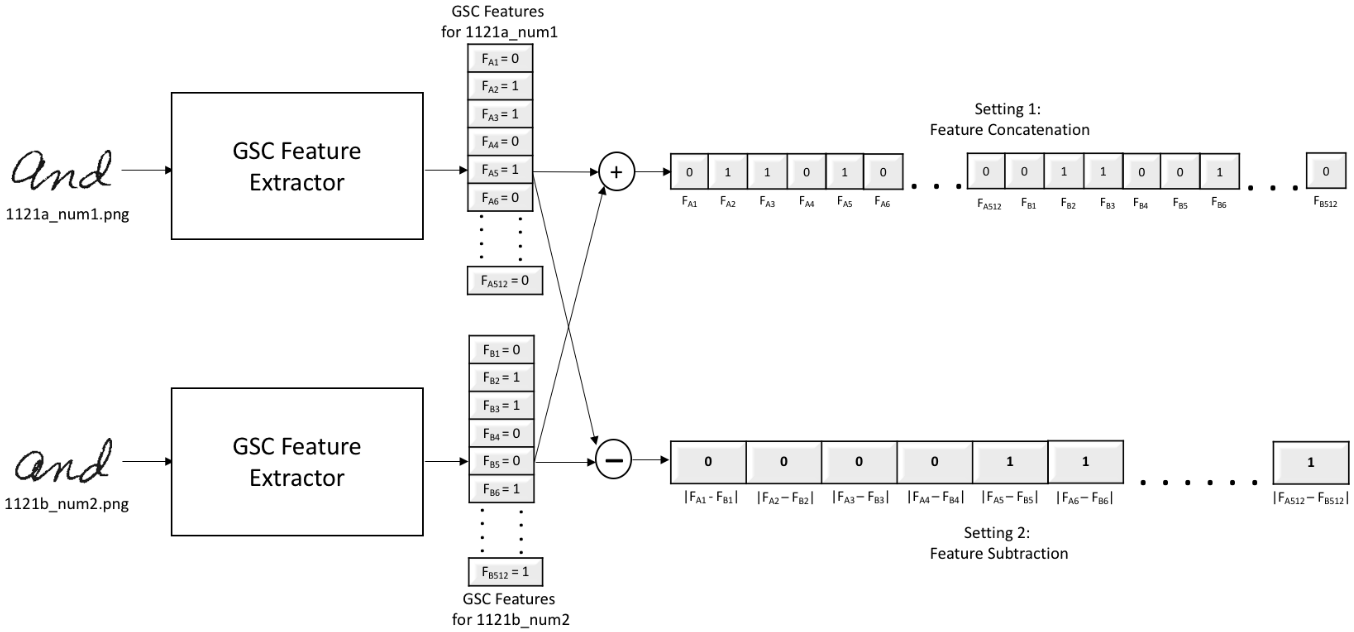
3 Plan of Work

1. **Extract features values and Image Ids from the data:** Process the original CSV data files into a Numpy matrix or Pandas Dataframe. Process the csv files to derive four datasets:

   (a) Human Observed Dataset with feature concatentation

   (b) Human Observed Dataset with feature subtraction

Figure 5: Feature Extraction For GSC Dataset



(c) GSC Dataset with feature concatentation

(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

4 Evaluation

Evaluate your solution on a test set using Accuracy and Root Mean Square (RMS) error. E_{RMS} defined as

$$E_{RMS} = \sqrt{2E(w^*)/N_V} \quad (1)$$

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