CSE 574: Introduction to Machine Learning (Fall 2018)
Instructor: Sargur N. Srihari

Project 2: Predictive Models for Detection of Crime

November 1, 2018

Report By:
Siddheswar Chandrasekhar

Objective

The project statement is 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).

The 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 (Center of Excellence for Document Analysis) "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}.

Linear Regression

In regression problems we try to predict continuous valued output.

Our linear regression function *y(x, w)* has the form:

$$y(x, w) = w^T \phi(x)$$

where,

$w = (w_0, w_1, w_2, \ldots\ldots, w_{M-1})$ is a weight vector to be learnt from training samples
$\phi = (\phi_0, \phi_1, \phi_2, \ldots\ldots, \phi_{M-1})^T$ is a vector of M basis functions.

Each basis function $\phi_j(x)$ converts the input vector x into a scalar value.

In this project, we use the Gaussian Radial Basis Fumction which has the form:

$$\phi_j(x) = exp\ [\ -\ ½\ (x - \mu_j)^T\ \Sigma_j^{-1}\ (x - \mu_j)\ ]$$

where,

μ_j is the center of the basis function, and
Σ_j decides how broadly the basis function spreads

Preparing the Datasets

The Human Observed Features Dataset has 1025 samples and each sample has 9 associated features. We follow two approaches to give pairs of samples to out model:

1. Concat Features: here we take two samples, concat their features together (hence we now have 18 features) and input it to our model.

2. Subtract Features: here we take two samples, subtract their individual respective features (hence we now have 9 features) and input it to our model.

We follow a similar approach for our GSC Dataset where we have 14,000 (approx.) samples and each sample has 512 associated features. Hence our concat dataset for GSC has 1024 features and the subtract dataset has 512 features.

We hence have four datasets:

1. Human Observed dataset with concatenated features
2. Human Observed dataset with subtracted features
3. GSC dataset with concatenated features
4. GSC dataset with subtracted features

Linear Regression Observations

Observation 1

Number of Basis Functions (M)	Regularization Term (λ)	Learning Rate
15	2	0.08

Human Observed Dataset

Concatenated Features
E_{RMS} Training = 0.53842
E_{RMS} Validation = 0.43268
E_{RMS} Testing = 0.51195

Subtracted Features
E_{RMS} Training = 0.54467
E_{RMS} Validation = 0.55423
E_{RMS} Testing = 0.52849

GSC Dataset

Concatenated Features
E_{RMS} Training = 0.51039
E_{RMS} Validation = 0.53611
E_{RMS} Testing = 0.4001

Subtracted Features
E_{RMS} Training = 0.50721
E_{RMS} Validation = 0.49483
E_{RMS} Testing = 0.50748

Observation 2

Number of Basis Functions (M)	Regularization Term (λ)	Learning Rate
20	2	0.08

Human Observed Dataset

Concatenated Features
E_{RMS} Training = 0.4974
E_{RMS} Validation = 0.47107
E_{RMS} Testing = 0.49889

Subtracted Features
E_{RMS} Training = 0.50074
E_{RMS} Validation = 0.40103
E_{RMS} Testing = 0.49923

GSC Dataset

Concatenated Features
E_{RMS} Training = 0.49971
E_{RMS} Validation = 0.48991
E_{RMS} Testing = 0.49889

Subtracted Features
E_{RMS} Training = 0.50057
E_{RMS} Validation = 0.47141
E_{RMS} Testing = 0.49889

Observation 3

Number of Basis Functions (M)	Regularization Term (λ)	Learning Rate
20	2	0.001

Human Observed Dataset

Concatenated Features
E_{RMS} Training = 0.62967
E_{RMS} Validation = 0.47334
E_{RMS} Testing = 0.60695

Subtracted Features
E_{RMS} Training = 0.50774
E_{RMS} Validation = 0.53041
E_{RMS} Testing = 0.5591

GSC Dataset

Concatenated Features
E_{RMS} Training = 0.49997
E_{RMS} Validation = 0.44222
E_{RMS} Testing = 0.47492

Subtracted Features
E_{RMS} Training = 0.49999
E_{RMS} Validation = 0.46694
E_{RMS} Testing = 0.48863

Observation 4

Number of Basis Functions (M)	Regularization Term (λ)	Learning Rate
25	2	0.001

Human Observed Dataset

Concatenated Features
E_{RMS} Training = 0.76249
E_{RMS} Validation = 0.52318
E_{RMS} Testing = 0.87939

Subtracted Features
E_{RMS} Training = 0.51201
E_{RMS} Validation = 0.51497
E_{RMS} Testing = 0.512

GSC Dataset

Concatenated Features
E_{RMS} Training = 0.50007
E_{RMS} Validation = 0.49855
E_{RMS} Testing = 0.49889

Subtracted Features
E_{RMS} Training = 0.50236
E_{RMS} Validation = 0.4899
E_{RMS} Testing = 0.4746

From the above observations, we conclude that Linear Regression isn't an appropriate choice of Machine Learning algorithm for this dataset. There can be several reasons for this such as, Linear Regression is too simple for a comprehensive dataset such as this, this dataset may require an activation function or a much deeper network to get better results.

We hence try to achieve better results using Logistic Regression.

Logistic Regression (Linear Regression vs Logistic Regression [1])

It's tempting to use the linear regression output as probabilities but it's a mistake because the output can be negative, and greater than 1 whereas probability cannot. As regression might actually produce probabilities that could be less than 0, or even bigger than 1, logistic regression was introduced.

- **Outcome**

  In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values.

  In logistic regression, the outcome (dependent variable) has only a limited number of possible values.

- **The dependent variable**

  Logistic regression is used when the response variable is categorical in nature whereas Linear regression is used when your response variable is continuous.

- **Equation**

  Linear regression gives an equation which is of the form $Y = mX + C$, means equation with degree 1.

  However, logistic regression gives an equation which is of the form $Y = e^{X} + e^{-X}$

- **Coefficient Interpretation**

  In linear regression, the coefficient interpretation of independent variables are quite straightforward (i.e. holding all other variables constant, with a unit increase in this variable, the dependent variable is expected to increase/decrease by xxx).

However, in logistic regression, depends on the family (binomial, Poisson, etc.) and link (log, logit, inverse-log, etc.) you use, the interpretation is different.

- **Error minimization technique**

  Linear regression uses ordinary least squares method to minimise the errors and arrive at a best possible fit, while logistic regression uses maximum likelihood method to arrive at the solution.

  Linear regression is usually solved by minimizing the least squares error of the model to the data, therefore large errors are penalized quadratically.

  Logistic regression is just the opposite. Using the logistic loss function causes large errors to be penalized to an asymptotically constant.

Logistic Regression Approach

1. Features → X
2. Target → y
3. Initialize weights (W) "normally" (i.e. using Normal Distribution)
4. $z = W^T X$
5. $$\sigma = \frac{1}{1 + e^{-x}}$$
6. Gradient Descent

   $X^T (\hat{y} - y)$ → / number of features → * learning rate → – W → new W
7. Repeat steps 4 thru 6 for N number of epochs
8. Test accuracy on unseen data

Logistic Regression Observations

With learning rate as 0.01 and 1000 epochs, we get the following results using Logistic Regression on our dataset:

Human Observed Dataset

Concatenated Features
E_{RMS} = 0.20664
Accuracy = 95 %

Subtracted Features
E_{RMS} = 0.46141
Accuracy = 71 %

GSC Dataset

Concatenated Features
E_{RMS} = 0.16177
Accuracy = 98 %

Subtracted Features
E_{RMS} = 0.244788
Accuracy = 90 %

Inferences

- We clearly observe that Logistic Regression provides much better results than Linear Regression.

- We observe that concatenated features provide better results than subtracted features.
  A possible explanation for this could be that the model has more features (double the features in concatenated than subtracted) to base its prediction on.

- We also observe that accuracy is better in the GSC dataset than in the Human Observed dataset.
  A possible explanation for this could be the fact that the GSC dataset has more features (512) compared to Human Observed dataset (9). Hence, like in the previous observation, our model has more features to base its prediction on.

References

1. Sayali Sonawane
   What is the difference between Linear Regression and Logistic Regression
   https://stackoverflow.com/questions/12146914/what-is-the-difference-between-linear-regression-and-logistic-regression