PROJECT REPORT

1 PROJECT SYNOPSIS

Project 2 is to find the similarity between the Learning to Rank problem using the two Machine Learning technique – Linear regression with Stochastic Gradient Descent and Logistic Regression.

The task is to find the similarity between the handwritten samples of the known and the questioned writer by using linear regression and logistic regression. The dataset from the CEDAR "AND" consists of input features with each hand written "AND" sample.

The Dataset has two parts, Human observed features and GSC features. The target value is a scalar which can take two values 0(different writers) or 1(same writers). Evaluation is done by with Erms and Accuracy.

2 IMPLEMENTATION

The data is preprocessed for concatenation and subtraction for both the datasets. The concatenated data and subtracted data each as to be trained and checked for Erms and Accuracy. Both the processed datasets are being solved with two techniques – Linear Regression and Logistic Regression. The Linear Regression is done via Stochastic Gradient Descent. For the Linear Regression, code 1.2 is recycled.

The hyper parameter tuning is done with grid search methodology.

2.1 DATA PREPROCESSING

The datasets given to us is concatenated and subtracted with the same and different pairs of writers with the human observed dataset and GSC Dataset respectively. Each of them will

2.1.1 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. 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).

2.1.1.1 Concatenation

Each sample in Human observed dataset will have 9 features. Therefor each pair of sample will have 18 features when concatenated. I have taken nearly 20% of same pairs and 80% of diffn pairs for balancing the dataset.

Here there are total of 791 data samples in same pairs and 293032 in diffn pairs. I have taken all the 791 pairs from same and 3000 pairs from the diffn pairs. I have set the Training data, Validation data and Testing data into 70%, 10% and 20% respectively.

1. At first the dataset ‘diffn pairs’ with random samples of 3000 rows is merged with human observed original dataset on image_id_A and image_id_B.
2. The dataset ‘same pairs’ with 791 sample is merged human observed original dataset on image_id_A and image_id_B.
3. The two parts are concatenated into one and saved as csv file with no index, and image_id’s.
4. The Raw data will have 18 features for each sample which is split into its required datasets.

2.1.1.2 Subtraction

The Concatenated data is used for the subtraction which will result in 9 features for each sample. The result is passed for linear and logistic regression.

2.1.2 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”. The entire dataset consists of 71,531 same writer pairs and 762,557 different writer pairs(rows). We have to build a dataset using GSC-FeaturesData.csv, same pairs.csv and diffn pairs.csv.

2.1.2.1 Concatenation

Each sample in GSC dataset will have 512 features. There each pair of sample will have 1024 features when concatenated. I have taken 50% each of same and diffn pairs for balancing the data.

Here the total of 10000 data samples in same pairs and 10000 in diffn pairs are taken randomly. And this dataset is further split into Training data, Validation data and Testing data into 70%, 10% and 20% respectively.

1. At first the dataset ‘diffn pairs’ with random samples of 10000 rows is merged with human observed original dataset on image_id_A and image_id_B.
2. The dataset ‘same pairs’ with 10000 sample is merged human observed original dataset on image_id_A and image_id_B.
3. The two parts are concatenated into one and saved as csv file with no index, and image_id’s.
4. The Raw data will have 1024 features for each sample which is split into its required datasets.

2.1.2.2 Subtraction

The concatenated data is used for subtraction which will result in 512 feature dataset. Which is further passed in the remaining steps.

2.2 Linear Regression – SGD

Stochastic gradient descent (SGD), also known as incremental gradient descent, is an iterative method for optimizing a differentiable objective function, a stochastic approximation of gradient descent optimization.

I have recycled the code of 1.2 project for linear regression in project2.

We have divided the training set into pairs of input values and target values. We use x to represent x input vector, and map to real valued scalar target, y(x, w). It is defined as:

Where w is the weight vector, represented as w = (w0, w1, .., wM−1) which will be learnt from the training, , is a vector of M basis functions. Here w0 is considered as bias term.

Gradient Descent is an optimization algorithm; the weights are incremented after each epoch. With large datasets the GD would make the algorithm slower since it as goes through the entire training dataset to update the weights. But with SGD, the updates of weights per-iteration, computational cost in gradient descent scales linearly with the training data set size n.

2.3 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. sigmoid function outputs the conditional probabilities of the prediction, the class probabilities.

At the center of the logistic regression analysis is the task estimating the log odds of an event. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\, of\, presence\, of\, characteristic}{probability\, of\, absence\, of\, characteristic}$$

3 Observations

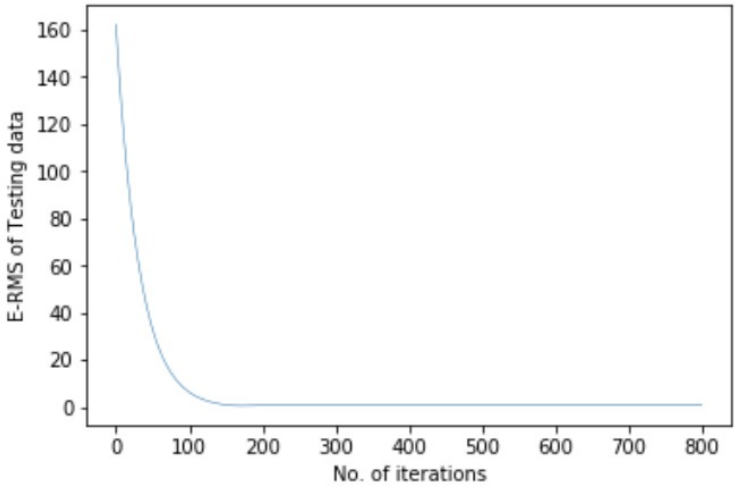
3.1 Linear Regression

Tuning the hyperparameters – Clusters(M) , Learning Rate ,Lamda and Epcoh. I have tabulated the accuracy with various values.

3.1.1 Concatenated Human Observed Dataset

Cluster(M)	Learning Rate	Lamda	Epoch	ERMS_Trai	Erms_Val	Erms_Test
10	0.03	2	400	0.409	0.397	0.394
10	0.1	2	400	0.427	0.399	0394
15	0.05	3	400	0.409	0.397	0.394
20	0.05	3	500	0.4133	0.399	0.395

The Best Results was for Cluster = 20 , Learning rate = 0.05 , lamda = 3 and Epoch = 500 with Erms_test = 0.395 .



3.1.2 Subtracted Human Observed Dataset

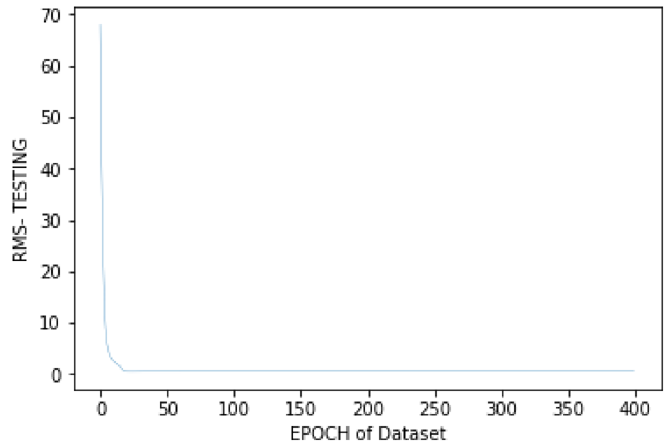
Cluster(M)	Learning Rate	Lamda	Epoch	ERMS_Trai	Erms_Val	Erms_Test
18	0.1	2	500	0.405	0.422	0.379
10	0.1	2	400	0.417	0.412	0380
15	0.01	3	500	0.401	0.410	0.409
12	0.1	2	400	0.407	0.422	0.379

The Best Results was for Cluster = 18 , Learning rate = 0.01 , lamda = 2 and Epoch = 500 with Erms_test = 0.379 .

3.1.3 Concatenated GSC

Cluster(M)	Learning Rate	Lamda	Epoch	ERMS_Trai	Erms_Val	Erms_Test
10	0.01	2	400	0.592	0.501	0.504
15	0.1	2	400	0.453	0.464	0.454
15	0.01	2	400	0.469	0.481	0.473
15	0.03	3	400	0.481	0.488	0.491

The Best Results was for Cluster = 15 , Learning rate = 0.1 , lamda = 2 and Epoch = 400 with Erms_test = 0.454.



3.1.4 Subtracted GSC

Cluster(M)	Learning Rate	Lamda	Epoch	ERMS_Trai	Erms_Val	Erms_Test
10	0.1	2	400	0.483	0.490	0.484
15	0.01	2	400	0.453	0.464	0.454
15	0.06	3	500	0.444	0.457	0.445
10	0.01	2	500	0.488	0.498	0.491

The Best Results was for Cluster = 15 , Learning rate = 0.06, lambda = 3 and Epoch = 500 with Erms_test = 0.445.

3.2 Logistic Regression

The Hyper parameters which needs to be tuned are learning rate and Number of Iterations.

3.2.1 Concatenated Human Observed Dataset

Learning Rate	Number of Iterations	Accuracy(%)
0.1	200000	79.19
0.01	200000	79.19
0.05	150000	80.4

The Best Results was for Learning rate = 0.01 and Epoch = 150000 with Acc = 80.4% .

3.2.2 Subtracted Human Observed Dataset

Learning Rate	Number of Iterations	Accuracy(%)
0.1	200000	81.21
0.01	100000	78.53
0.05	150000	78.86

The Best Results was for Learning rate = 0.1 and Epoch = 200000 with Acc = 81.21% .

3.2.3 Concatenated GSC

Learning Rate	Number of Iterations	Accuracy(%)
0.01	15000	80.1
0.001	10000	81.6
0.1	10000	78.9

The Best Results was for Learning rate = 0.001 and Epoch = 10000 with Acc = 81.6% .

3.2.4 Subtracted GSC

Learning Rate	Number of Iterations	Accuracy(%)
0.1	20000	81.21
0.01	10000	78.53
0.05	150000	78.86

The Best Results was for Learning rate = 0.1 and Epoch = 20000 with Acc = 81.21% .

4 REFERENCES

1. https://en.wikipedia.org/wiki/Stochastic_gradient_descent
2. https://medium.com/we-are-orb/multivariate-linear-regression-in-python-without-scikit-learn-7091b1d45905
3. https://machinelearningmastery.com/logistic-regression-for-machine-learning/
4. https://towardsdatascience.com/machine-learning-part-3-logistics-regression-9d890928680f