UNIVERSITY OF MISSOURI, KANSAS CITY

PYTHON & DEEP LEARNING – CS 5590

PART – 2 - DEEP LEARNING

LAB – 1 REPORT

BHAVYA TEJA GURIJALA
STUDENT ID: 16220446
CLASS ID: 14

TASK:

Implement the Logistic Regression using the Tensor Flow in python.

1. INTRODUCTION:

Logistic Regression is a machine learning algorithm used to predict the values which are Boolean in nature i.e., which are of the form True/False, 0/1, Yes/No. The computation is done in such way that a set of independent variables are used to predict the target variable which is dependent on all the other variables.

Logistic Regression is same as that of the Linear Regression but the only difference is the prediction is for only the Boolean values and is of a sigmoid function which is either 1 or 0. The prediction is done based on the value which is in favor of the occurrence of the event. The value lies in the range of 0 to 1. We use the round function to know whether the it supports 0 or 1 for the prediction. It is like probability which should be between 0 and 1 and cannot exceed 1.

The dataset that I chose to perform the logistic regression identifies the risk that is associated with giving birth to the child who is low in weight (less than 2500 kilograms). The logistic regression algorithm is chosen as we should predict whether the child is a low weight baby or not which is a Boolean situation.

2. OBJECTIVES:

The objective of this model is to predict whether the model able to predict whether the child is a low birth weight baby or not. The prediction is done using the independent variables like Age of mother, Weight during the last menstrual period, Race, Smoke, History of pre-mature labor, Hyper Tension, Uterine Irritability, Birth weight in grams. The performance of the model depends on well the model predicts the low birth weight situation of the children based on the listed variables.

3. APPROACH:

The approach I used to implement the logistic regression to build the model is by applying the sigmoid function to the regular line equation as it results in a Boolean function.

$$\mathbf{Y = Sigmoid(A * X + B)}$$

Where

X – Matrix of all the independent variables.

Y – Target or dependent variable which is to be predicted.

Tensor Flow has a built in cross entropy function called 'tf.nn.sigmoid_cross_entropy_with_logits()' which lists the loss every time it iterates. And as it is an iterative process which depends on the batch size we calculate the mean of all the losses using the loss function.

$$\mathbf{loss = mean(- y * log(predicted) + (1-y) * log(1-predicted))}$$

The main objective behind calculating the mean of the loss is to minimize the loss by the time all random batches are iterated.

4. WORKFLOW:

The workflow of the entire model building is as follows:

a. The dataset is read using csv.reader and each row is appended to a list by converting every value in it to float.
b. Creating a matrix of all the independent variables which are used for prediction and a single row matrix for the dependent variable.
c. Setting the random.seed for generating random batches every time the loop runs.
d. Splitting the dataset into train and test sets which is of the ratio 80:20.
e. Applying feature scaling to all the columns by normalizing the data to avoid huge differences in the data.

f. Now defining the tensor flow computational graph by giving the batch size as 25.
g. Initializing the placeholders and creating the variables for the data and the target variables.
h. Declaring the model operation by using the sigmoid function.
i. Declare the loss function and the gradient descent optimizer.
j. Training the loop by running the session for the randomly generated batches and calculating the predictions using the output.
k. Now plotting the graph for the cross entropy and the test, train accuracy for every generation of the batch.

5. DATASET:

The dataset consists of 189 rows and 8 columns of which 7 are independent variables and one dependent variable which is the target variable. The data were collected from 189 women of which 59 women gave birth to low weight babies and rest normally.

The variables are:

a. Age of mother
b. Weight during the last menstrual period
c. Race
d. Smoke
e. History of the pre-mature labor
f. Hyper Tension
g. Uterine Irritability
h. Birth weight in grams

By using the above variables, we should predict whether the child born in a low weight baby or not.

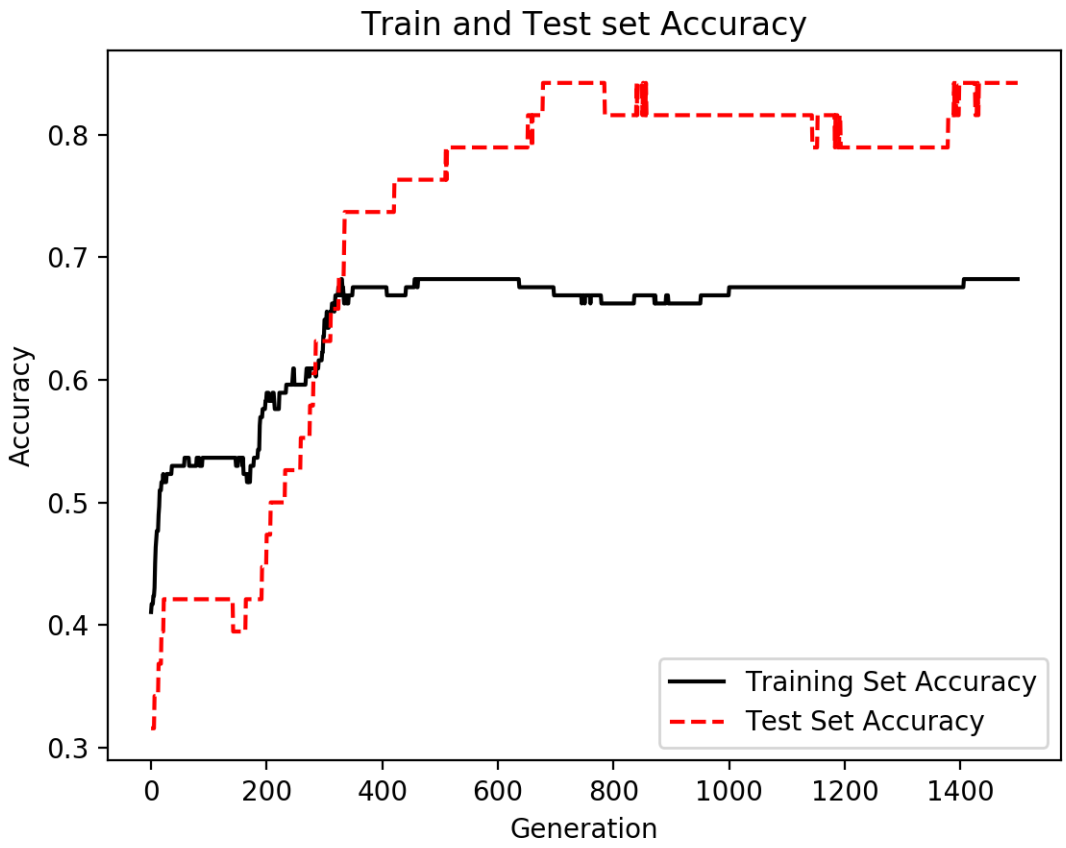
6. PARAMETERS:

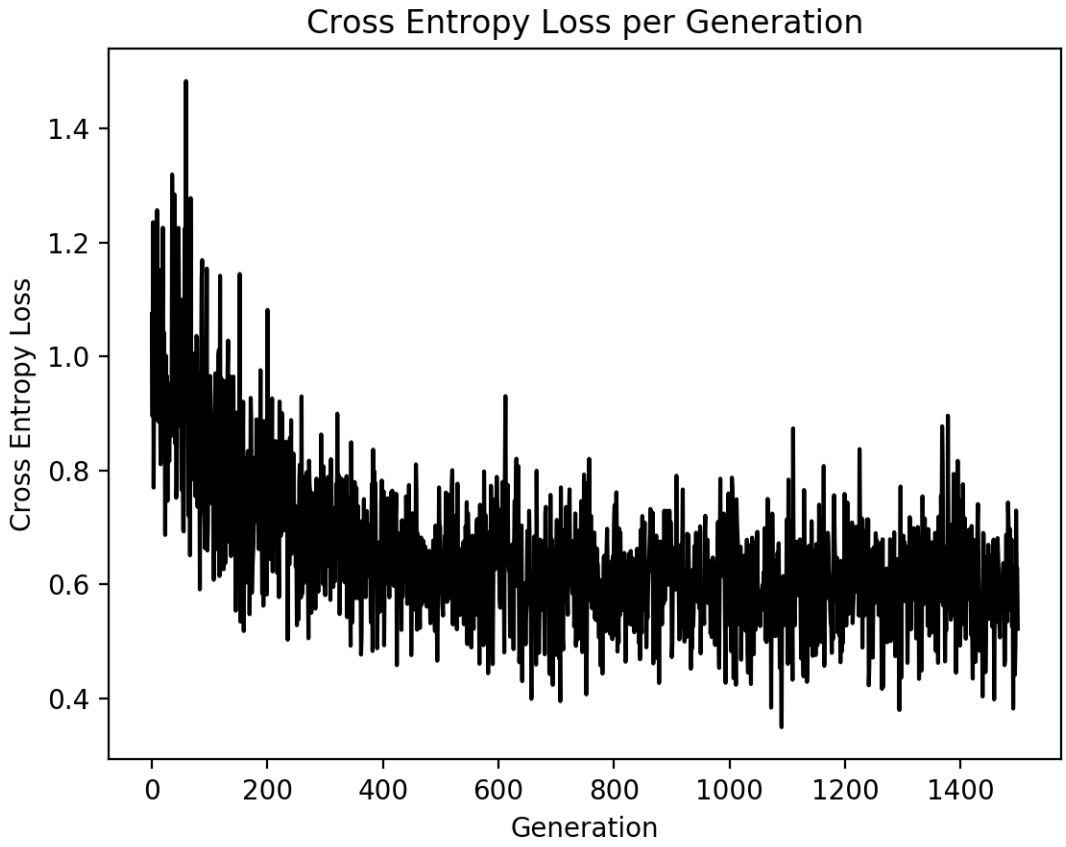
The parameters to be considered while building the model are:

a. Avoiding over fitting of the data while building the model.
b. Data preprocessing by feature scaling the data.
c. Make sure there are no NAs in the dataset.
d. Setting the random.seed value for random generation of the sets.
e. Batch size should be low so that the loss could be minimized.
f. Using the gradient descent optimizer for better results

7. EVALUATION & DISCUSSION:

The model can be evaluated by the performance graphs portray the loss and the accuracy of the model.





By looking at the graph we can say the accuracy of the model is very high as the loss gradually decreases as the batch generation happens.

8. CONCLUSION:

Eventually we conclude that low weight baby birth is better predicted using the logistic regression by considering various parameters which affect the performance of the model.

References:

https://github.com/nfmcclure/tensorflow_cookbook
https://www.statcrunch.com/5.0/shareddata.php?keywords=birth+weight
http://jrmeyer.github.io/tutorial/2016/02/01/TensorFlow-Tutorial.html