DEEP LEARNING -LAB1

TASK:

To implement the Logistic Regression in python using Tensor Flow.

1. INTRODUCTION:

Logistic Regression is a model which is used to predict the binary classified output i.e., True/False, 0/1, Yes/No. The calculation is basically the prediction of target variable i.e. dependent on other variables, using a set of independent variables.

The difference between Logistic and Linear regression is that in the former predicts the values of Boolean and Sigmod function which results according to the occurrence of the event. The value lies in the range of 0 to 1. Round function is used to check the prediction if it is 0 or 1. Also, it cannot exceed 1.

The dataset that I chose to perform the logistic regression identifies the risk that is associated in predicting the quality of wine. The logistic regression algorithm is chosen as we should predict whether the quality of wine is good or bad.

2. OBJECTIVES:

A range of regression techniques have been developed for analyzing data with categorical dependent variables, including logistic regression and discriminant analysis.

The objective of this model is to predict whether the model able to predict the quality of wine. The prediction is done using the independent variables like fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free Sulphur-dioxide, density of wine, alcohol. The performance of the model depends on well the model predicts the quality of wine.

3. APPROACH:

Logistic Regression generates a value where is always either 0 or 1.

-- Interpreting hypothesis:

- $h_\theta(x) = P(y=1|x ; \theta)$
- Since this is a binary classification task we know y = 0 or 1
- So the following must be true
  - $P(y=1|x ; \theta) + P(y=0|x ; \theta) = 1$
  - $P(y=0|x ; \theta) = 1 - P(y=1|x ; \theta)$

Logistic regression approach has been used to build the model applying the sigmoid function on the regular line equation resulted in Boolean function.

Y_Predicted = Sigmoid(W * X + Bias)

Where

X – Matrix of all the independent variables.

Y _Predicted – 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 list out the loss for every iteration. The batch size is considered for calculating the mean of all the losses resulted using the loss function.

```
x_entropy = tf.nn.sigmoid_cross_entropy_with_logits(y_pred, y_)

 loss = tf.reduce_mean(x_entropy)
train_step = tf.train.GradientDescentOptimizer(0.01).minimize(loss)
```

The learning value chosen is 0.01 and the Gradient Descent Optimizer function is used for the for reducing the loss. The loss is being minimized at every iteration and 1500 iterations are chosen for building the logistic regression model.

The mean of these losses are calculated in order to minimize the loss till all the random batches have iterated.

4. WORKFLOW:

The workflow is as follows:

a. The dataset is read using by using csv.reader
b. Preprocess the data and split the data into training data and the test data.
c. Defining the parameters for the model.
d. Initialize the placeholders
e. Create the variables for the data and target variables.
f. Creating weights and bias.
g. Predicting the output variable by using the independent variables along with weights and bias.
h. Construct the model by using softmax.
i. Declaring the loss function
j. Declaring the gradient descent optimizer.
k. Minimize the loss using the cross entropy.
l. The loop is trained to calculate the predictions on running the session .
m. Finally, plotting the graph for all the batch generations.

5. DATASET:

The dataset consists of 1600 rows and 15 columns of which 14 are independent variables and one dependent variable which is the target variable and the dataset is wine mixture.

The variables are:

a. Fixed acidity
b. Volatile acidity
c. Citric acid
d. Residual Sugar
e. Chlorides
f. Free Sulphur Dioxide
g. Density
h. pH
i. Sulphates
j. Alcohol
k. Quality

By using the above variables, we should predict the quality of the alcohol.

6. PARAMETERS:

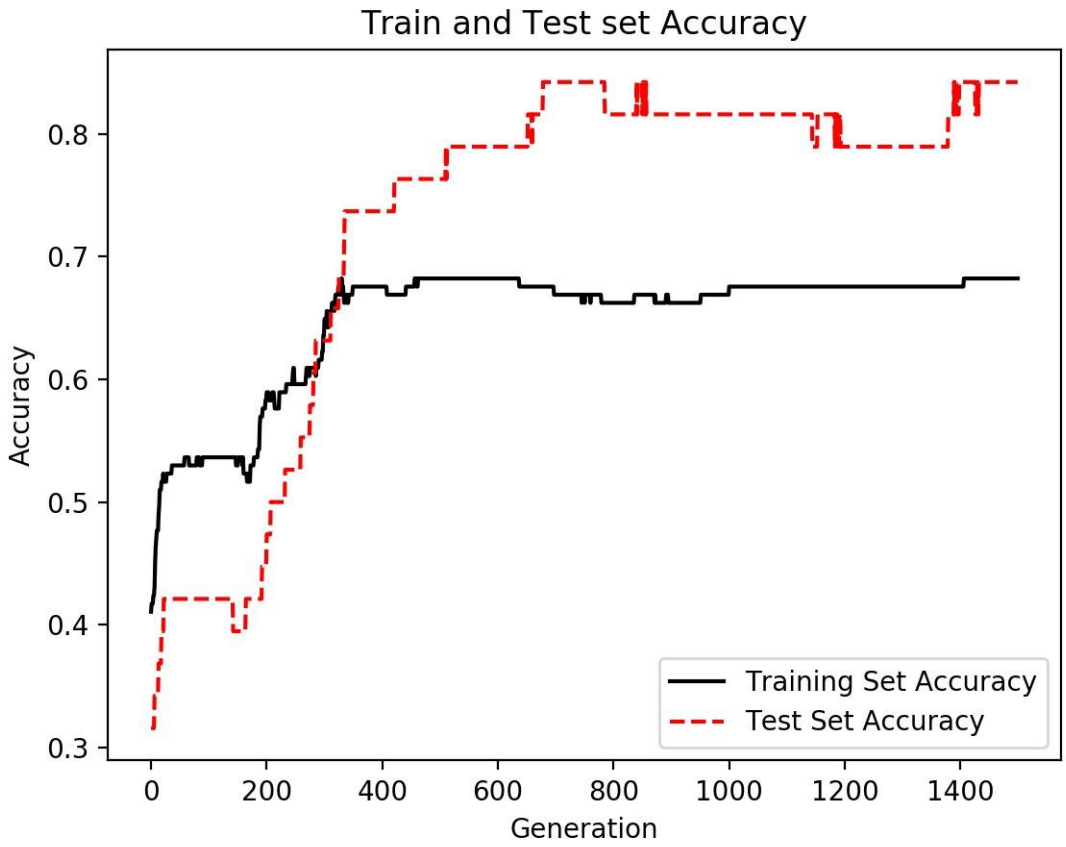
The parameters assumed are:

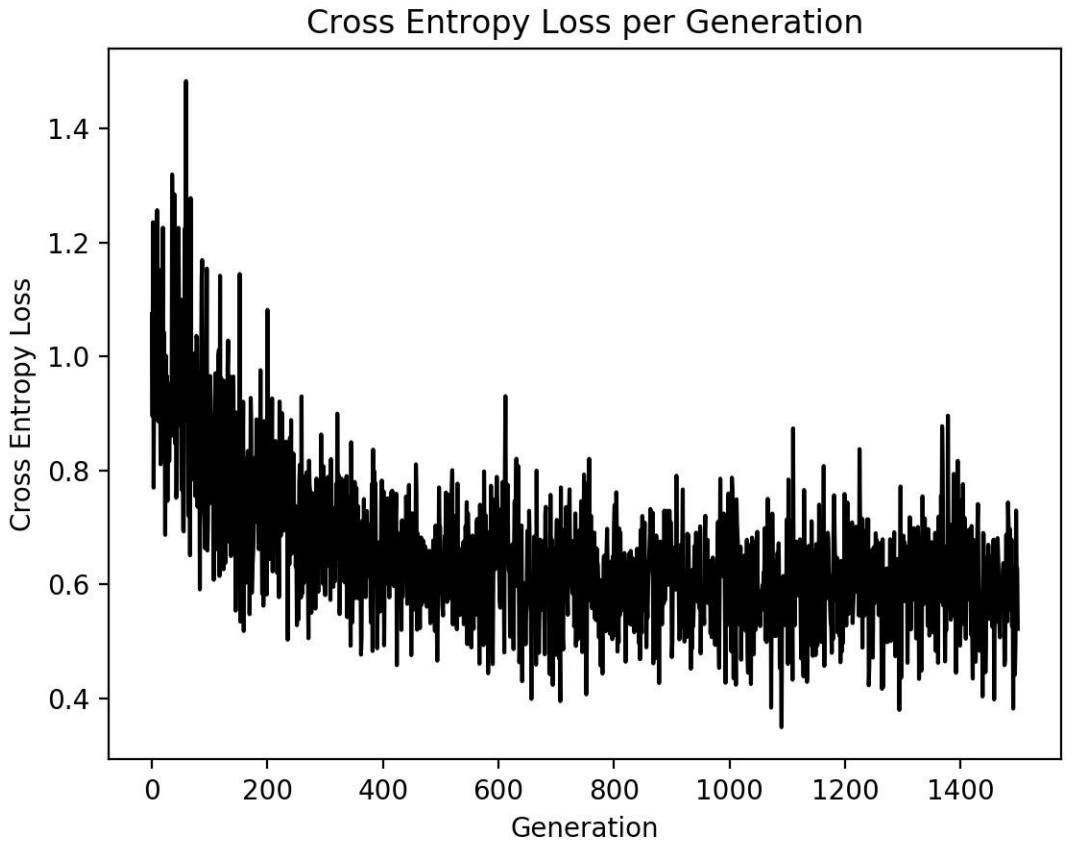
a. To avoid over fitting and underfitting of the data while building the model.
b. Preprocessing of Data by bringing all the data into a predefined range.
c. There should not be any null values in the dataset.

d. For random generation of the sets we need to set the random.seed value.
e. For the minimal loss the batch size should be very low.
f. The gradient descent optimizer is used for efficient results.

7. EVALUATION & DISCUSSION:

The model is evaluated by the performance graphs which shows the amount of loss and the accuracy of the model predicted.





The above graph shows that the accuracy of the model is very high as the loss gradually decreases for every batch generation.

8. CONCLUSION:

Eventually we conclude that the quality of alcohol can be predicted using the logistic regression by considering various parameters which affect the performance and accuracy metrics of the model.

References:

https://github.com/nfmcclure/tensorflow

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality