Classification using Logistic Regression

Deepanshu Yadav
UB Person No:
50321285
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
dyadav@buffalo.edu

Abstract

This study demonstrated use of logistic regression as a classifier to classify suspected fine needle aspirate (FNA) cells to Benign (class 0) or Malignant (class 1). The study was performed on dataset containing 569 instances with 30 real-valued input features. Python 2.7 and related libraries were used to manipulate the data and implement the model. It was observed that after tuning the hyperparameters using training and validation datasets, the model was able to classify the FNA cells to correct class with significant accuracy and precision.

1 Introduction

Many real-world classification problems call for the analysis and prediction of a dichotomous outcome: whether a student will succeed in college, whether a child should be classified as learning disabled (LD), is this picture of a dog or not, and so on. There are various ways of addressing such classification problems. Two of them are linear discriminant function analysis and logistic regression. Linear analysis has stricter statistical assumptions than logistic regression so is less preferred now a days. Logistic regression is increasingly becoming more popular for dichotomous outcome problems.

In this study, a logistic regression model is developed from scratch using Wisconsin Diagnostic Breast Cancer (WDBC) dataset to classify whether the FNA cells are malignant or benign. The dataset contains 569 samples with 30 attributes of each cell like radius, texture, symmetry, fractal dimension, etc. It also contains associated state (Malignant or Benign) of each cell. Model is developed using 80% of data as training set and 10% as validation set and is tested on rest 10% of unseen data. Results show significant accuracy and very low loss values for all the partitioned datasets. It establishes the importance of logistic regression in addressing real world classification problems with the use of machine learning.

2 Related Work

The work in this report includes reading the dataset and preprocessing it before performing the logistic regression operation on it. The model uses gradient descent approach to refine the weights and bias and reach optimum value with dependencies on hyperparameters namely epochs and learning rate. The updated weights and bias are used to predict the outcome again in every epoch. The predicted value is then compared with actual target value to find out the cost and accuracy. An optimum solution is reached when loss is minimized and accuracy is significant. The results are plotted against epochs to obtain the loss and accuracy curves showing to check improvements with every epoch. The following sections of the report include dataset reading, pre-processing the data, developing model architecture and then the results. These present detailed analyses of each step involved in developing the model and testing it.

3 Dataset

Wisconsin Diagnostic Breast Cancer (WDBC) dataset was used for training, validation and testing. The dataset contains 569 instances with 32 attributes (ID, diagnosis (B/M), 30 real-valued input features). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Computed features describe the following characteristics of the cell nuclei present in the image:

Table 1: Characteristics of Cell Nuclei Present in the Image

S. No.	Description
1	Radius (mean of distances from center to points on the perimeter)
2	Texture (standard deviation of gray-scale values)
3	Perimeter
4	Area
5	Smoothness (local variation in radius lengths)
6	Compactness (perimeter2/area − 1.0)
7	Concavity (severity of concave portions of the contour
8	Concave points (number of concave portions of the contour)
9	Symmetry
10	Fractal dimension (“coastline approximation” - 1)

The mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

4 Pre - Processing

Preprocessing of the dataset was done using Pandas, NumPy and sklearn libraries of Python. First of all, the dataset was read and stored using read_csv () method available in Pandas library for further operations. As the dataset didn’t have headers, a parameter ‘header = None’ was passed in the method call to keep first raw of dataset intact as shown in code snippet below.

fullDataSet = pd.read_csv("/Users/deep/Downloads/CSE574/Projects/wdbc.csv", header = None)

The data was then cleaned by dropping the first column ID then mapping the Malignant and Benign column to 1 and 0 by using replace () method. The data was then partitioned into training, validation and test sets using train_test_split () method from sklearn library. The target values were stored in another array and then dropped from the original partitioned files. The rest of the dataset containing the 30 attributes was normalized to remove the bias which could have occurred due to large variations in values between the columns.

5 Model Architecture

Overview: The goal of this logistic regression model was to predict the state of a cell (whether malignant or not) correctly by minimizing the cost and maximizing accuracy using machine learning. The first challenge was to preprocess the given dataset and manipulate it in a way it fits the equations used in developing the model. Also, the hyperparameters used (epochs and learning rate) also need to be tuned by trial and error method to reduce the cost.

The below mentioned figure represents the work flow of the model development for the given dataset with 30 attributes. As shown below, every attribute has a weight associated with it along with 1 bias for all attributes.

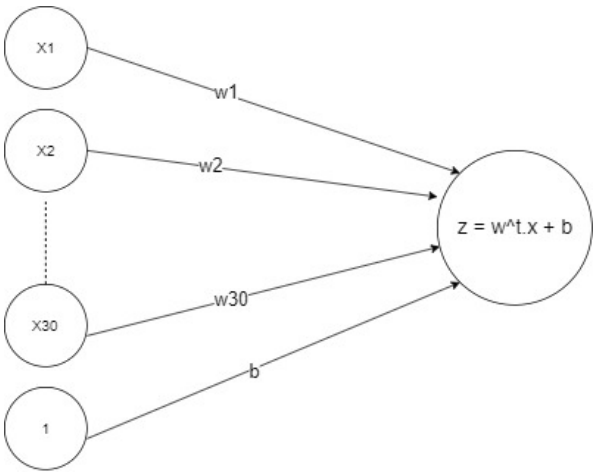

Fig 1: Computational Graph for Logistic Model for Dataset with 30 Attributes

The equation below summarizes the linear regression model represented in the above diagram.

$$z = w^t.x + b$$

where w and b are the weights and bias, x is the input vector and z is the output.

Logistic Regression model introduces an extra non-linearity over the linear classifier (eq above), by using a logistic (or sigmoid) function, σ().

Sigmoid function σ() is defined as

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

As z goes from $-\infty$ to ∞, σ(z) goes from 0 to 1 in the manner mentioned in following graph.

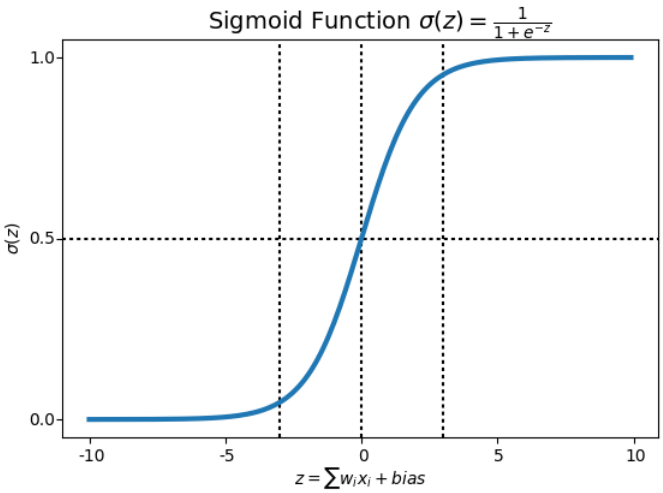

Fig 2: Sigmoid Function Plot

As the calculation is performed, for each sample, there will be an associated loss. The average of losses associated with each sample will give us the overall cost incurred in the model which is represented be the equation below

$$C = -\frac{ylog(\sigma(z)) + (1-y)\log\big(1-\sigma(z)\big)}{m}$$

where
C = total cost incurred
y = actual target value
σ(z) = predicted value as of result of calculations performed by the model
m = sample size

To minimize the cost, Gradient Descent algorithm is used which is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.
Gradients for weights and bias are calculated in the below mentioned way. First, we take the difference between predicted value σ(z) and target value (y).

$$dz = \sigma(z) - y$$

Then the difference between these values is used to calculate gradient for weights $dw = \frac{x.dz^T}{m}$ and that for bias $db = (\sum dz)/m$.
where
dw = gradient vector for weights
x = input vector
dz = difference between predicted value σ(z) and target value (y)
m = sample size

These gradients are used to calculate the updated weights and bias with every epoch as mentioned below using learning rate which can be described as size of steps taken to minimize the cost. This has to be optimum as if it's too low, calculating gradient will be time consuming. If it's too high, we risk overshooting the lowest point since slope of curve is constantly changing.

Updated weights and bias are represented by below equations.

$$w = w- \propto dw$$
$$b = b- \propto db$$

where $\propto$ is the learning rate.

This process is repeated for a number of iterations known as epochs, which is also a hyperparameter in the present model along with learning rate. No. of epochs are chosen such that cost gets minimized with significant accuracy and precision also obtained for given learning rate. In this study, epochs are chosen to be 10000 and learning rate is varied to get optimum values for weights and bias.

The predicted values from the sigmoid function are mapped to 0 and 1 using the following criteria.

$$\sigma(z)\begin{cases}\geq 0.5 & p = 1\\ < 0.5 & p = 0\end{cases}$$

Here, p is the predicted value.

These predicted values are then compared with actual target values to calculate the evaluation metrics explained below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

i.e.

$$Accuracy = 1 - \frac{\sum_{k=1}^{m} difference_k}{m}$$

Where difference = predicted value (p) – target value (y) for each data sample.

Precision and Recall metrics are calculated using following formulas.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Where
TP = True Positive
TN = True Negative
FP = False Positve
FN = False Negative

6 Results of Experimentation

The task to be performed is predicting the state of cell (Malignant or Benign) with some confidence.

Evaluation metrics. I have used Accuracy, Precision and Recall metrics for each dataset for evaluation of the efficiency of the model. Also, the cost values for training and validation datasets have been evaluated with respect to epochs. As the weights are initialized with random values before running the model every time, the model gives varied values of the evaluation metrics every time the model is run.

The hyperparameters for the model, i.e. epochs and learning rate were set to get the optimum results by trial and error method. Cost and accuracy were calculated for various learning rates and the one with minimum cost and maximum accuracy was chosen as the final learning rate for the model. Figures below represent loss and accuracy plots for four learning rates as 5, 0.05, 0.1 and 0.01.

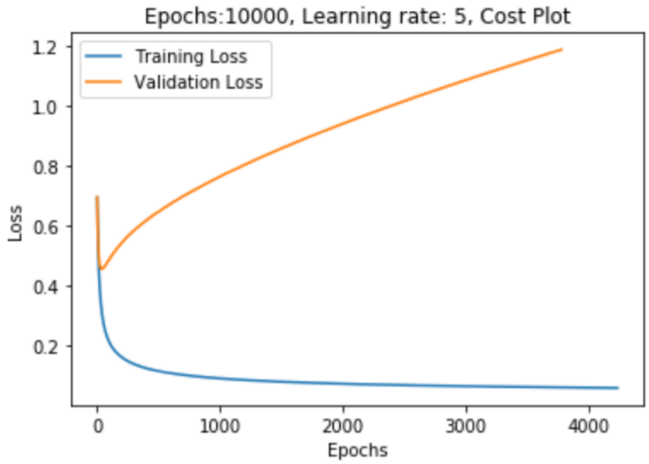

Fig 3: Training and Validation Cost for learning rate of 5

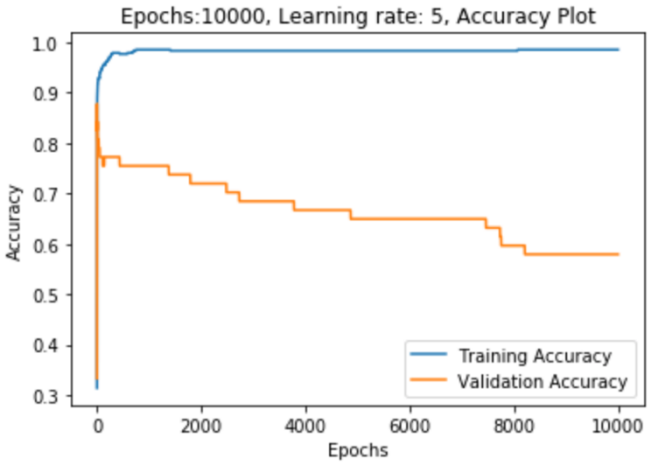

Fig 4: Training and Validation Accuracy for learning rate of 5

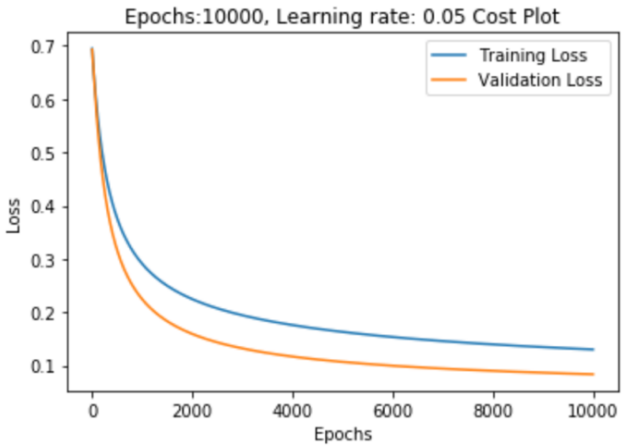

Fig 5: Training and Validation Cost for learning rate of 0.05

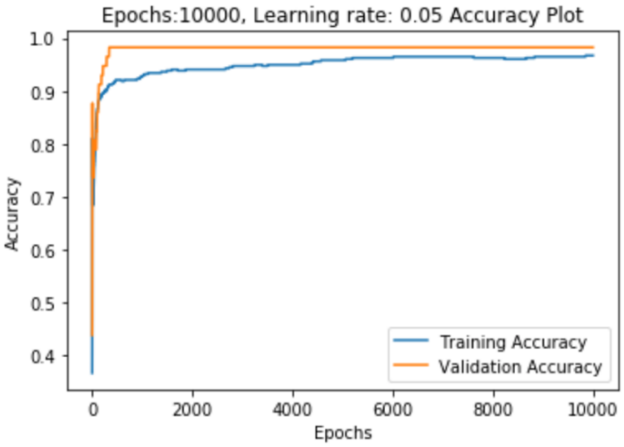

Fig 6: Training and Validation Accuracy for learning rate of 0.05

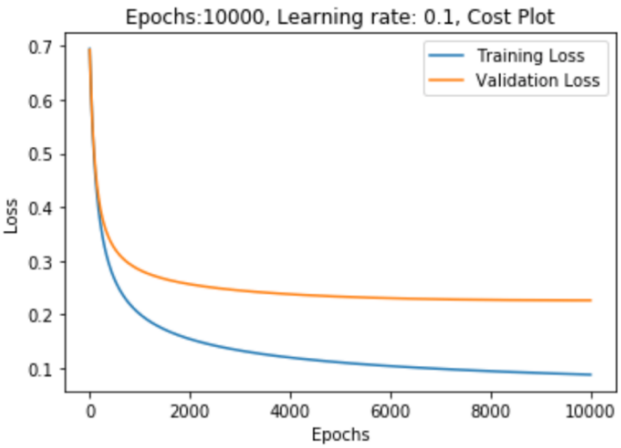

Fig 7: Training and Validation Cost for learning rate of 0.1

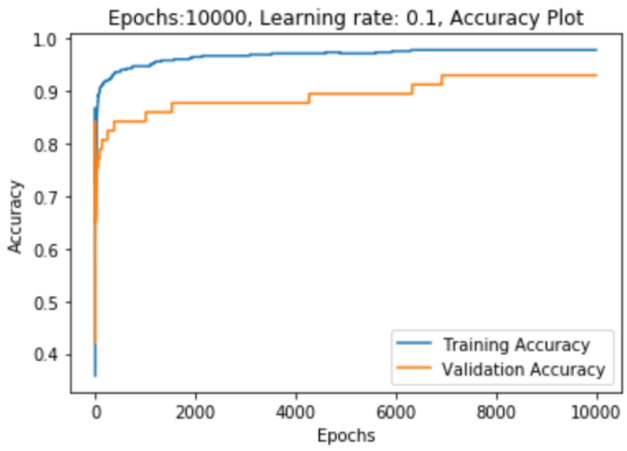

Fig 8: Training and Validation Accuracy for learning rate of 0.1

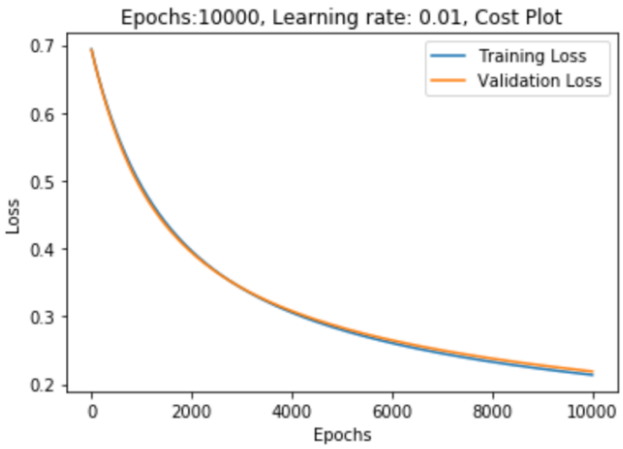

Fig 9: Training and Validation Cost for learning rate of 0.01

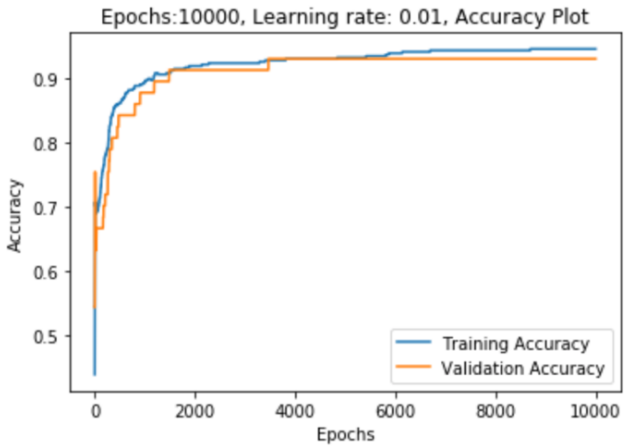

Fig 10: Training and Validation Accuracy for learning rate of 0.01

Table 2. Evaluated metrics for the respective datasets for learning rate of 0.01

For Training dataset:

Accuracy is: 94.5%

	precision	recall	f1-score
Benign	0.93	0.99	0.96
Malignant	0.98	0.87	0.92
macro avg	0.95	0.93	0.94

For Validation dataset:

Accuracy is: 93.0%

	precision	recall	f1-score
Benign	0.90	1.00	0.95
Malignant	1.00	0.82	0.90
macro avg	0.95	0.91	0.92

For Test dataset:

Accuracy is: 91.2%

	precision	recall	f1-score
Benign	0.94	0.91	0.93
Malignant	0.88	0.91	0.89
macro avg	0.91	0.91	0.91

In table 2, we can observe the accuracy, precision and recall metrics for learning rate of 0.01 and 10000 epochs for each dataset. The accuracy for test set is found to be 91.2%, precision, recall and f1-score were found to be 91% which implies that the logistic regression model developed is working with significant efficiency on unseen data.

7 Conclusion

In this study, logistic regression model was developed to be used as a classifier to classify suspected fine needle aspirate (FNA) cells to Benign (class 0) or Malignant (class 1). Model was run for various values of hyperparameters, i.e. epochs and learning rate and finally these were tuned to 10000 and 0.01 respectively after trial and error method. The evaluation metrics obtained showed significant efficiency of the model in classifying the cells into malignant and benign classes when tested on unseen data, i.e. test set.

References

[1] CHAO-YING JOANNE PENG, KUK LIDA LEE and GARY M. INGERSOLL. An Introduction to Logistic Regression

[2] http://www.robots.ox.ac.uk/~az/lectures/ml/2011/lect4.pdf

[3] https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html