Classification of Breast Cancer by Logistic Regression

Jyoti Sinha
Department of Computer Science
University of Buffalo
Buffalo, NY,14221
Email:-jsinha@buffalo.edu

Abstract

Breast cancer is one of the life-threatening and common cancer diseases among women and it has been increasing for several years. Although, it is a fatal disease but can be cured by early diagnosis. The early diagnosis is important as it increases the survival rate of the patients. Moreover, the accurate prediction of malignant and benign tumor is very important for the treatment and the prevention of unnecessary treatment of the patient. In this study, I used logistic regression to classify the suspected FNA cells to Benign (class 0) or Malignant (class 1). First, I analyze the dataset, perform the feature selection, build the logistic model and then made the prediction using the testing dataset. I achieved 96.491 percent accuracy on the testing dataset.

Introduction

There are various types of breast cancer like Ductal Carcinoma, Invasive Ductal Carcinoma, Metastatic Breast cancer, Inflammatory Breast cancer, Triple Negative Breast cancer and others and it varies with different stages or spread. Breast cancer is curable if it is detected at an early stage. Breast cancer detection is most commonly performed by mass screening in which the radiologist compare the new mammograms with the old mammograms. If the new mammograms are same as the old mammograms then there aren't likely to be cancer. But, this method increases the caseload of the radiologist and raises the chance of improper diagnosis. The prediction of breast cancer using Logistic Regression would help the radiologist to detect cancer in a much easier way.

There are various stages to detect breast cancer in women. First, the patient's history like age of the patient, trauma in the breast, lumps in the breast, menstrual cycle history, menopause history, patient's family history about any type of cancer etc is used to predict cancer or not. The result of the patient's history is used for the next step of the detection procedure. The next step is the clinical examination of both the breasts. This is a physical process in which the doctor examines the condition of the lumps, neighbouring area of the lumps and examine the entire breast. Doctors pay attention to the location of the lumps, shape and texture of the breast to predict the chance of the breast cancer. The next step of the breast cancer detection is Mammographic screening which is performed when the patient has an abnormal clinical examination. Mammographic screening is widely used for the prediction of breast cancer in most of the developed countries. This method is very common for cancer detection and gives the condition of calcification of the mass, their shape and texture information.

Logistic regression is most commonly used in the biomedical field. It is an algorithm to solve the classification problems. In this method, the model predicts the output based on the provided information. The model predicts an output after the training of the model is accomplished. Logistic regression is all about fitting the data into the logistic curve for the prediction. In this, the variables can be binary or multinomial. Logistic regression is a model in which the independent variables consist of different levels of size whereas the dependent variables are linear and fulfil the requirement needed for this method. The data which is used in the study determine the patient condition: (1) Negative of breast cancer, (2) Positive of breast cancer.

The outcome of the logistic regression can be used to verify the prediction of the doctor and can correct the wrong prediction made by them. The result achieved by the logistic model is compared with the prediction of the doctor. The prediction made by the logistic models based on various features of the patients can avoid the various steps involved to detect breast cancer. The logistic regression model can determine the level of breast cancer without performing the mammogram screening.

Dataset

The dataset that I am using in the study is Wisconsin Diagnostic Breast Cancer(WDBC) and is publicly available at the UCI Machine Learning Repository. This dataset is created by Dr. William H.Wolberg. He used the fluid samples which are extracted from the solid breast mass of the patient. He extracted ten features with the help of the fluid and the graphical computer program, which is capable to analyse the features based on the digital scans.
The dataset contains 569 instances and 32 attributes(Target:-B/M, ID),30 real-valued input features). He calculated the mean, standard error and worst of these features for each Images resulting in 30 attributes.

Attribute Information:-
1)Patient Id number
2)Target(B=benign, M=malignant)

The ten real-valued features are computed for each cell nucleus:-
1)radius (mean of distances from centre 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 (/ area-1.0)
7)concavity
8)concave points{number of concave portions of the contour)
9)symmetry
10)fractal dimension ("coastline approximation" -1)

Pre-processing

It is a method to convert the raw data into the clean or to the usable data. It is one of the important steps as the useful information and quality of data is fully derived from it. It affects the ability of our model to learn. Hence, it is important to pre-process the data before feeding into the model.

1) Splitting the dataset

I divided the dataset into two parts: -
1) First part contains 30 attributes
2) Second part contain the target variable

I dropped the Patient ID number as it will not contribute to the prediction of the output. Then, I divided the two different part of dataset into training, validation and testing in the ratio 8:1:1. Splitting of dataset is performed using SciKit-Learn library using train_test_split method.

Feature Normalization

The dataset contains features whose magnitudes vary highly. Thus, I performed the normalization of features to make the values of each feature in the data have zero mean and unit variance. The goal of normalization is to change the numeric value of the column to a common scale.

The formula for normalization: -

$\hat{x} = \frac{x-\ddot{x}}{\sigma}$

Where, $\ddot{x}$=mean of the feature vector
σ= standard deviation
x= feature input

Architecture:-

Hypothesis

Linear regression model is represented by the equation:-

h(x)= θ^Tx

Logistic regression uses sigmoid function that gives the output between 0 and 1 for all values of x to generate probabilities.

Thus, Apply the sigmoid function to the output of the linear regression

h(x)= $\sigma(\theta^T$x)

Where the sigmoid function is given by,

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

Thus, the hypothesis of the logistic regression becomes,

h(x)=$\frac{1}{1+e^{-\theta Tx}}$

h(x)= > 0.5, if $\theta^T x$ > 0
< 0.5, if $\theta^T x < 0$

If the value of the (θ^Tx) is greater than zero then the predicted class is 1
And if the value of the (θ^Tx) is less than zero then the predicted class is 0

Loss Function:-

The loss function is given by:

Loss(h(x),y)= (-y *log(h(x)) – (1-y)*log(1-h(x)))

The loss function for all the training examples can be calculated by taking the average of the loss function for all the training examples.

J(θ)= -$\frac{1}{m}\sum_{i=1}^{m}$($y^i$log(h($x^i$))+(1-$y^i$)log(1-h($x^i$)))

Where m =number of training sample

Our goal is to minimize the loss function and I am using Gradient Descent to minimize the loss function. Thus, gradient descent can be given by

$$= g(\theta) = \frac{dJ(\theta)}{d\theta_J} = \frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)x_j^i$$

Weights will be updated by subtracting the derivative (gradient descent) times the learning rate,

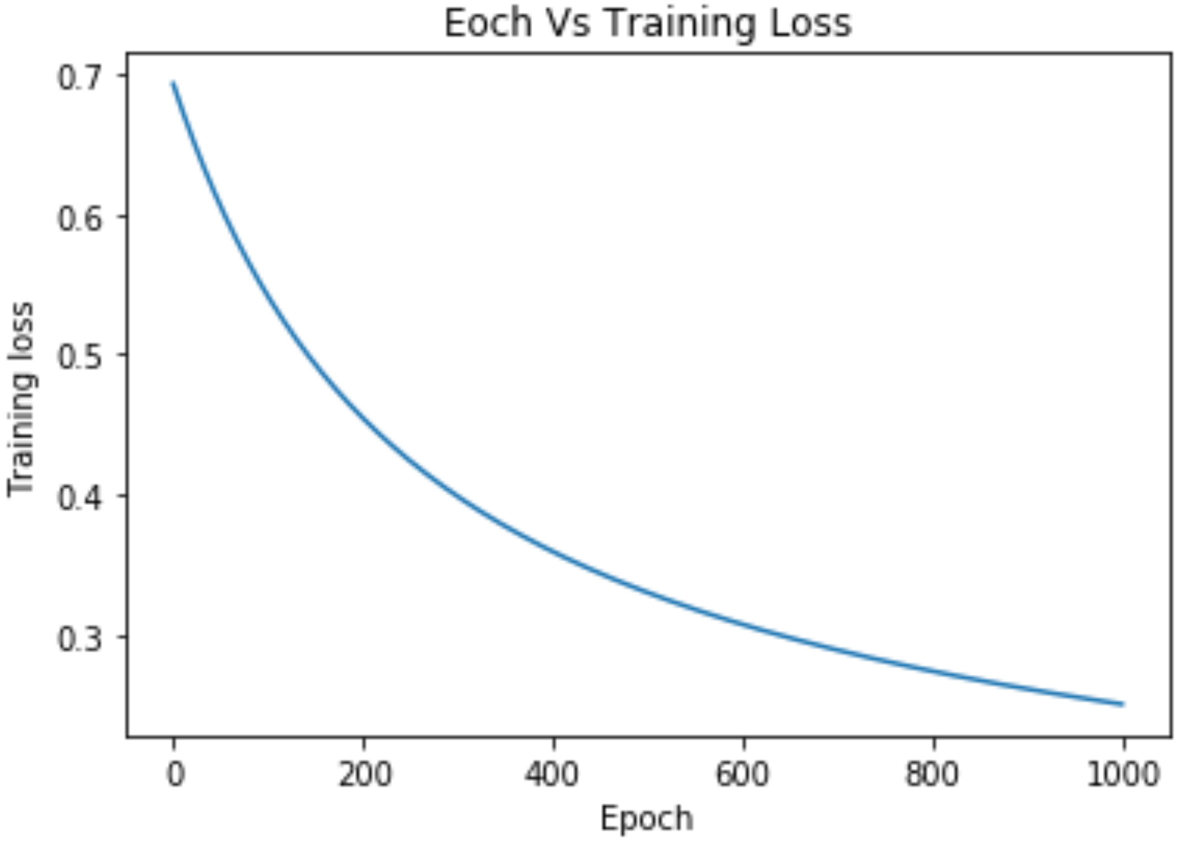
$$\theta = \theta - \alpha\frac{dy}{dx}(g(\theta))$$

Where α =learning rate
In my study, learning rate is 0.01

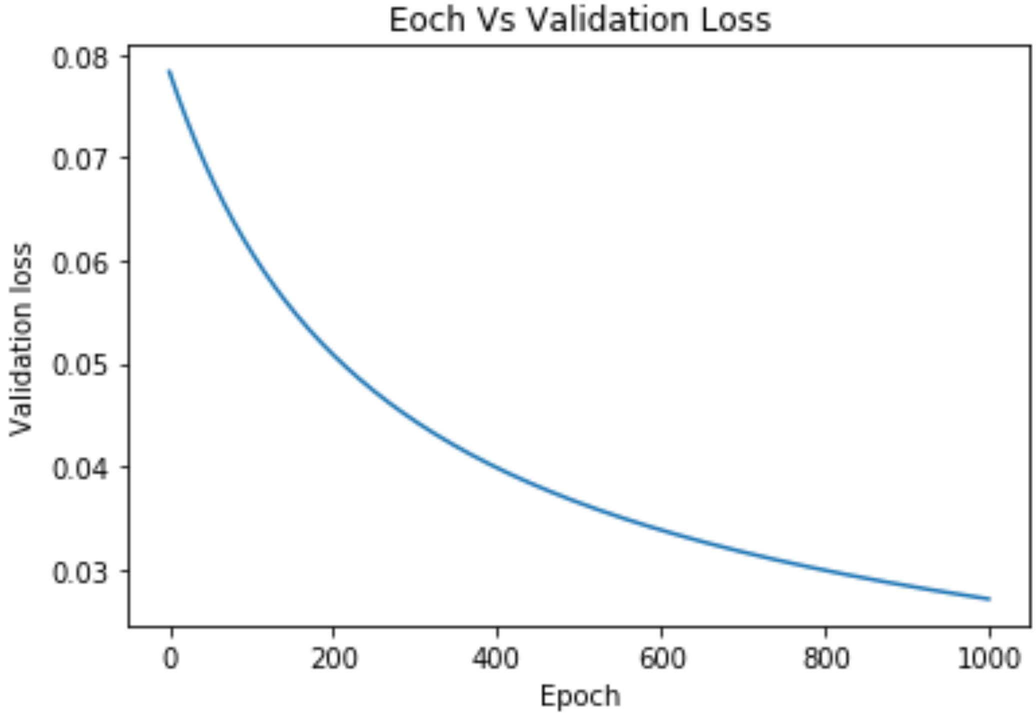
Result:-

In the study, I calculated the loss function for 1000 epoch and found that the loss function decreased for the entire tenure of the epochs. Further, I also calculated the training loss and the validation loss for the same number of epochs and plotted the graph between the epoch and loss.

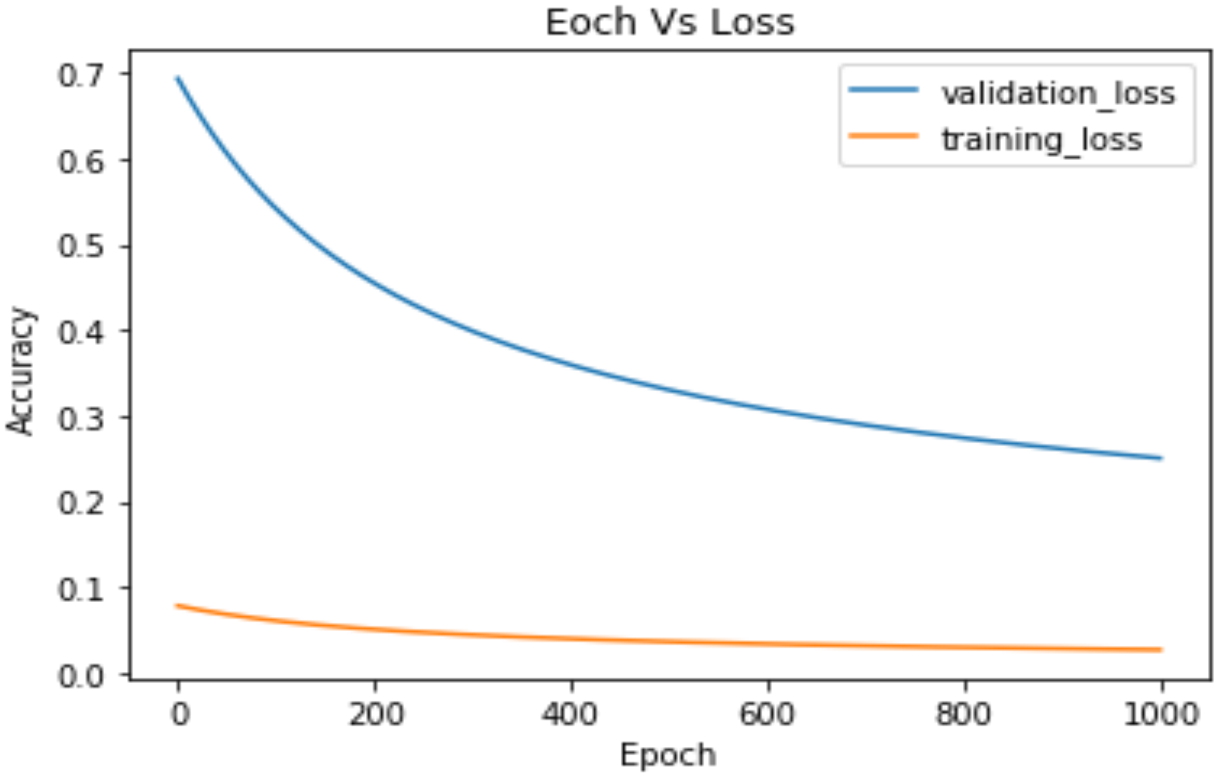
For Learning rate= 0.001
And Epochs= 10000



Graph:- Between Epoch and Training Loss

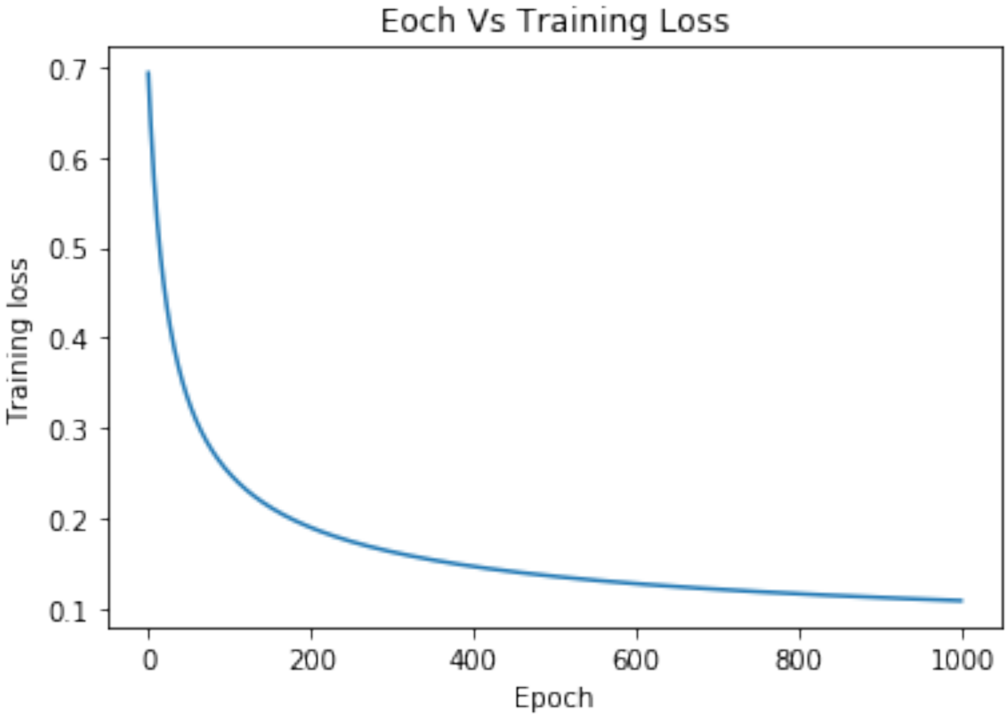


Graph :- Between Epoch and Validation Loss

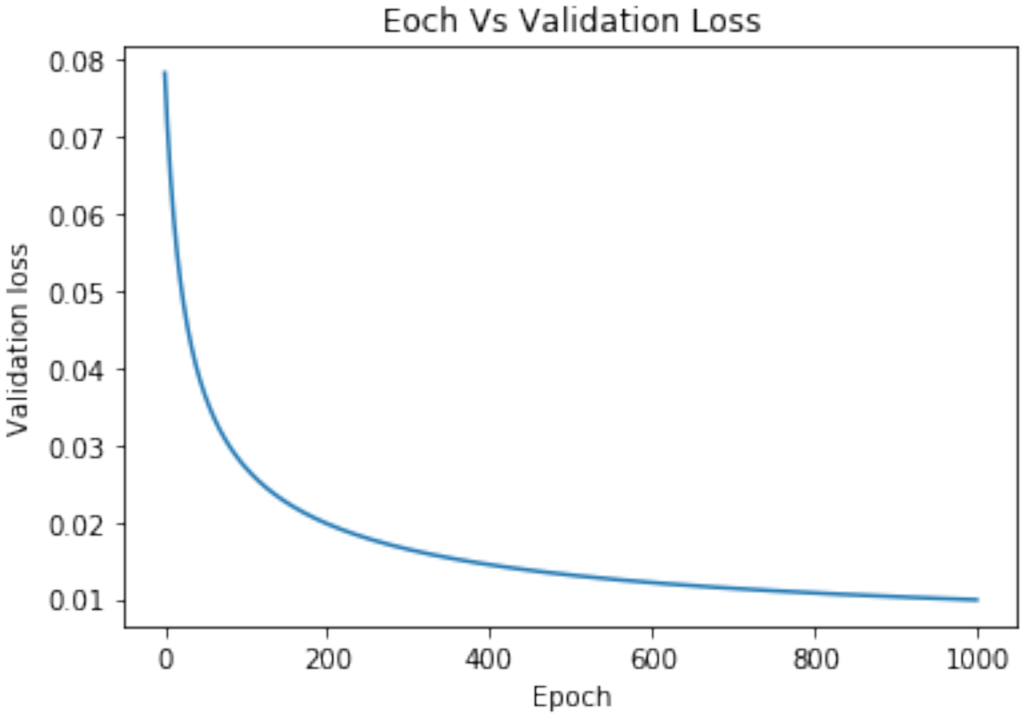


Graph: Between (Epoch) Vs (Training Loss and Validation loss)

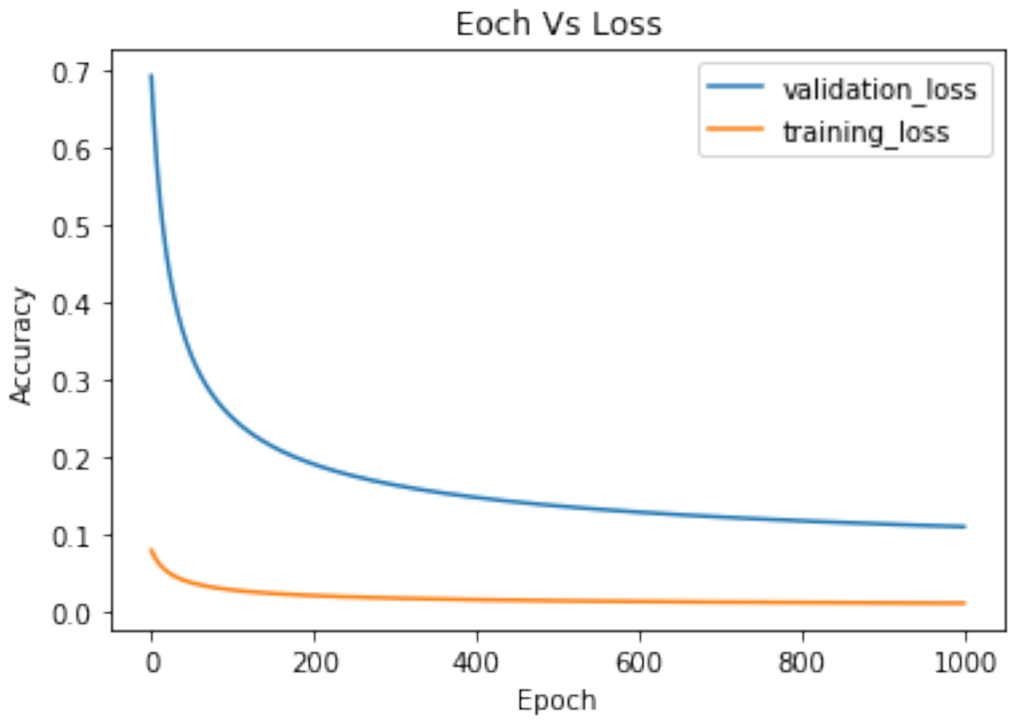
For learning rate=0.01
And Epoch= 10000



Graph:- Between Epoch and Training Loss



Graph :- Between Epoch and Validation Loss



Graph: Between (Epoch) Vs (Training Loss and Validation loss)

Calculation of Accuracy, Precision and Recall:-

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

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

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

Where TP = True Positive: - Both observation and predicted value is positive
TN= True Negative: - Both observation and predicted value is positive
FP= False Positive: - Observation is negative but predicted value is positive
FN= False Negative: -Observation is positive but predicted value is negative

For Learning rate= 0.001
And Epochs= 10000

Accuracy: - 0.9649122807017544 (Calculated in the program file)
Precision: - 1.0 (Calculated in the program file)

Recall: - 0.913043478260895 (Calculated in the program file)

Conclusion:-

I classified the suspected FNA cells to Benign (class 0) or Malignant (class 1) using logistic regression in python from scratch. I trained the model using training dataset and then validated the model over validation dataset. Passed the testing dataset over the trained model and calculated the testing accuracy. I obtained 96% accuracy with this trained model. By seeing the value of accuracy, prediction and recall it is concluded that this model fits with our requirement. s

References:-

- Slides of the professor
- UCI Machine Learning repository
- https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3
- https://medium.com/@martinpella/logistic-regression-from-scratch-in-python-124c5636b8ac
- https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
- Project Description pdf
- https://towardsdatascience.com/building-a-logistic-regression-in-python-301d27367c24