CSE 574 - Introduction to Machine Learning
Project 1 Report

Ajith Kumar Natarajan
UB Person number: 50318505
Fall 2019
ajithkum@buffalo.edu

Abstract

I have built a logistic regression model to classify the suspected fine needle aspirate (FNA) of a breast mass as benign or malignant. The model was trained on 455 instances consisting of 30 features. Built with a learning rate of 0.1 and the number of epochs set as 10000, the developed model classifies at an accuracy of 98.24%.
To understand the working of the logistic regression technique (and any machine learning technique in general), the effect of number of epochs and learning rate has been studied by varying them and observing the convergence rate.

1 Introduction

Breast cancer is the most common type of cancer in female worldwide, representing nearly a quarter (25%) of all cancers and is the second leading cause of cancer deaths in women. Thus, premature discovery of the condition can lead to more efficient care thereby reducing the fatality of the disease. An advanced and reliable technique to detect the cancer can help us expedite as well as reduce the cost of cancer screening.

In this work, a simple yet powerful classification method of logistic regression has been used. No regularization technique has been used. Cross entropy was used to evaluate the loss.

2 Dataset

The dataset in used is the Wisconsin Diagnostic Breast Cancer dataset. The dataset contains 569 instances with 32 attributes (ID, diagnosis (B/M) and 30 real-valued input features). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. It has been divided into training, validation and test data in the ratio of 8:1:1.

3 Preprocessing

There were two major preprocessing steps required before using the data to build the model. The label that was used for diagnosis in the original data was 'B' for benign and 'M' for 'malignant'. Since the computer can only interpret mathematical values, it was mapped onto 0 and 1 respectively.

The second preprocessing that was required is the normalization of data. This is a common step in almost all the machine learning model building exercise. In order to avoid the effect of one particular column, having high values from inhibiting the other features, all the columns have been normalised so that the value of the columns are in the range 0-1 consistently.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4 Architecture

Since the input from the dataset consists of 30 features or attributes, the number of input nodes in the case was 30 (plus bias). There is no hidden layer. The weights were initialized randomly from a Gaussian distribution with mean 0 and variance 1. This is a sample of the computation diagram used.

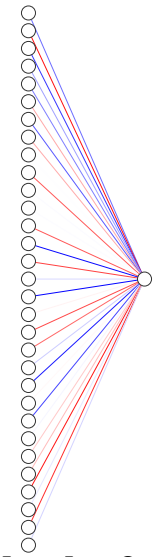

Figure 1: Architecture diagram w/o weights

5 Results

This section consists of the results obtained by implementing the model described.

Graphs have been plotted to show the variation of loss (or cost) as a function of epochs. This is done to observe the convergence of the loss with increase in epochs. As expected, even with different values for the learning rate, the shape of the graph is roughly similar. But the effect of learning rate can be clearly seen in the initial stages. Increase in the learning rate leads to steep declination of loss which then flattens out. The number of epochs is constantly kept at 10000.

5.1 Learning rate: 0.1, epochs: 10000

Figure 1 shows how the loss increases with epoch when the learning rate is set at 0.1. It can also be

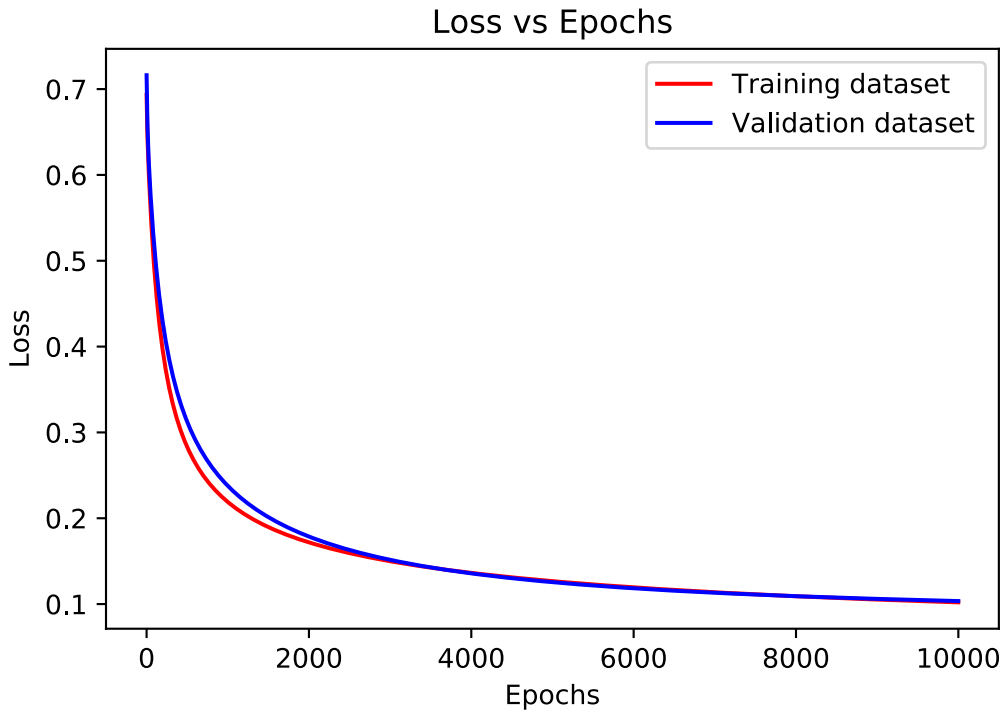

Figure 2: Loss vs epoch when learning rate is 0.1

seen that the nature of the curve produced is a very curve and decreases gradually. For this case, the observed highest loss for the training dataset was 0.7324 and the lowest was 0.1036. Whereas for the validation data, the highest loss in training data was observed to be 0.6823 and the lowest was 0.1019. With a learning rate of 0.1, the accuracy, precision and recall values that have been obtained are 0.9824, 1.0 and 0.9729 respectively

5.2 Learning rate: 1, epochs: 10000

Figure 2 shows the same details, but the learning rate here is 1. It can be seen that initially the loss decreases at a faster rate but after a certain period, the loss start to increase rather than continuing

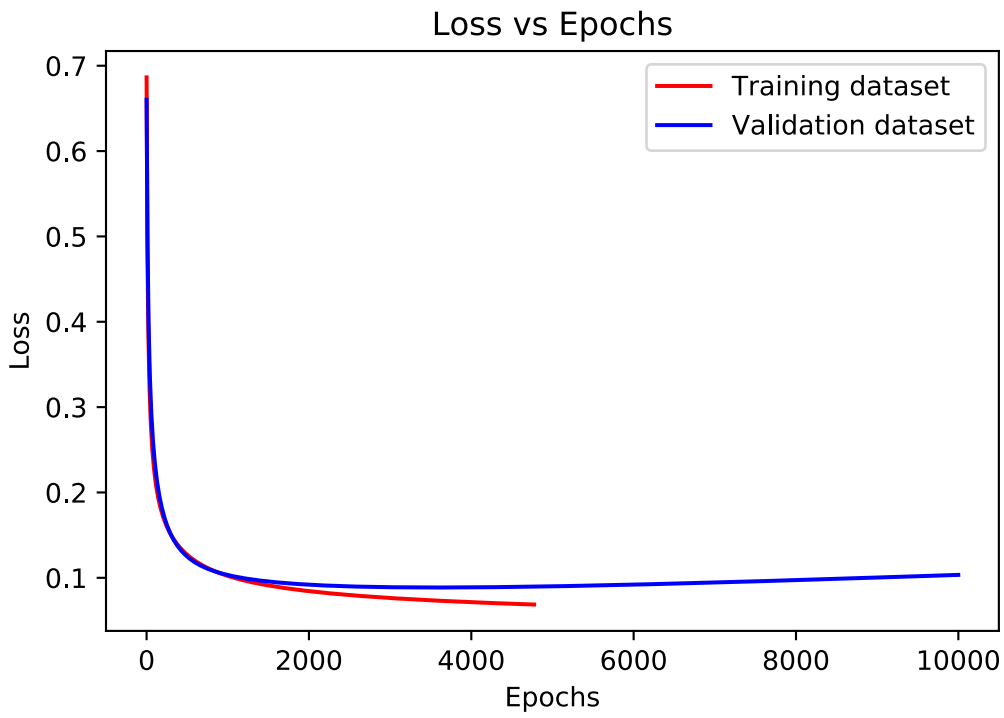

Figure 3: Loss vs epoch when learning rate is 1

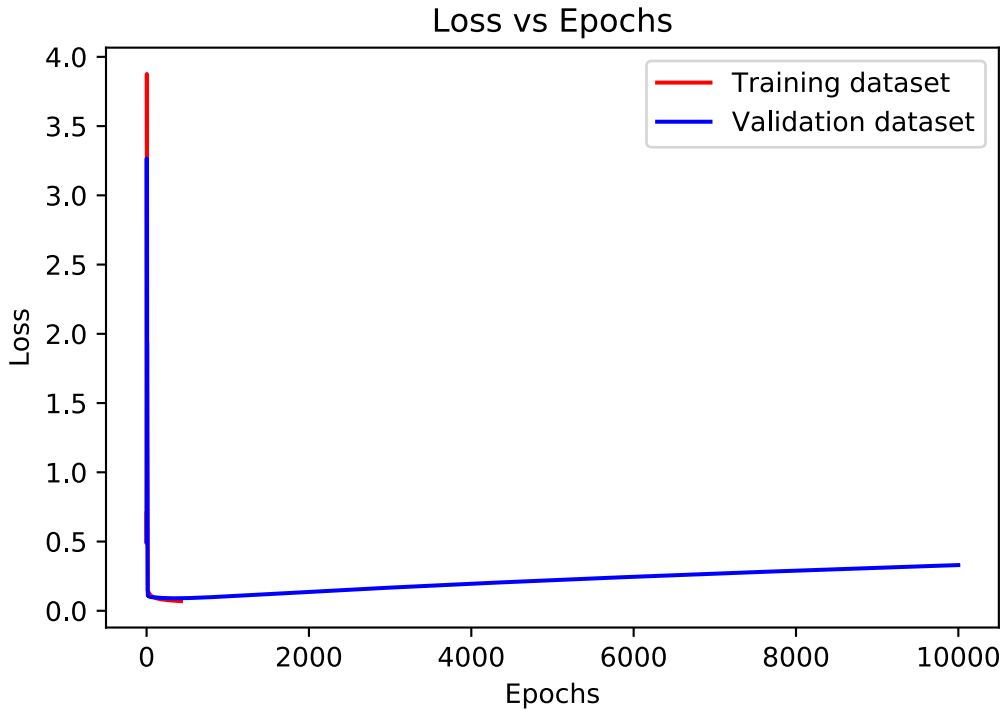

Figure 4: Loss vs epoch when learning rate is 10

Table 1: Confusion matrix

Confusion matrix	
36	0
1	20

to decrease as it starts diverging out. In this case, the highest observed loss in the training dataset was 0.6768 and the lowest was 0.12. The corresponding values were 0.6996 and 0.522 in the case of validation data. The accuracy, precision and recall values that have been obtained are 0.9122, 1.0 and 0.8780 respectively.

5.3 Learning rate: 10, epochs: 10000

The final study was made by setting the learning rate to 10 and retaining the number of epochs as 10000.

The initial decrease of loss is even more sharper in this case, and is almost perpendicular. Like the previous case, here also the model starts diverging subsequently and so the loss starts increasing. While the accuracy, precision and recall values obtained are 0.9122, 1.0 and 0.8780 like in the previous case, the observed maximum loss was 3.88 in the case of training data and 3.021 in the case of validation data.

5.4 Variation of number of epochs

With a variation in the number of epochs at a constant learning rate (say 0.1), it is observed that the convergence, though it slows down with increase in the epoch count, it decreases continuously before it saturates at a point. Less number of epochs means that the time required to train is less but the convergence and hence the accuracy of the model may not be good enough.

5.5 Confusion matrix

Refer Table 1 for the confusion matrix obtained for a learning rate of 0.1 and epoch count of 10000. It can be seen that the number of false negatives is only 1.

6 Conclusion

A logistic regression model was built for the given dataset. An accuracy of 98.24% was achieved. It was also observed that while a low learning rate might slow the process of convergence, a high learning rate but not at all and end up diverging. Another hyper-parameter, the number of epochs determines the extent to which the model must converge or that how low the loss should be. It should be decided as a trade off between the time for obtaining the parameters of the model and the accuracy.

References

[1] Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006.

[2] Srihari, Sargur N. Lecture Slides for Machine Learning