Logistic Regression

Wisconsin Diagnostic Breast Cancer Data Set

Snigdha Gupta
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
snigdhag@buffalo.edu

Abstract

Predicting if a cell has grown a tumor can help save a patient's life considerably. The project aims to utilize logistic regression to predict whether a breast cell is malignant or not. The model classifies a cell as class 1 (malignant) or class 0 (benign) by reducing the loss, adjusting weights and biases to reach optimum minima. A practical implementation has been performed to better understand the problem and verify the results.

1 Introduction

1.1 Supervised Learning – Regression and Classification

Supervised learning refers to training a model when we already know what is the data set and how our correct output should look like. We already have an idea about the relationship between the input and the output. Supervised learning can be of two type – Regression and Classification.

1.2 Linear Regression

In order to find a linear relationship between target and one or more predictors, Linear Regression is used. In regression, we try to map the input to a continuous function. This means that we try to predict the result within a continuous output. For example, consider a scenario to predict the price of a house based on a data set with the size of houses on the real estate market. Here we train the model to predict the price of a house, when it's size and features are known keeping in mind the real estate market.

The Hypothesis Equation (Eq 1):

$$h_\theta(x) = W_0 + W^T * x$$

The hypothesis equation for Linear Regression tries to draw a straight line through the data point such that the sum of squared errors is minimized i.e. we reach the local minima.

The Cost Function (Eq 2):

$$J(W_0, W) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$$

However, consider that we want the model to predict whether a house is an apartment or not? This can be done using linear regression, but the hypothesis function does not take into account many factors for classification tasks. Hence the predicted output might not be accurate and therefore, we have the need for another hypothesis function to direct this problem.

1.3 Logistic Regression

Logistic regression hypothesis function models the probabilities. Since probabilities are continuous, we call it logistic **regression.** Furthermore, if a decision boundary is selected then we use logistic regression as a **classifier**.

Therefore, logistic regression is a classification algorithm that allows us to assign observations to discrete classes. In the previous example of houses, we discussed that to predict whether a house is an apartment or not will be a classification problem. Similarly, for predicting whether a breast cell is malignant or not we use logistic regression along with a decision boundary.

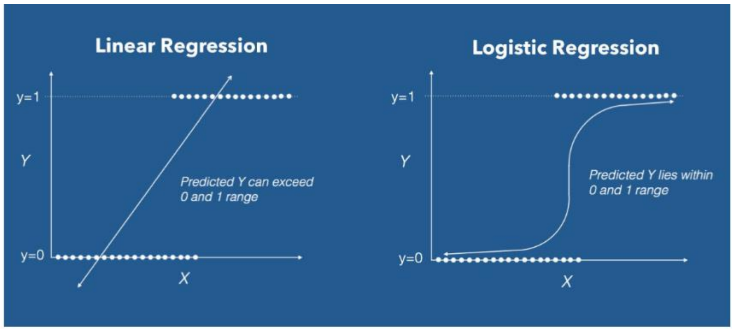


Figure 1 Linear Regression VS Logistic Regression Graph | Image: Data Camp

In figure 1, we see that the predicted value can exceed a range when using Linear Regression. However, when using Logistic Regression, we get the predicted values within a range. This is because the hypothesis equation of logistic regression uses the 'Sigmoid Function'. Let us first understand the hypothesis equation of logistic regression.

The Hypothesis Equation (Eq 3):

$$h\theta(x) = \sigma(W_0 + W^T * x)$$

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

Now, if we consider the loss function of linear regression (Eq 2), then we will get a non-convex function with many local minima which will be of no use. Hence, minimizing the cost function to find the local minimum will be a difficult task, as represented in figure 2.

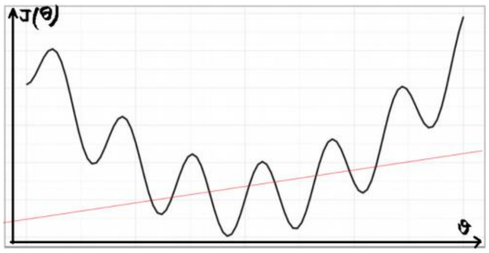


Figure 2 Non-Convex Function

This brings Decision Boundary into the picture. We pass our hypothesis through a prediction function (sigmoid function) and return a probability score between 0 and 1. We decide a threshold value above which we classify the values into Class 1 and if the values go below the threshold, we classify the value in Class 2.

For the WDBC data set we have set the threshold value to 0.5. So, let's say if our prediction returns a value of 0.88, then we classify it to class 1 and if our prediction returns a value of 0.22 then we classify it to class 0.

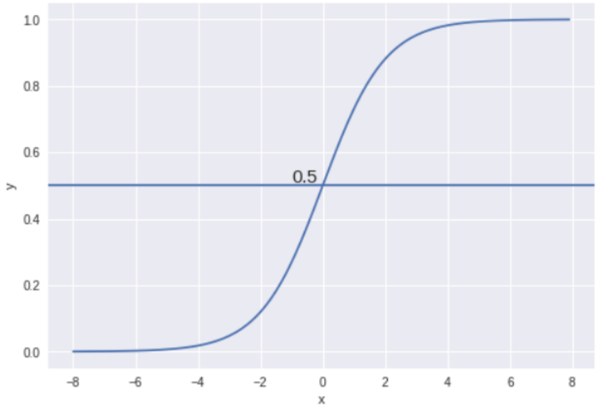


Figure 3 Sigmoid function as a decision boundary

For the above reasons we now define the cost function for logistic regression. If we assume our class density to be Gaussian and apply Bayes' Theorem, we can prove the function to be sigmoid.

The two rules of probabilities can be defined as:

Sum Rule (Eq 4):

$$p(X) = \sum_{Y} p(X,Y)$$

Product Rule (Eq 5):

$$p(X,Y) = p(Y|X).p(X)$$

Symmetry Property (Eq 6):

$$p(X,Y) = p(Y,X)$$

Now, using equation 4, 5 and 6:

$$p(X,Y) = p(Y,X)$$

$$p(Y|X).p(X) = p(X|Y).p(Y)$$

$$p(Y|X) = \frac{p(X|Y).p(Y)}{p(X)}$$

$$p(Y|X) = \frac{p(X|Y).p(Y)}{\sum_Y p(X,Y)}$$

$$p(Y|X) = \frac{p(X|Y).p(Y)}{\sum_Y p(X|Y).p(Y)}$$

The deduced equation is called the *Bayes' Theorem* (Eq 7):

$$p(Y|X) = \frac{p(X|Y).p(Y)}{\sum_Y p(X|Y).p(Y)}$$

For the given breast cancer problem, we have two classes – Malignant (Class 1) and Benign (Class 2). Therefore, using equation 7 for C_1 and C_2, we get:

$$p(Y|X) = \frac{p(X|Y).p(Y)}{p(X|C_1).p(C_1) + p(X|C_2).p(C_2)}$$

Dividing by numerator:

$$p(Y|X) = \frac{1}{1 + \frac{p(X|C_2).p(C_2)}{p(X|C_1).p(C_1)}}$$

Using natural log:

$$p(Y|X) = \frac{1}{1 + e^{-ln\frac{p(X|C_2).p(C_2)}{p(X|C_1).p(C_1)}}}$$

$$z = ln\frac{p(X|C_2).p(C_2)}{p(X|C_1).p(C_1)}$$

Hence, the sigmoid activation deduced is (Eq 8):

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

We can say that since p(Y|X) is linear, similar values of x will have similar probabilities. Hence (Eq 9):

$$z = ln\frac{p(X|C_2).p(C_2)}{p(X|C_1).p(C_1)} = W_0 + W^T * x$$

The cost function for logistic regression is defined as:

$$-\log\left(h_\theta(x), if\ y = 1\right.$$

$$-\log\left(1 - h_\theta(x), if\ y = 0\right.$$

Taking the summation, we get the Cost Function (Eq 9):

$$J(W_0, W) = -\frac{1}{m}\sum y^i \log(h_\theta x^i) + (1 - y^i)\log(1 - h_\theta x^i)$$

2 Data Set

2.1 Breast Cancer

What is breast cancer? It starts when cells in the breast begin to grow out of control. The cells grow into a tumor. A tumor can often be seen in an X-ray or felt as a lump. The tumor is malignant (cancer) if the cells can invade into surrounding tissues or metastasize to distant areas of the body. Breast cancer occurs almost entirely in women, but men can get breast cancer, too.

Breast cancer is the most common cancer amongst women. It is also the second most common type of cancer. In less developed countries, it is the leading cause of cancer death and it is the second leading cause of cancer death amongst American women which is exceeded by lung cancer. Approximately 1.7 million cases of breast cancer were recorded in 2012. 1 in 4 women have breast cancer around the world. In the United States, there are more than 3.5 million breast cancer survivors, including women who are still being treated and those who have completed the treatment.

It is estimated that in 2019, there will be 271,270 new cases of invasive breast cancer diagnosed in women and 2,670 cases diagnosed in men.

2.2 Wisconsin Diagnostic Breast Cancer (WDBC)

The Wisconsin Diagnostic Breast Cancer data set [2] or WDBC set contains 569 instances with 32 attributes (ID, diagnosis, 30 real-valued input features). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes. The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in [1].

The attribute information is as follows:
1) ID number
2) Diagnosis (M = malignant, B = benign)
3) Computed features describe the following characteristics of the cell nuclei present in the image:

- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter^2 / area - 1.0)
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

Several papers have been published that contain detailed descriptions of how these features are computed.

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.

Missing attribute values: none
Class distribution: 357 benign, 212 malignant

3 Preprocessing

The transformation applied to the data before feeding it to the algorithm is called pre-processing. This is required to convert the raw data into a clean data set, since the raw data is not feasible for the analysis.

The data set comprises of 32 columns. The first column, i.e. the ID Column and the second column is the Target Vector column. The target vector column comprises of M – malignant and B – benign values. The following steps were carried out while processing the data:

1) ID column is dropped from the data set.
2) Each column is given a name or label. Columns in this data set are named as col_1, col_2 and so on till col_31.
3) A total of 569 samples with 31 features is used as a data set. The values in the target vector column are mapped to 0 and 1. If value is 'M' its mapping is equivalent to 1 and if value is 'B' its mapping is equivalent to 0.
4) The data set is split into three parts – Training, Validation and Test data sets. 80% of data is used as Training set (X_train, Y_train). 10% data is used as Validation set (X_validate, Y_validate) and the rest 10% data is used as Test set (X_test, Y_test).

   The X terms contain 30 features – this means that the target column is not included. The Y terms contain only the target column.

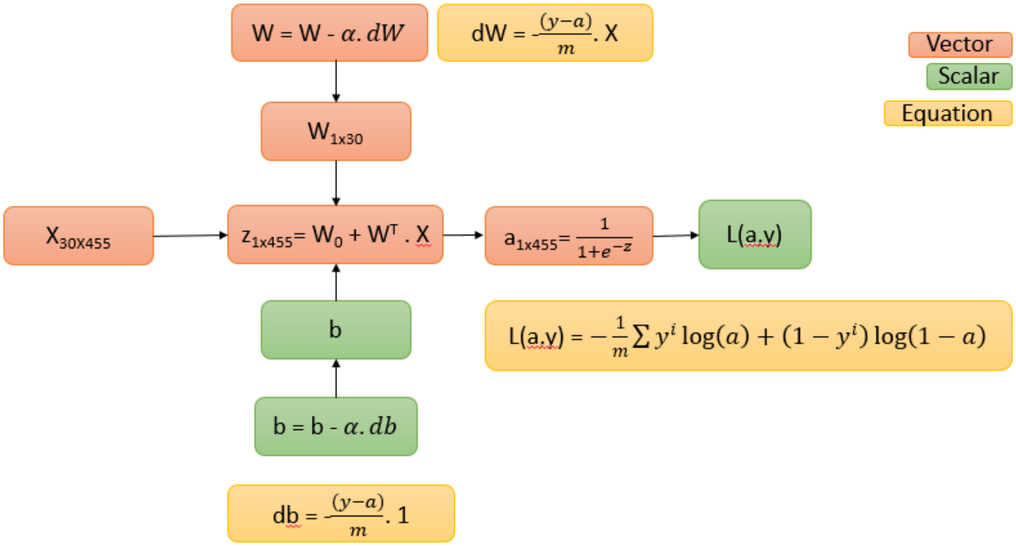
5) Normalize the data set using StandardScaler. It transforms the attributes with Gaussian distribution so that each value has a mean of 0 and a standard deviation of 1.
6) Initialize the weights and biases to a dummy value. We initialize the weight vector

weight to 0 and bias b to 0. The learning rate is also initialized to an arbitrary value.

7) Set number of epochs to 10000

4 Architecture

The computational graph is designed below:



$W = W - \alpha . dW$
$dW = -\frac{(y-a)}{m} . X$
Vector
Scalar
Equation
W_{1x30}
X_{30X455}
$z_{1x455} = W_0 + W^T . X$
$a_{1x455} = \frac{1}{1+e^{-z}}$
L(a,y)
b
$L(a,y) = -\frac{1}{m}\sum y^i \log(a) + (1 - y^i)\log(1-a)$
$b = b - \alpha . db$
$db = -\frac{(y-a)}{m} . 1$

Figure 4 Computational Graph of Logistic Regression

5 Results

The following results were obtained:

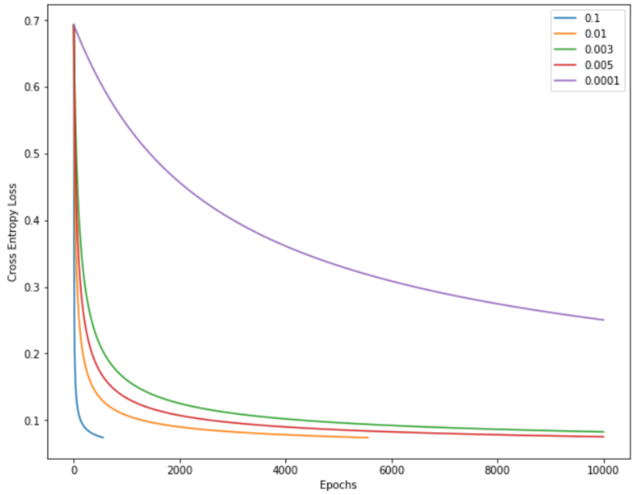


Figure 5 Cross Entropy Loss vs Epochs (hyperparameter)

In figure 5, we observer that for learning rate 0.1 the graph converges rapidly and hence we are unable to reach our optimum minima. Furthermore, we observe that when learning rate is 0.0001, the graph converges too slowly. This again restricts us to reach the optimum minima for the problem. Hence through a series of hit and trial, the best fir for hyperparameter is when learning rate is 0.003 and number of epoch is 10000.

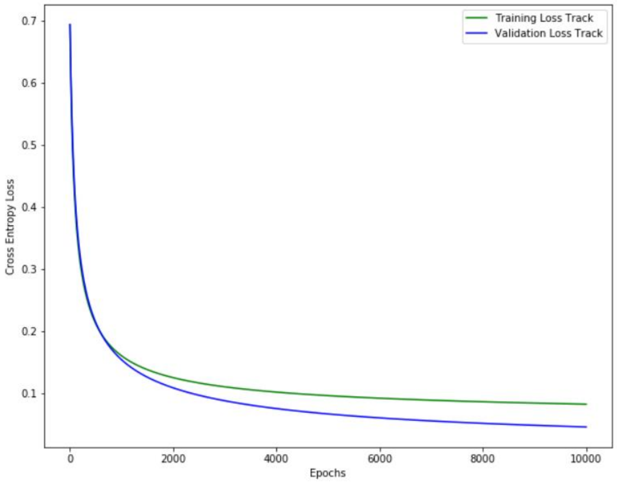


Figure 6 Cross Entropy Loss vs Epochs (10000), Learning Rate 0.003

With a learning rate of 0.003 and 10000 epochs we see in figure 6 that the model converges to the optimum minima successfully over the training set. We also observe that since the validation curve is below the training curve, the model does not overfit and hence successfully predicts the result for unseen data.

The evaluation metric is as follows:

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

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

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

In figure 7, we compare the accuracies of training and validation set and the accuracy of validation is more than that of training. This means that the model is not overfitting.

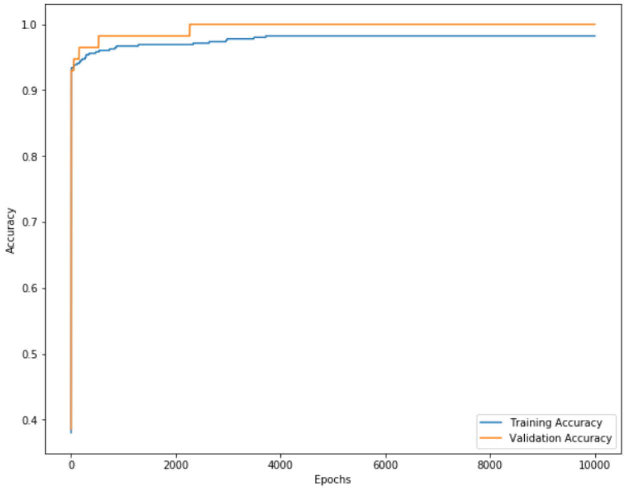


Figure 7 Accuracy vs Epochs

6 Conclusion

It can be concluded from our results that logistic regression can be used to classify breast cells as malignant or benign efficiently. After training on approximately 450 samples, we were able to create a model that classified cells with an accuracy of 98.25%, precision – 100% and recall 94.12%. With a greater number of samples, the model can be made even more robust.

References

[1] K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992

[2] https://data.world/health/breast-cancer-wisconsin/workspace/file?filename=DatasetDescription.txt

[3] https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[4] https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

[5] https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

[6] https://www.bcrf.org/breast-cancer-statistics-and-resources

[7] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[8] https://d1b10bmlvqabco.cloudfront.net/attach/jzsomflkppr3r4/isamd3soc56z/k0oa54u3szi5/Equation.pdf

[9] Andrew ng Machine Learning Course: https://www.coursera.org/learn/machine-learning/home/welcome

[10] https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/