Logistic Regression Based Breast Cancer Prediction

Vineeth Thayanithi
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
vthayani@buffalo.edu

Abstract

Breast cancer is a common type of cancer that can affect women and sometimes men. There are over 200,000 cases of breast cancer per year in the United States alone. There are several symptoms that could indicate the presence of cancer cells in the case of breast cancer. These symptoms although not definitive will lead us to an idea of whether a person has been affected by breast cancer or not. Using these symptoms and Logistic Regression, we could build a machine learning model that could help us predict the likelihood of presence of cancerous cells.

1 Introduction

Breast cancer is the most common type of cancer that occurs amongst women in the united states. It occurs when few cells present in the breast starts to grow abnormally. Cancer cells usually start growing in the milk producing ducts known as invasive ductal carcinoma or invasive lobular carcinoma or within other tissues of the breast. These cells accumulate as solid lumps in that area by dividing at a rate faster than the healthy cells. These cells may then spread to other parts of the body.

Doctors estimate that about 5 to 10 percent of breast cancers are linked to gene mutations passed through generations of a family. It has been identified that several environmental, hormonal and lifestyle factors tend to increase the risk of being affected by breast cancer.

There are several signs and symptoms that can be used for identifying the presence of cancerous cells. Using these signs and symptoms and the records of previous patients who have been diagnosed for breast cancer, we can build a machine learning model using logistic regression that would help us classify the patient to have or not have breast cancer. Although the results from the model cannot be certain at all times, it can give insights to radiologists which can lead to a better diagnosis.

2 Dataset

The dataset used for training the model is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The data set contains around 569 records with 32 attributes for each record. 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:

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 (perimeter*2/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)

3 Data Preprocessing

The dataset is imported as a data frame using pandas in python 3. The diagnosis in the dataset is denoted by either “M” for Malignant or “B” for Benign. This label cannot however be used for your model. Hence, we convert it into a Boolean state with 1 denoting the presence of Malignant cells and 0 denoting Benign. The columns 3 to 33 are extracted as features that are used for predicting the target values that are present in column 2. The 30 features that were extracted contains ranges that are specific to each of these features. This could cause unintended training results as one trivial feature with a very large value can override a significant feature whose range is small. To avoid such problems, the dataset is normalized.

To normalize the data, we use the following.

$$x = \frac{x - \text{Minimum}(x)}{\text{Maximum}(x) - \text{Minimum}(x)}$$

Once the dataset undergoes normalization, the data is randomly split into training, test and validation sets. 80 percent of the data is reserved for training the data, 10 percent for testing the model and 10 percent to the validation set for tuning the hyper parameters. The randomness in the dataset can sometimes change the precision of the model.

4 Architecture

Logistic Regression is a machine learning model that is used for classification-based problems. The name “Regression” is used to indicate the although the model solves classification-based problems, the underlying strategy that’s used in the model is same as linear regression. Logistic regression is best used when there is a need to classify an instance into one of the two specified problems. Logistic Regression can also be used to solve a multi-class classification problem but however, it would require a lot of computational and storage power.

The basis function of logistic regression takes the form

$$h(x) = \sigma\,(w^T x + b)$$

Logistic regression uses the sigmoid function (logistic function) which is an S Shaped curve that returns a value between 0 and 1 and never the limits. The logistic function can be written as:

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

Where,

$\sigma(z)$ - Sigmoid function
e - Base of nature log
Z - Input to the function

The weights are added to each of the features of the data and these weights keeps altering until the model has finished learning. These weights are what that determine the outcome of the target values. The bias term is introduced to determine the default log odds when all of the features have the value 0.

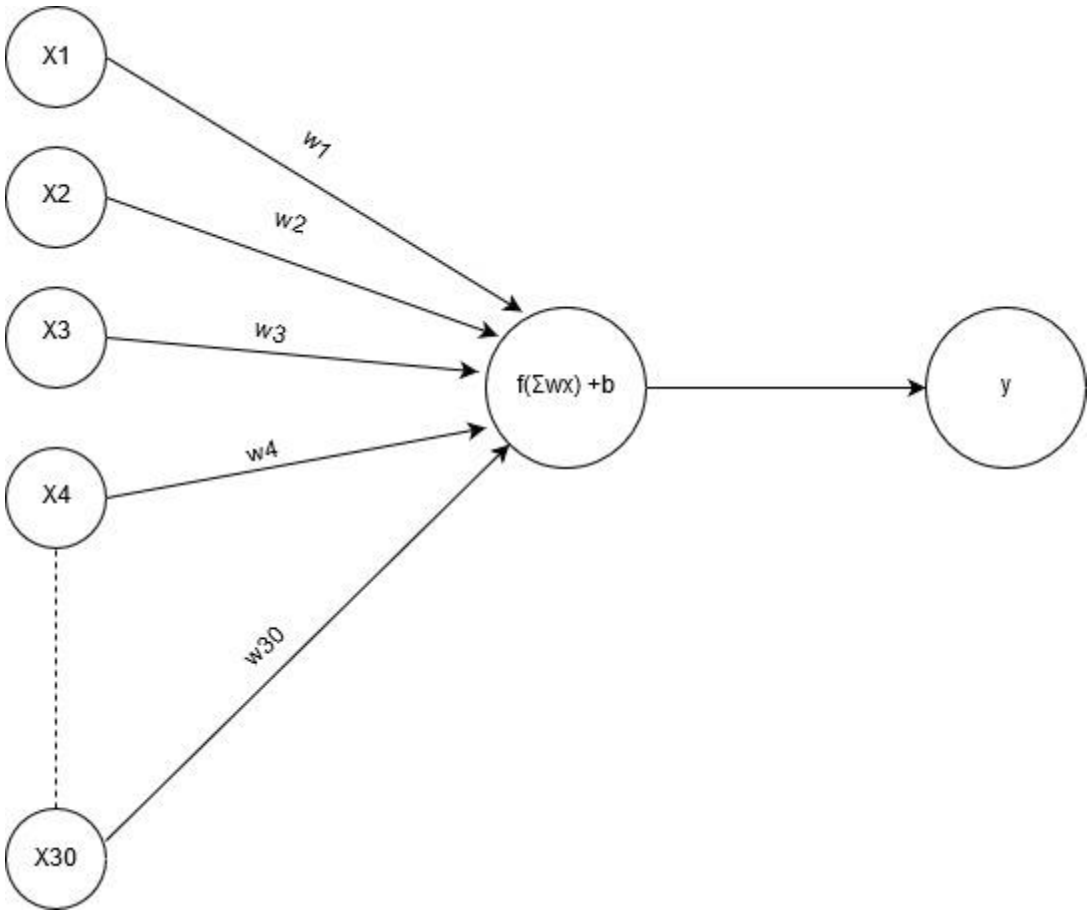

Fig 4.1 Architecture diagram

Once we arrive at a probability value, we move on to find the cost of the model which is otherwise known as the loss. We use cross- entropy to determine the cost. There exists two cost functions, when y=1 and y=0. Fortunately, these functions can be combined and written together.

$$J(\theta) = \frac{1}{M}\sum \text{cost}(h, y)$$

$$\text{Cost}(h, y) = -\log(h(x)) \quad \text{if } y=1$$
$$\text{Cost}(h, y) = -\log(1-h(x)) \quad \text{if } y=0$$

Combined Cost,

$$J(\theta) = -\frac{1}{M}(y \log(h(x)) - (1-y) \log(1-h(x)))$$

Similar to our linear regression problems, we use gradient descent in order to minimize the cost by moving in the direction of steepest descent as defined by the negative of the gradient. Gradient descent is the strategy that we use to update the weights and the bias introduced.

$$C' = (s(z) - y)$$

$$W = W - \eta C'$$

$$B = B - \eta C'$$

is the learning rate which determines the rate at which the gradient descent proceeds with a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing.

5 Results

To perform evaluations on the resultant model. We consider three metrics, 1. Accuracy, 2. Precision and 3. Recall. Generally, Accuracy is defined as the total number of test results that our model predicted right. However, In order to determine the metrics using the formal definition we need to construct what is called as the confusion matrix. The confusion matrix is a 2x2 matrix with its elements being number of True positives, False positives, False negatives and True negatives for (0,0), (0,1), (1,0) and (1,1) respectively. The metrics are formally defined as below:

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

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

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

Where,

TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

With Learning rate set to 0.06 and the number of epochs being 1000, the model returns an accuracy of 98%, a precision of 97 % and a 100% recall on the test set. A graph is plotted for loss against epochs and in that we can observe a gradual reduction in the loss function as the epochs reach towards 1000.

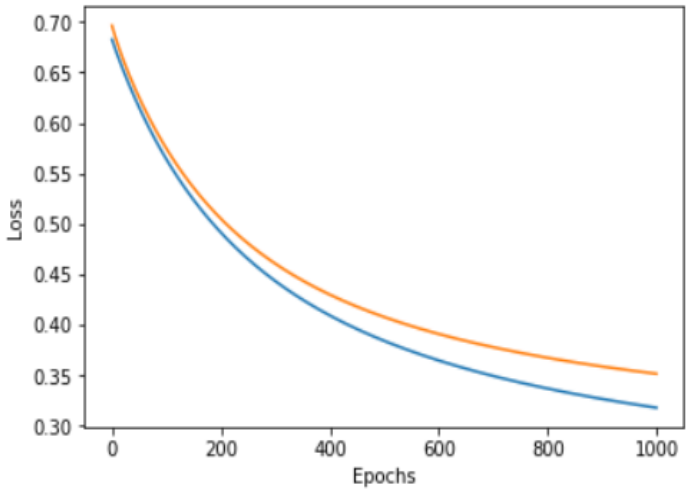

Fig 5.1 Loss versus Epochs 1

In the figure above, the yellow and blue curves denote the loss vs epochs for the validation set and train set respectively. It is also evident from the graph that with the learning rate of 0.01 and 100 epochs, the model reaches a very bad accuracy of 28%.

```
Accuracy of the Model is: 29.82456140350877
Precision of the Model is: 100.0
Recall of the Model is: 11.11111111111111
```

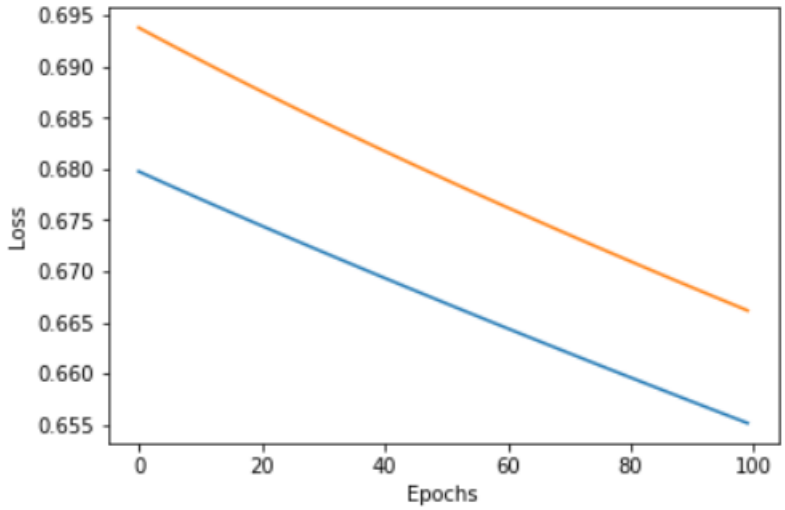

Fig 5.2 Loss versus Epochs 2

6 Conclusion

Logistic Regression was performed on the Wisconsin Diagnostic Breast Cancer Dataset. The basic idea of this study was to classify between two groups having found their probabilities of occurrences in those groups. This approach led us to a 98% accuracy with the given dataset. Further diagnosis using this data would provide very useful insights to the radiologists trying to diagnose breast cancer.

7 References

[1] Logistic Regression, https://ublearns.buffalo.edu/bbcswebdav/pid-5108334-dt-content-rid-25438520_1/courses/2199_23170_COMB/4.3.2-LogisticReg.pdf

[2] Hypothesis Representation, https://www.coursera.org/lecture/machine-learning/hypothesis-representation-RJXfB

[3] Train/Test Split and Cross Validation in Python, https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

[4] Breast Cancer Wisconsin (Diagnostic) Data Set, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[5] What is the Difference Between Test and Validation Datasets?, https://machinelearningmastery.com/difference-test-validation-datasets/