Breast Cancer Classification using Logistic Regression

Akshara Santharam
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
aksharas@buffalo.edu

Abstract

Considerable benefits will be provided to the medical system, when machine leaning algorithms can automatically identify and classify the cancer cells. The task of this project is to perform classification using machine learning for a two-class problem. The features to be used for classification are obtained from images of a fine needle aspirate (FNA) of a breast mass. In this project, the classification of suspected FNA cells to Benign (Class 0) and Malignant (Class 1) is complemented by Logistic Regression. The dataset in use is the Wisconsin Diagnostic Breast Cancer which is collected from the UCI Machine Learning Repository. This dataset contains 569 instances with 32 attributes and will be used for training, testing and validation. After training the model using logistic regression, the model's performance can be tested on the test data by calculating accuracy, precision and recall.

1 Introduction

Breast cancer is the cancer that is formed in the cells of the breasts and is an increasing public health problem. It is one of prime source of cancer among women. Many lives can be saved in an effective manner if the breast cancer is early detected. The incidence is rising in most countries and is projected to rise further over the next 10 years despite current efforts to prevent the disease [1].

Breast cancer can be classified into three different categories such as benign breast cancer, invasive breast cancer and in-situ cancers. Benign breast cancer are the unusual growths in the breast tissue that are non-cancerous. Invasive breast cancer occurs when cancer cells break out into the nearby breast tissue. In-situ breast cancer is non-invasive and has a low risk of becoming invasive [2].

Machine learning is used to develop classification and prediction models that can predict outcomes in cancer patients. It can predict interpretations and conclusions which could not be made using standard statistical approaches [3]. Logistic Regression is a statistical method that is used for analyzing data and explaining the relationship between one dependent variable and one or more independent variable. Logistic regression is one of the common models for breast cancer classification and prediction [4]. Since, logistic regression can directly predict probabilities unlike linear regression, logistic regression is used to classify the breast cancer to Benign and Malignant.

In this project, Wisconsin Diagnostic Breast Cancer (WDBC) dataset is first collected and processed into the Pandas Dataframe. After assigning the features to the data, the data is preprocessed and normalized. The normalized data is then partitioned into training, validation and testing, where 80% of the data is the training data, 10% of the data is the validation data and the rest 10% of the data is the testing data. Training of the data is done using the Gradient Descent for Logistic Regression using a group of hyperparameters. Then, the regression performance of the model is validated on the validation data. Finally, the model's performance is tested on the test data and is measured by calculating accuracy, precision and recall. The validation and training accuracy graph vs epochs & validation and training loss graph vs epochs are plotted.

2 Dataset

In this project, we use the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from UCI Machine Learning Repository. The Wisconsin Breast Cancer dataset will be used for training, validation and testing. The dataset contains 569 instances with 32 attributes. The features are obtained from a digitized image of a fine needle aspirate (FNA) of a breast cancer mass. The attributes present in the Wisconsin Breast Cancer dataset are:

- ID
- Diagnosis (B/M)
- Radius Mean
- Texture Mean
- Perimeter Mean
- Area Mean
- Smoothness Mean
- Compactness Mean
- Concavity Mean
- Concave Point Mean
- Symmetry Mean
- Fractal Dimension Mean
- Radius Standard Error
- Texture Standard Error
- Perimeter Standard Error
- Area Standard Error
- Smoothness Standard Error
- Compactness Standard Error
- Concavity Standard Error
- Concave Point Standard Error
- Symmetry Standard Error
- Fractal Dimension Standard Error
- Radius Worst
- Texture Worst
- Perimeter Worst
- Area Worst
- Smoothness Worst
- Compactness Worst
- Concavity Worst
- Concave Point Worst
- Symmetry Worst
- Fractal Dimension Worst

3 Data Preprocessing

Data preprocessing is an important step in the data mining process that involves transforming raw data into an understandable format suitable for several classifiers to perform classification. Real world data is dirty, incomplete and inconsistent and it most likely to contain many errors. Data preprocessing can be achieved by various means like data reduction, data cleaning, data integration, data transformation and data discretization. Data cleaning is the process of detecting and correcting the corrupt data. Data integration is the process of combining data from other resources and providing users with a unifies view. Data transformation involves converting data from one form or structure to another. Data discretization is the process of converting a large data into a smaller one so that data management and evaluation becomes easier.

In this project, the most important step if to identify and remove null values from the dataset. If empty values are found, the dataset is updated by deleting all the rows where empty values are present. Data Normalization is performed to organize the data in accordance to normal forms which can help reduce data redundancy and improve data integrity.

4 Architecture

Logistic Regression follows the equation Y = WX + b, where X is the input array and W represents the weights and b represents the bias. Logistic Regression uses an activation function called sigmoid function to perform classification and prediction. Sigmoid equation can be calculated by using the following formula:

$$Ø = 1/ (1 + (e \wedge -z))$$

Y is then passed to the sigmoid function. Forward propagation is performed by assigning and feeding weights and bias into the computational graph. Then, Loss function is calculated by

$$h = g\,(X\,Ø)$$

$$J(Ø) = 1/m.\ (y^{T} \log(h) - (1 - y)^{T} \log(1-h))$$

We need to update weights and bias, to decrease cost. The technique gradient descent is used to make our model learn the parameters which can decrease cost function like weights and bias.

$$\delta\, J(Ø) / \delta\, Ø_j = 1/m\ X^{T}\,(g\,(X\,Ø) - y)$$

Computational graph of logistic regression shows that the input array is multiplied with the weights and then is sent to the Sigmoid function.

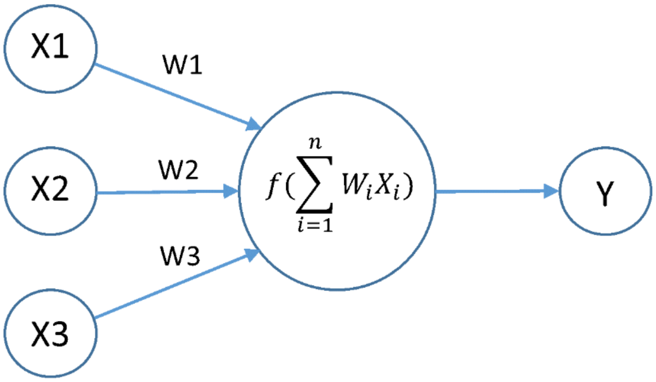

Figure 1: Computational graph of Logistic Regression [5]

In the above figure, X1, X2, X3 are the input arrays and W1, W2, W3 are the weights.

5 Result

The task of the project is to classify the fine needle aspirate (FNA) to Benign (Class 0) and Malignant (Class 1) using logistic regression as the classifier and plot training and validation accuracy vs epochs & training and validation loss vs

Evaluation Metrics: We have evaluated the model using Accuracy, Precision and Recall.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = (TP) / (TP + FP)$$

$$\text{Recall} = (TP) / (TP + FN)$$

Where, TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negatives.

Using Logistic Regression, the accuracy, precision and recall of the model is 96.4%, 97.01% and 97.05% respectively, where TP = 33, FN = 1, FP = 1 and TN = 22.

Table 1: Confusion Matrix

TP = 33	**FN = 1**
FP = 1	**TN = 22**

The model is evaluated by plotting graph of training data accuracy and validation data accuracy & Training loss vs Validation loss against 15000 epochs with learning rate as 1.

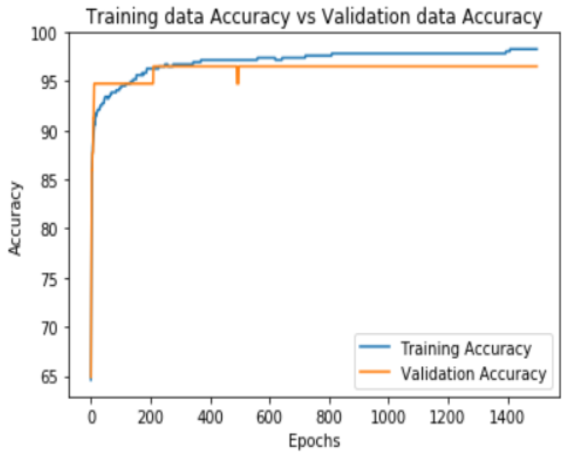

Figure 2: Training data vs Validation data Accuracy

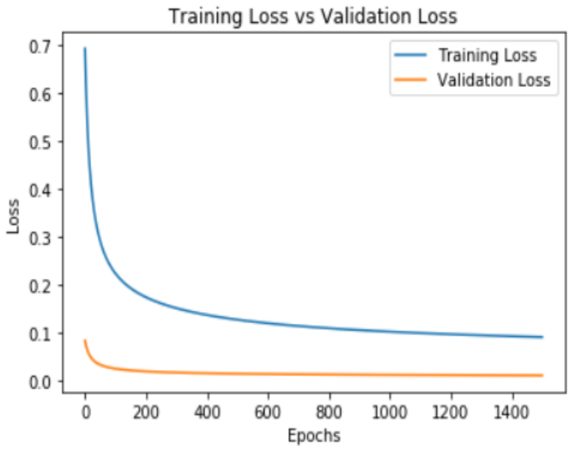

Figure 3: Training loss vs Validation loss

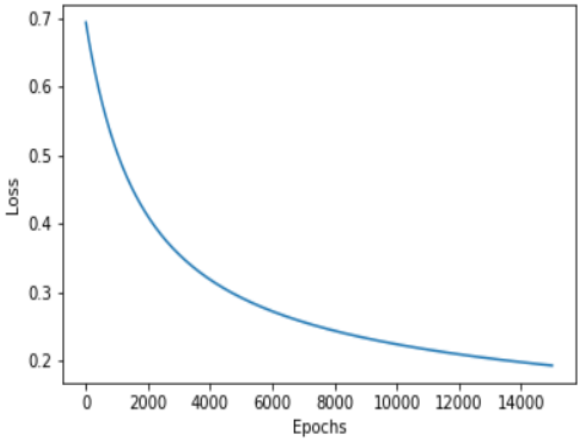

Figure 4: Training loss when learning rate is 0.05

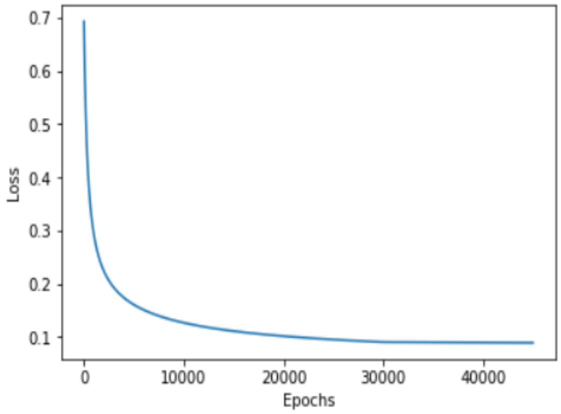

Figure 5: Training loss when learning rate is 0.005

6 Conclusion

To classify the fine needle aspirate (FNA) cells to Benign (Class 0) and Malignant (Class 1), Linear Regression is used. The model's performance is tested on the test data and is measured by calculating accuracy, precision and recall. The accuracy, precision and recall achieved by the model is 96.4%, 97.05% and 97.05% respectively.

References

[1] Anthony Howell, Annie S Anderson, Robert B Clarke, Stephen W Duffy, D Gareth Evarc, Montserat Garcia - Closas, Andy J Gescher, Timothy J Key, John M Saxton & Michelle N Harvie (2014) Risk determination and prevention of breast cancer, *BMC Part of Springer Nature*, pp. 1-19.

[2] Jabeen Sultana & Abdul Khader Jilani (2018) Predicting Breast Cancer Using Logistic Regression and Multi- Class, *International Journal of Engineering & Technology*, pp. 22-26.

[3] Cruz JA, Wishart DS (2006) Application of Machine Learning in Cancer Prediction and Prognosis, *Departments of Biological Science and Computing Science, University of Alberta Edmonton, AB, Canada.Vol.2*, pp. 2-21.

[4] H. Yusuff, N. Mohamad, U.K.Ngah & A.S.Yahaya (2012) Breast Cancer Analysis using Logistic Regression. *Vol.10*, pp. 14-22.

[5] https://medium.com/@martinpella/logistic-regression-from-scratch-in-python-124c5636b8ac