Logistic Regression

Akash Sharma[*]
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
as475@buffalo.edu

October 4, 2021

Abstract

Diabetes is very common disease nowadays. It is a additional inventor to other disease. To diagnose it, patient needs to visit the clinic or hospital. This Project aim to predict whether the female patient have the diabetes by observing the other features like blood sugar, pregnancies etc. This project will predict diabetes of the female patient with the help of logistic Regression.

1 What is Logistic Regression?

Logistic Regression predict data by analysing its features and outcomes. It is the algorithm that use in supervised problem which results in 0 or 1, yes or no.On the contrary to the name it is a classifier because it classify the data according the probability of the outcomes.

2 Logistic Regression Overview

2.1 Formulas for Logistic Regression

$$Z = W^T * Dia\_x\_train + b$$

$$A = 1/(1 + e^{-Z}))$$

$$Cost(w, b) = \sum_{n=1}^{m}$$

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

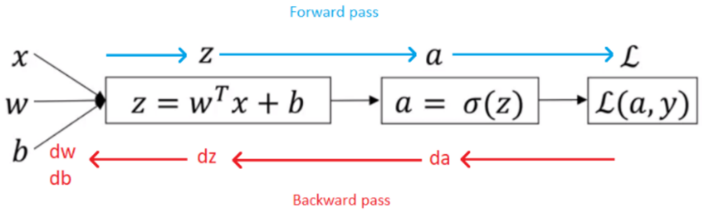


Figure 1: Logistic regression

$$y_n log(predicted_n) + (1 - y_n) log(1 - predicted_n) \frac{-1}{m}$$

$$dZ = A - Y$$

$$= X * dZ^T \frac{1}{m}$$

$$W = W - (learningrate) * dW$$

$$b = W - (learningrate) * db$$

3 Data Set and its features

Pima Indians Diabetes Database dataset will be used for training, and testing. The dataset contains medical data of female patients above the age of 21 and 768 instances with the diagnostic measurements of 8 features. The 8 features are as follows:
1 Glucose (Blood Glucose level)
2 Pregnancies (The number of pregnancies the patient has had)
3 Blood Pressure(mm Hg)
4 Skin Thickness(Triceps skin fold thickness (mm))
5 Insulin level
6 BMI (Body Mass Index : weight in kg/(height in m)2)
7 Diabetes Pedigree Function
8 Age (In years)

4 Preprocessing of Data

Data should be refined or transformed before feeding to the algorithm. So It is necessary to convert the raw data into clean and transformed according to the requirement of algorithm.

The data set contain 8 features in column and 768 rows containing values of each female patient. Below steps are used to preprocessing of data.
1. Data is divided in to features and outcomes as x and y respectively
2. Row containing column name dropped.
3. Standard scaler function is used to scale or normalise the data.
4. Data is divided into train, test and validation, 60%, 20% and 20% respectively.
5. Data sets shape changed, so that it can feed to the algorithm.
6. Sigmoid and Costfunction defined to calculate the cost of data.
7. Iteration set to 100000 and learning rate set .0003.

4.1 Implementation of Logistic Regression

Logistic regression uses equation same as linear regression but give output in 0 or 1. Input features x combined using weights to predicts the outcome value y.

It uses sigmoid function to get the output in 0 or 1.

$$sigmoidfunction = 1/(1+e^{-Z}))$$

where is base to natural log and z is linear expression.

$$Z = W^T * Dia_x train + b$$

Each column in input data have coefficient and should be calculated from the training data.

4.1.1 Gradient Descent

Gradient Descent function is used to minimize the error of the function.It is the way to optimize the algorithm so that the output should be more accurate.

$$Cost(w,b) = (\sum_{n=1}^{m}(y_n log(predicted_n) + (1-y_n)log(1-predicted_n))\frac{-1}{m}$$

4.1.2 Estimating the cost

Gradient descent use to estimate the cost of the function. Cost is nothing but the error observed between actual data and predicted data.

There is 2 parameters to get cost.
1. Iteration- Number of times to run training the data.
2. Learning rate- to limit the coefficient of function.

Loop is required to train the data.

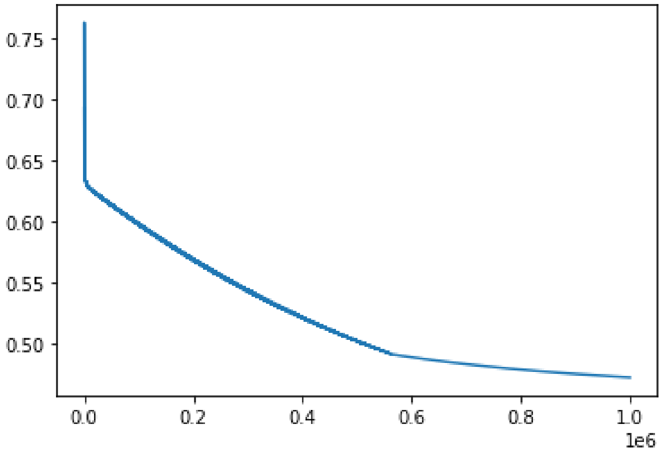
In the project 100000 iteration is used and .0003 learning rate to get cost. Each time coefficients will be updated based on the error of the model.

4.2 Calculating the accuracy of data

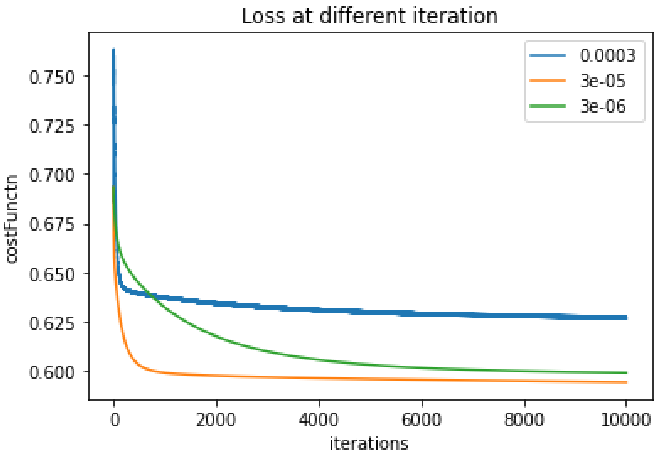
To obtain the accuracy data should be calculated as training data, test data and validation data. Training data is used to train the algorithm, test data is used for testing the algorithm and validation data is used to validate the algorithm.

All training, testing and validation data should be iterated to get the predict the the accuracy of the algorithm.

5 Result



Iteration Graph represents the cost decreasing with increasing iteration.



Loss at different iteration.

5.1 Conclusion:

Accuracy given below for train, test and validation when iteration 100000 and learning rate is .0003.
Accuracy for train data: 77.61 %
Accuracy for test data: 73.38 %
Accuracy for Validation data: 77.92 %

References

1. https://en.wikipedia.org/wiki/Logistic_regression
2. https://www.coursera.org/learn/neural-networks-deep-learning/home/welcome