Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

Using Multivariate Linear Regression to Estimate the Probability of Having a Heart Attack

Parameters used: Age, Cholesterol Levels

Neel Adwani
First Year,
BTech Computer Science with Specialization in Artificial Intelligence and Machine Learning
University of Petroleum and Energy Studies
Dehradun, India

***Abstract*—Heart attack due to high cholesterol level is a new growing problem in the Health industry. For problems like this, Machine Learning can be of great use, when it is put into action. To estimate the probability of having a heart attack, I have written a multivariate linear regression algorithm, which is a part of Machine Learning.**

Keywords—Cholesterol; Age; Machine Learning; Linear Regression

I. INTRODUCTION

Linear Regression in an approach of plotting data on a graph and drawing a straight line that is the best fit for the data. Using that trend, the next value can be predicted easily with the help of slope (theta). In Univariate Linear Regression, one input (x) is fed into the program and it is trained on the basis of y. Then the value of 'x' can be entered to predict the value of 'y' at that point.

In statistics, linear regression is known to be a linear approach to model the connection between a scalar response (or dependent variable) and one or additional informative variables (or independent variables). The case of 1 informative variable is termed univariate linear regression. For quite one informative variable, the method is termed multiple rectilinear regression. This term is distinct from variable linear regression, wherever multiple related to dependent variables are foreseen, instead of one scalar variable.

Multivariate Linear Regression is a technique in which multiple inputs are given, denoted by X(x1, x2, x3,…, xn) and the value of 'y' is fed to train the model. Using the training dataset, a graph is plotted and the value of 'y' can be further predicted by multiplying X with theta.

II. SOFTWARES USED

A. *GNU Octave*

It is an open source software that is compatible with MATLAB commands and is open source, featuring a high level open source programming language named Octave. The Octave language is an interpreted programing language. it's a structured programing language (similar to C) and supports several common C commonplace library functions, and additionally bound UNIX system calls and functions. However, it doesn't support passing arguments by reference. Octave programs accommodates a listing of perform calls or a script. The syntax is matrix-based and provides numerous functions for matrix operations. It supports numerous information structures and permits object-oriented programming. Its syntax is extremely almost like MATLAB, and careful programming of a script can permit it to run on each Octave and MATLAB. As a result of Octave is formed out there beneath the wildebeest General Public License, it can be freely changed or modified. The program runs on Microsoft Windows and in most operating systems and Unix-like operating systems, together with macOS.

B. *KAGGLE*

KAGGLE is a website that is a home to a numerous amount of datasets freely available for research purposes. It is an internet community of information scientists and machine learners, closely-held by Google. Kaggle permits users to seek out and publish knowledge sets, explore and build models in an exceedingly web-based data-science setting, work with different knowledge scientists and machine learning engineers, and enter competitions to unravel knowledge science challenges. Kaggle got its begin by providing machine learning competitions and currently additionally offers a public knowledge platform, a cloud-based work table for knowledge science, and short kind AI education.

III. ALGORITHM

1. Load the whole dataset, containing 3 columns namely, Age, Cholesterol Level and if the heart attack has happened or not in the form of 0 or 1. Store that dataset in a variable.
2. Store the data from first and second column inside the variable 'X'.
3. Store the value of third column inside the variable 'y'.
4. Plot the first column of the dataset on the X-axis and 'y' on the Y-axis of the first figure.
5. Store the data from first and second column inside the variable 'X'.
6. Store the value of third column inside the variable 'y'.
7. Plot the first column of the dataset on the X-axis and 'y' on the Y-axis of the first figure.Plot the second column of the dataset on the X-axis and 'y' on the Y-axis of the second figure.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

```
data =

    28   132     0
    29   243     1
    29   220     1
    30   237     1
    31   219     1
    32   198     0
    32   225     1
    32   254     1
    33   298     1
    34   161     0
    34   214     1
    34   220     1
    35   160     0
    35   167     0
    35   308     1
    35   264     1
    36   166     0
    36   340     1
    36   209     1
    36   160     0
    37   260     1
    37   211     1
    37   173     0
    37   283     1
    37   194     0
    37   223     1
    37   315     1
    38   275     1
    38   297     1
```

Figure 1: Dataset

7.

8. Plot the first column of the dataset on the X-axis and 'y' on the Y-axis of the first figure.Plot the second column of the dataset on the X-axis and 'y' on the Y-axis of the second figure.

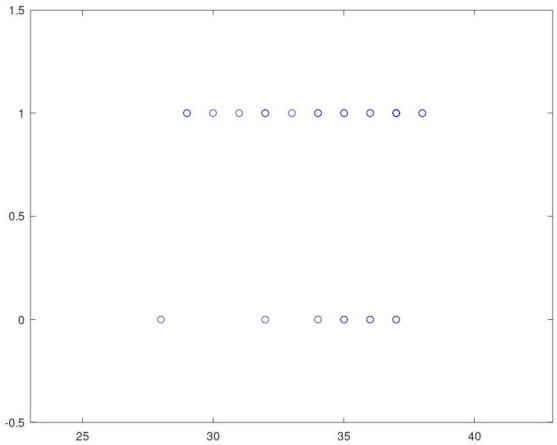

Figure 2: Graph 1

9.

10. Assign the value of 'm' equal to the length of 'y'.Normalize the value of all the elements of the matrix.
11. Update the matrix 'X' and add another column containing all the ones.

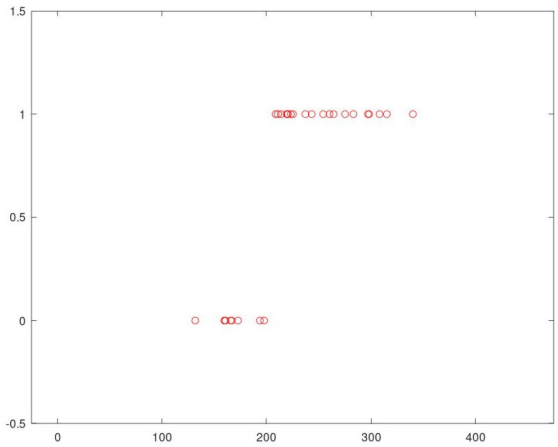

Figure 3: Graph 2

12. Set a learning rate 'alpha'.Set the number of iterations 'num_iters'.
13. Initialize the slope 'theta' to a column zero matrix.
14. Keep updating the value of 'theta' until the value of cost function becomes minimum, depending upon the number of iterations.
15. Plot a convergence graph.

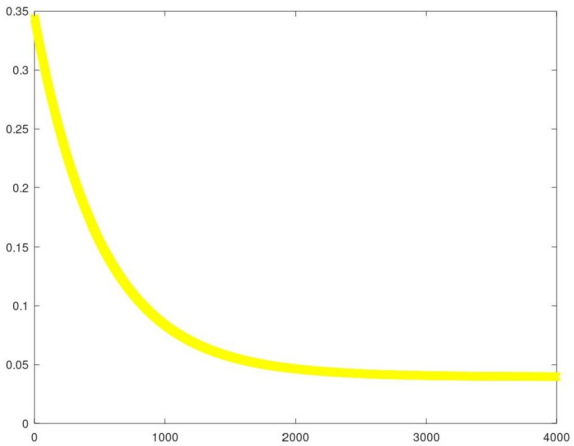

Figure 4: Convergence Plot

16. Input the age and the Cholesterol level.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

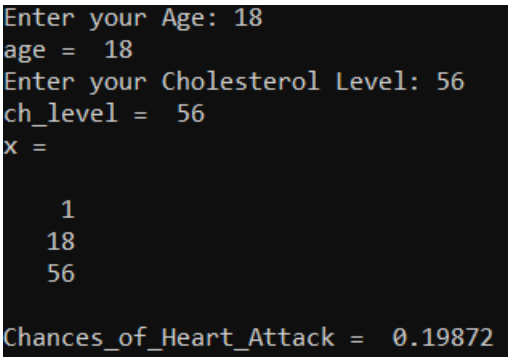

Figure 5: Probability of having a heart attack at age 18 with cholesterol level 56

17. Predict the Probability by multiplying the transpose of theta with the transpose of 'x'.

IV. CODE

```
data = load('heartdata.txt')
X = data(:,1:2);
y = data(:, 3);
a = data(:, 1);
b = data(:, 2);
m = length(y);
figure(1);
plot(a, y, 'bo');
figure(2);
plot(b, y, 'ro');
[X mu sigma] = featureNormalize(X);
X = [ones(m,1) X];
alpha = 0.001;
num_iters = 4000;
theta = zeros(3, 1);
[theta, J_history] = gradientDescentMulti(X, y, theta, alpha, num_iters);
figure(3);
plot(1:numel(J_history), J_history, 'xy', 'LineWidth', 2);
age = input("Enter your Age: ")
ch_level = input("Enter your Cholesterol Level: ")
x = [1 age ch_level]'
Chances_of_Heart_Attack = (theta' * x) / 100
```

FUTURE SCOPE

This algorithm can be used for various purposes in the future, after a lot of improvement. This model is a bit inaccurate because of the lack of data, but once the correct data set is fed into it, it'll be able to find the probability more accurately. Also, more parameters like heart rate need to be considered to increase the accuracy of this model.

REFERENCES

[1] Asmaa Shaker Ashoor , Ali Abdul Karim Kadim Naji, 2019, Statistical Analysis of the Fish Death in Babylon Province by using an Interactive Network of Simple and Multiple Linear Regression/Iraq, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 05 (May 2019).

[2] Girraj Singh, D. S. Chauhan, Aseem Chandel, Deepak Parashar, Girijapati Sharma, 2014, Factor Affecting Elements and Short term Load forecasting Based on Multiple Linear Regression Method, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 03, Issue 12 (December 2014),

[3] Dr. Jihad Alfarajat, Dr. Mohammad Alalaya, 2017, Factors Affecting Heart Diseases through Logistic Linear and Nonlinear Regression, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 06, Issue 07 (July 2017),