**LINEAR REGRESSION AND LOGISTIC REGRESSION**

**Linear Regression**

**INTRODUCTION**

Linear Regression is an Ml algorithm which is based on supervised learning. It is a regression model where it predicts the future outcome based on the current trend. It establishes a relationship between two variables. One of the variables is an independent variable while the other is a dependent variable. The simplest form for an example is the establishment of the height of the individual to their respective weight. Based on these relationships a model is fitted. Care must be taken to test the existence of the relationship between the two parameters.

Linear Regression constructs a relationship between two variables. Let us take the following Equation as Example,

Y = M(X) + C

Here x is an independent variable, it can take any value without any external interference, while y is a dependent variable.

Regression is a method of modelling the output values based on input values. It gives the value based on the value of x. Here, x is the input variable and y is the output variable. The regression models might differ for every model based on the relationship between the variables and the number of independent variables being considered. Here, the model gives an output, Y based on the relationship with the input variable, X. It finds a linear relationship between the two variables, hence Linear Regression.

Fig 1 Fig 2

Fig 1 represents a positive linear relationship, while fig 2 represents a negative linear relationship.

**USES OF LINEAR REGRESSION IN REAL WORLD**

* Applications in trend lines : It can be used to represent the variation in quantitative data with respect to time can be detected using a trend line. A linear relationship is established between them. Therefore, linear regression can be applied to analyse the trend and predict future values (like GDP, oil prices etc.)
* Applications in Economics :  It is used to predict consumer spending, fixed investment spending, expenditures in various fields, inventory investment, purchases of a country’s exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply. It can help us predict the future values based on the current trend line.
* Applications in Finances : Linear regression is applied in capital price asset model in detecting and quantifying the risks of investment etc.

**Types of Linear Regression**

Linear Regression is generally classified into two types which are as follows:

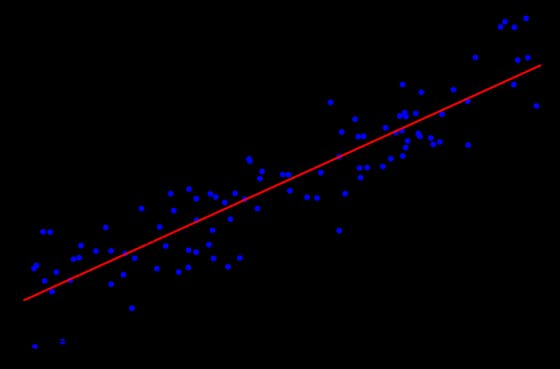
1. Simple Linear Regression
2. Multiple Linear Regression

**Simple Linear Regression**

Here the output is between a single output dependent variable, y and an independent input variable, x. Simple Linear regression in comparison is often less used because usually in any model the number of input variables is generally more than one. Consider,

Y= M(X)+C

Is a simple Linear Regression where the input and output values are single.



Above is an example of Simple Linear Regression.

Applications of Simple Linear Regression:

1. Determining the amount of crop yield based on the total rainfall. Here rainfall is the independent variable and crop yield is the dependent variable.
2. Marks scored by a student based on the number of hours spent on studying. The number of hours spent on studying is the independent variable and marks scored being the dependent.
3. Predicting the salary of the person based on years of experience. The years of experience is the independent variable and the salary is the dependent variable.

**Multiple Linear Regression**

Unlike the Simple Linear Regression, Multiple Linear Regression marks a relationship between two or more independent or input variables and an output or dependent variables. The input variables can be continuous or categorical data.

The equation for Multiple Linear Regression is given as follows:

Y = α0 + α1(X­­1) + α2(X2) + α3(X3)+…..+ αn(Xn) + µ

Here:

* Y -> Output variable/Dependent Variable
* X1, X­2.. -> Input variables/Independent variables
* α -> Coefficients
* µ -> Error

Above is an example for Multiple Linear Regression.

Applications of Multiple Linear Regression are as follows:

1. It will help us determine future values and also to predict trends. It can be used to find point estimators.
2. It will help us understand the effect or intensity of changes. We can visualize how much the output value will change when the input value is changed.
3. It helps us analyse the strength of the effect of independent values have on dependent values.

**Implementation**

Let us start with importing all the necessary libraries.

****

Next step is to read the data set and store the relevant values of the columns in the variables.

****

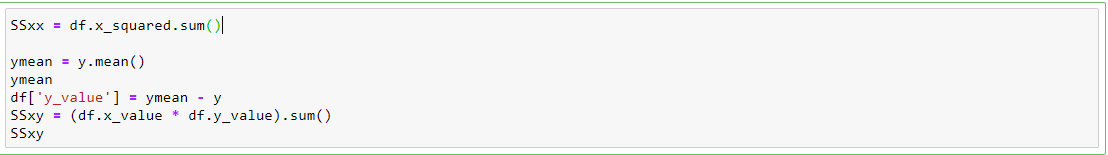
We have to then find the values of x and y in the equation,

Y = M(X) + C

This can be done by finding the mean and subtracting it with the respective x or y values.

****

In order to find the slope, we have to divide the squared sum of x by the product of ‘x\_value’ and ‘y\_value’ and find the sum of it.

****

The slope and the intercept is found as shown below.

****

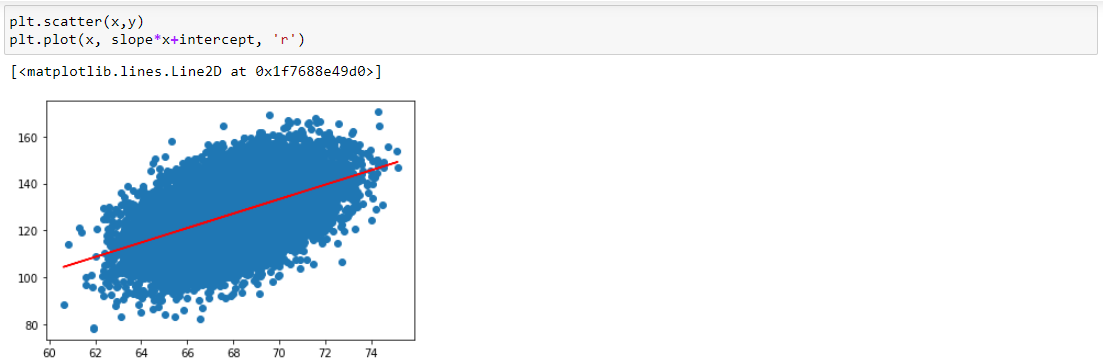
A function to predict the values is implemented where the values are substituted in the above formula and a prediction is made.

****

The value of the slope and the intercept for the input variable is as shown below.

****

The data and Linear regression can be visualized as shown below.

****

**Logistic Regression**

**INTRODUCTION**

Unlike Linear regression, logistic regression is applied for categorical target variable. Linear regression is applicable when target values are continuous, i.e. the output can take any value and is not restricted. But, for a logistic regression model, the output belongs to a specific category. It cannot take any random value.

To do so we have to implement sigmoid function of the hypothetical linear regression.

This regression technique uses logistic function to compress the output values to between 0 and 1. The equation of logistic function is defined as:

**Logistic(n) = 1**

**1 – e-1**

The graph can be given as:

It can also be given as, h( x ) = sigmoid( wx + b ), where sigmoid function is the above equation of Logistic(n).

Input values (x) are combined linearly using weights or coefficient to predict an output value (y). A key difference from linear regression is that the output value generated is a binary values (0 or 1) rather than a numeric value.

Below is a representation of logisctic function,

y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))

The cost function of a linear regression is a non-convex functions of weights, which cannot be used for logistic regression. Hence gradient descent, an optimizing algorithm, gradient descent, can converge the convex function into a global minimum. This can be done by using the following formula,

J = - ylog( h(x) ) - ( 1 - y )log( 1 - h(x) )

Where y -> target value and

h( x ) = sigmoid( wx + b )

**APPLICATIONS**

The applications of logistic regression are as listed below.

1. To detect if a message is spam or not. The output will be either a yes or a no.
2. In medicine, this analytics approach can be used to predict likelihood of disease or illness for a given population, which means that preventative care can be put in place.
3. A manufacturer’s analytics team can use logistic regression analysis as part of a statistics software package to discover a probability between part failures in machines and the length of time those parts are held in inventory.

**TYPES OF LOGISTIC REGRESSION**

1. Binary Logistic Regression: The response variable can belong to either one or two categories.

Example: Spam detection (Spam or not spam)

NBA draft (drafted or not drafted)

1. Multinomial Logistic Regression: The response variable can belong to one of three or more categories and there is no natural ordering among the categories.

Example: Political Preference (different parties)

**Sports Preferences (different types of sports)**

1. Ordinal logistic regression: The response variable can belong to one of three or more categories and there *is* a natural ordering among the categories.

Examples: School Rating (1 -5)

Movie Rating (1-10)

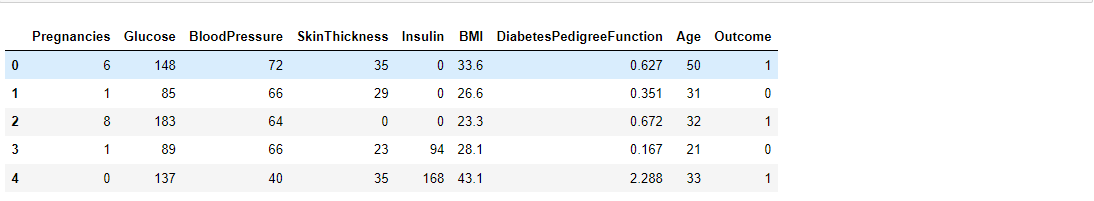
**IMPLEMENTATION**

We start by importing all the necessary libraries as illustrated below.

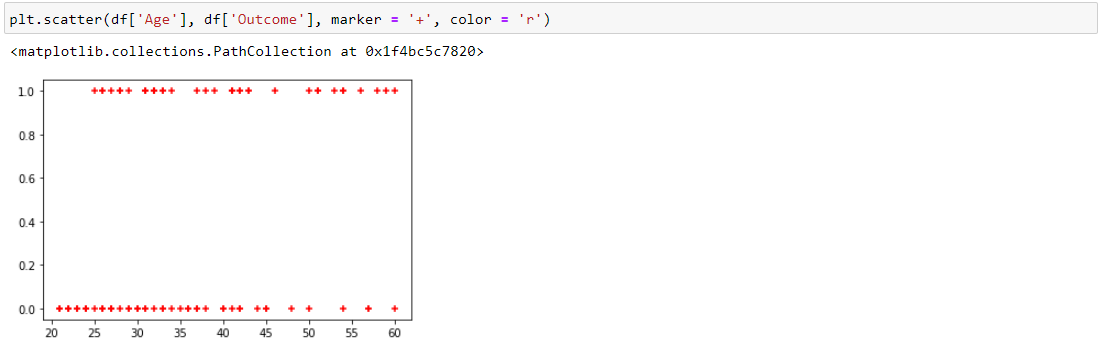
From sklearn, train\_test\_split is imported because we need to split the data for training and testing.

The csv file is read using the pandas library and the first five rows of the dataset is as shown.





We can visualize the data with the help of a scatter plot as shown below.



Under class LogisticRegression(), the code to implement the Logistic Regression is shown.

We define four functions:

The first function is initialize the variables for LogisticRegression(). Here we store the values of the variables of learning rate and iterations. Learning rate is the value of how much the slope and the intercept is being changed. Iterations is the number of times the entire process has to be done.

Fit() is a function where the actual process of fitting the model takes place. The values are initialized and update\_weights function is called.

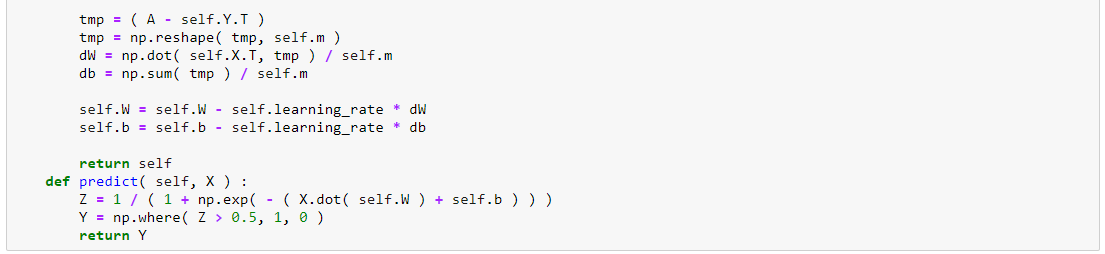
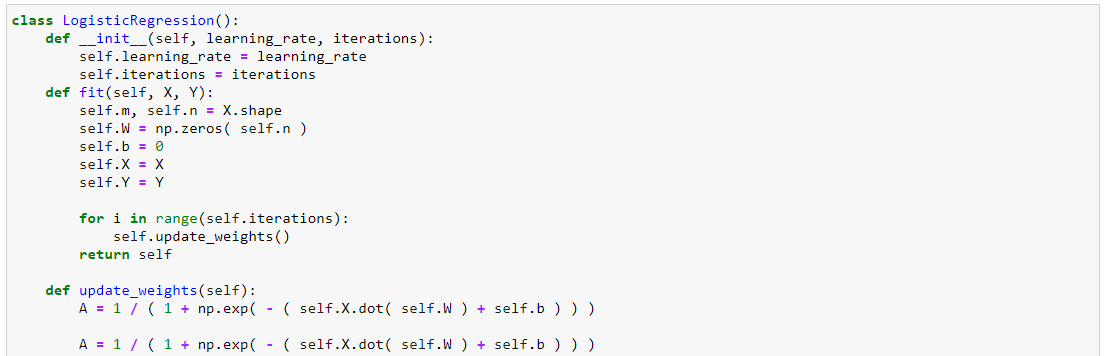
Update\_weights function does the process of changing the bias and the weights. Gradient descent an optimizing algorithm as discussed above, is used to to do this.

https://media.geeksforgeeks.org/wp-content/uploads/20200915115103/chainrule.png

The derivation above is the chain rule of derivation which is used in the calculation of the

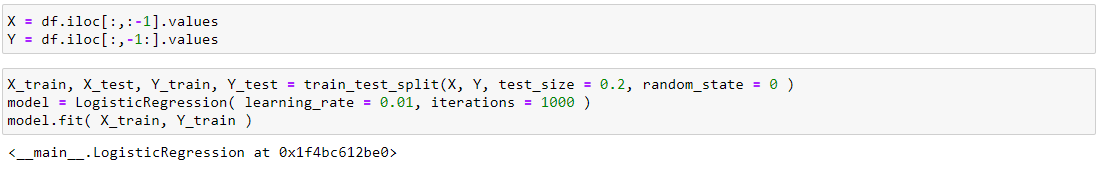
Gradient descent value. Loss function and gradient descent is required because when we train the model, we have lower the cost function value to increase the probability of predicting the correct values.

When the predict function is called, the entire process of execution of the model is computed.

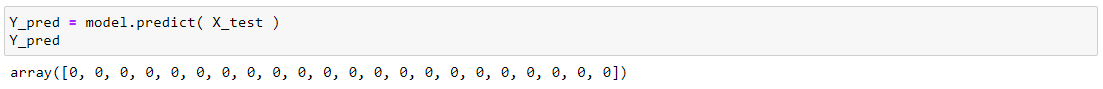


The memory locations of the data are stored in the X and Y variables.

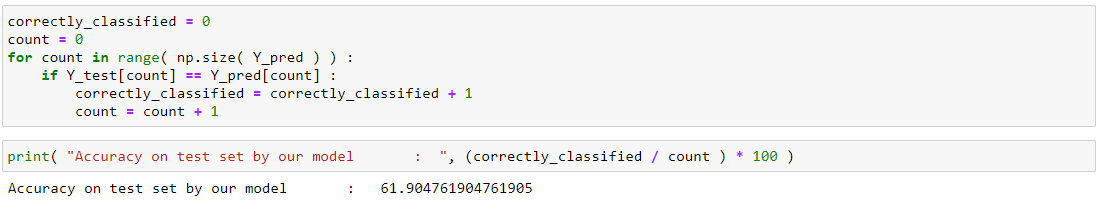
The dataset is split such that 20 per cent of it is used to test our model. We then call the LogisticRegression() and pass in the input parameters. Later, we finally fit the model.



Code snippet to predict test the values in test dataset.



We can check the accuracy score of our model as shown below.



**TERMS AND TERMINOLOGIES**

In order to begin with the algorithms, we need to have a basic understanding of few terminologies.

**COST FUNCTION**

When we have to fit a line, there are many possibilities. Cost function basically means how far is our predicted value is from the actual points that are plotted and given. We will have to fit a line that has the least cost function value. For example, when we find the values of m and y in the linear regression formula, Y = M(X) + C, a line is plotted for these values of m and c and is verified if the line is the best fit for the data. If not, the values of m and y are changed such that when plotted again, there is a better fit of the line.

The cost function can be found out with formula below,

**Cost function = 1 ∑1m­­ (h(x)I - yi)2**

**2m**

Where,

1/2m -> Constant,

m-> number of data points

h(xi) -> the output of our hypothesis for particular value of I using y = m\*x +c

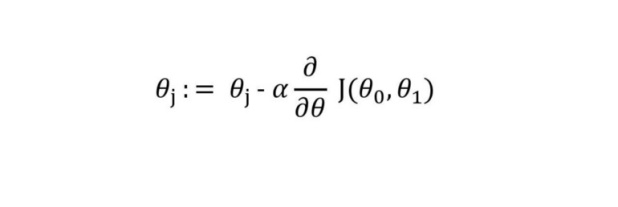
yi -> the value of the dataset we have

**GRADIENT DESCENT**

When the line generated is not a best fit, we have to modify the m and c values. This can be done using cost function reduction method called ‘gradient descent’. The cost function value must be reduced.

With the help of gradient descent we can change the values of m and c little by little such that we can get closer to a better fit of the points.

The gradient descent is calculated using,



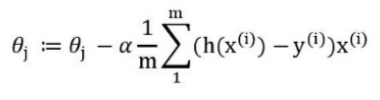
Where,

α —> with how much magnitude you are reducing your value.

Theta -> individual data points

Therefore, every time we fit a model and it is not accurate enough, we have to check the cost function and change the values of m and y until we obtain a desired fit.

A final equation can be derived from these equations,



**CONVEX FUNCTION**



A real-valued function is called convex if the line segment between any two points on the graph of the function does not lie below the graph between the two points. a function is convex if its epigraph (the set of points on or above the graph of the function) is a convex set. A twice-differentiable function of a single variable is convex if and only if its second derivative is nonnegative on its entire domain.

References:

[1] [Linear Regression In Real Life. Real world problems solved with Math | by Carolina Bento | Towards Data Science](https://towardsdatascience.com/linear-regression-in-real-life-4a78d7159f16)

[2] [Linear Regression for Machine Learning (machinelearningmastery.com)](https://machinelearningmastery.com/linear-regression-for-machine-learning/)

[3] [ML | Linear Regression - GeeksforGeeks](https://www.geeksforgeeks.org/ml-linear-regression/)

[4] [Beginner: Cost Function and Gradient Descent | by Kshitiz Sirohi | Towards Data Science](https://towardsdatascience.com/machine-leaning-cost-function-and-gradient-descend-75821535b2ef)

[5] [Logistic Regression — Detailed Overview | by Saishruthi Swaminathan | Towards Data Science](https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc)-

[6] [5.2 Logistic Regression | Interpretable Machine Learning (christophm.github.io)](https://christophm.github.io/interpretable-ml-book/logistic.html)