Linear Regression Report

Donal Loitam(AI21BTECH11009)

July 16, 2022

Contents

1 Introduction

- **Regression** shows a line or curve that passes through all the data points on a target-predictor graph s.t the vertical distance between the data points and the regression line is minimum.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression
- The below graph gives the linear relationship between the dependent and independent variables.

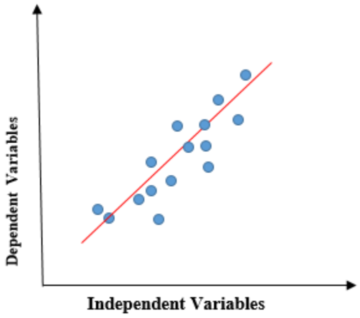

Figure 1: Linear graph $y = mx + c$

- The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.
- To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$h(x) = mx + c = a_0 + a_1x \tag{1}$$

2 Key Points of the Algorithm

Let for our simplicity, we have only 2 parameters a_0 and a_1 i.e we work on 2-Dimensions
The goal of the algorithm is to get the best values of a_0 and a_1, to find the best fit line. The best fit line should have the least error i.e the error between predicted values and actual values be minimized. The **cost function** $(J(a))$ in Linear Regression is defined by **Mean Squared Error(MSE)**

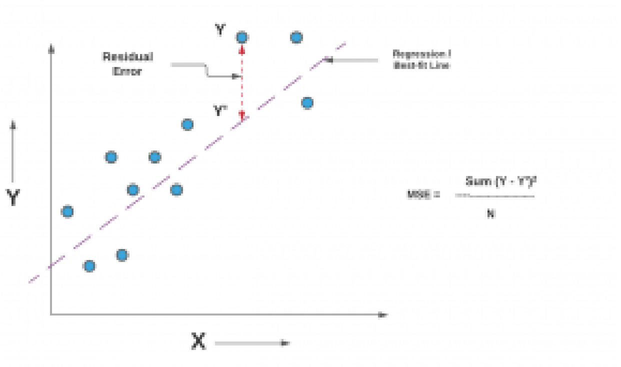

Figure 2: Linear graph $y = mx + c$

$$J(a) = \frac{1}{2}\sum_{i=1}^{m}(h_a(x^{(i)}) - y^{(i)})^2 \tag{2}$$

where our hypothesis is $h_a(x)$, a = weights/parameters, x = inputs be defined as :

$$h_w(x) = a_0 + a_1x_1 + ... \tag{3}$$

and m = no. of training samples, $(x^{(i)}, y^{(i)})$ = i'th training sample
But how do we get the values of a_0, a_1 which minimises the cost function.For this we use a method called **Gradient Descent**
A regression model uses gradient descent to update the coefficients of the line $(a0, a1 \implies xi, b)$ by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function. We repeat the algorithm until it converges to the global

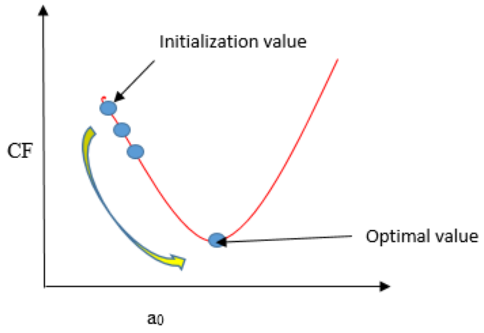

Figure 3: Linear graph $y = mx + c$

minima i.e update the parameter as follows: (every iteration)

$$a^{(i)} = a^{(i)} - \alpha.(\frac{\partial cost}{\partial a^{(i)}}) \tag{4}$$

$$\tag{5}$$

where, α = learning rate
For simplification, assume only one training sample($m = 1$)

$$\begin{aligned}
\frac{\partial cost}{\partial a^{(i)}} &= \frac{\partial}{\partial a^{(i)}}\frac{1}{2}(h_a(x) - y)^2 && (6)\\
&= (h_a(x) - y)\frac{\partial}{\partial a^{(i)}}(h_a(x) - y) && (7)\\
&= (h_a(x) - y)\frac{\partial}{\partial a^{(i)}}(a_0 + a_1x_1 + ...a_ix_i.... - y) && (8)\\
&= (h_a(x) - y).x_i && (9)
\end{aligned}$$

Now since derivaive of the sum = sum of derivative , we can generalise this for m trianing samples
Repeat unitl convergence ,

$$w_j = w_j - \alpha\sum_{i=1}^{m}(h_w(x^{(i)} - y^{(i)}).x_j^{(i)} \tag{10}$$

$$\tag{11}$$

3 Why does it work ?

It turns out that when you plot the cost function $J(a)$ for linear regression, it is a concave up quadratic graph as in 4 and hence the only local minima is a **global minima.**
Imagine a pit in the shape of U. You are standing at the topmost point in the pit, and your objective is to reach the bottom of the pit. In gradient descent algorithm, we are always travelling in the direction of gradient (essentially the direction of steepest descent) and hence at one point of time we will converge to the minima.
The partial derivates are the gradients, and they are used to update the values of a0 and a1. Alpha

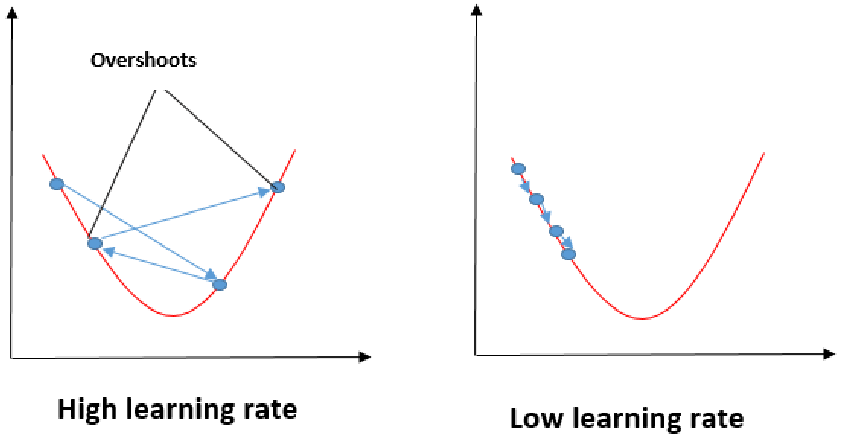

Figure 4: $J(a)$/ cost function

is the learning rate.
Note: As seen inthe fig α should not be too big because we may never converge to the global minima if we take huge steps Similarly α shouldn't be too small as it will increase computatipnal power and time.

4 Some Questions

1.When is Linear regression suitable and how can we know that from a given dataset?
Ans. Generally for 1 independent variable, a Scatter plot is used to see if linear regression is suitable for any given data. And if the plot looks somewhat linear then we can go for linear regression but for more than one independent variable, then two-dimensional pairwise scatter plot can be used .
2. What are the basic assumptions of Linear Regression ?
Ans.

- The relationship between the features(independent variables) and target(dependent variable).
- There is no multicollinearity between the features i.e There does not exist a linear dependency between the independent variables
- The error(residuals) are normally distributed.

3. In linear regression, what is the value of the **sum of the residuals/distance** for a given dataset? Justify !
Ans Since the error/reiduals/distance is assumed to follow normal distribution in case of linear regression, expected value or mean equal to 0.

$$Mean = \frac{1}{m}\sum_{i=1}^{m}(residuals) \tag{12}$$

$$\because Mean = 0 \implies \sum(residuals) = 0 \tag{13}$$

The sum of the residuals in a linear regression model is 0
4. Why do we square the residuals(M.S.E) instead of using absolute residuals?
Ans. In mathematical terms, the squared function is differentiable everywhere, while the absolute error is not differentiable at all the points in its domain(its derivative is undefined at 0). So while optimising squared error, we can differentiate it and equal it to 0, but we will have more computational headache for absolute residuals.
5.) What is Multicollinearity and how do we detect them ?
Ans. Multicollinearity happens when independent variables in the regression model are highly correlated to each other i.e. one variable can be linearly predicted with the help of other variables. To detect it :
Correlation coefficient: The first simple method is to plot the correlation matrix of all the independent variables.
VIF : With the help of Variance inflation factor (VIF) for each independent variable
VIF = 1, no correlation between the independent variable and the other variables.
VIF > 5 or 10 indicates high multicollinearity between this independent variable and the others.