Page No.
Date

Assignment A1

Title :- To find the best fit line for the given data using Linear Regression.

Date of Completion : 27th January 2021

Problem statement :-

The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. Find the equation of the best fit line for this data.

(x) Number of hours spent driving

(y) Risk score on a scale of 1-100

Objectives :-

To understand how linear regression works on the given dataset.

Outcome :-

Students will be able to :

- Find the best scenario for the result to be achieved for a given data set using linear regression.

Software and Hardware requirements

Core 2 DUO / i3/i5/i7 64 bit processor OS - Linux 64 bit OS

Editor - gedit / Eclipse
Software - Jupyter Notebook / Python.

Theory:-

Linear Regression

In statistics linear regression is linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variable is called simple linear regression.
For more than one explanatory variable the process is called multiple linear regression.

Real-time example

a dataset which contains information about relationship between 'number of hours spent driving' and 'risk score'. Many drivers have been observed and their hours of driving are recorded.
If we give number of hours driven by driver as an input our model should predict the risk with minimum error.

$$\hat{y} = b_0 + b_1 x$$

where b_0 is constant
b_1 is regression coefficient
x is value of independent variable
y is dependent variable.

Page No.
Date

The values b_0 and b_1 must be chosen so that they minimise the error. To assess the model usually RSE (Residual Standard Error) and R^2 statistics are used.

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1} (y - \hat{y})^2}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where $TSS = \sum (y_i - \hat{y}_i)^2$

The formula to find b_0 and b_1 are:

$$b_0 = \hat{y} - b_1 \hat{x}$$

$$b_1 = \frac{\sum (x_i - \hat{x}) * (y_i - \hat{y})}{\sum (x_i - \hat{x})}$$

$$r = \frac{\text{covariance}(x, y)}{\text{std.dev}(x) * \text{std.dev}(y)}$$

Algorithm:

(i) Gota bunch of points in R^2, $\{(x^i, y^i)\}$
(ii) Want to fit a line $y = ax + b$ that describes the trend.

(iii) We define a cost function that computes the total squared error of our predictions with respect to observed values.

(iv) See it as a function of a and b

(v) The coefficient you get give you the minimum squared error.

(vi) Can do this for specific points, or in general and find the formulas.

(vii) More general version in R^n

Result:

Regression line's Equation

$$y = 4.59x + 12.58$$

Conclusion:-

Successfully applied linear regression model on the given data set, and calculated a best fit equation.