Assignment No : 1

Title : Linear Regression

Problem statement : The following table shows the result of recently conducted study on the correlation of no. of hours spent driving with the risk of developing acute backache. Find the equation of best fit line for this data.

No. of hours spent driving (X)	Risk score on scale of 0 - 100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

Objective : Students should be able to build a linear model for the given data.

Outcome : After completion of this assignment, students are able to understand how to find correlation between variables and how to calculate accuracy of linear model.

CLASSMATE
Date :
Page :

S/w requirements: Anaconda, python, sklearn, linux os

H/w requirements: i3 processor, 4 gb ram

Concept related theory:

1] Linear regression: Regression analysis is one of the most widely used statistical techniques. It estimates relationship among a dependant variable and independent variables.

What is linear regression -

In cause and effect relationship, the independent variable is the cause and dependent variable is the effect. Least square regression is the method of predicting the value of dependent variable based on value of independent variable. x.

2] The least square regression line:

Given the random sample of observations population regression line is estimated by:

$$\hat{y} = b_0 + b_1 x$$, where b_0 is constant

b_1 is regression coefficient,
x is value of independent variable
y is value of dependent variable.

— To find the relationship between dependant and independant variable we find correlation.

— The formula to find correlation using pearson correlation is -

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

3] How to define regression line?

We know regression model :

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

— The formula for calculating b_1 & b_0.

$$b_1 = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

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

4] Accessing the model

— To access the model, usually RSE (Residual Standard Error) and R^2 statistics are used.

classmate
Date :
Page :

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

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

where $TSS = \Sigma (y_i - \hat{y_i})^2$

i) The first error matrix is simple to understand, the lower the residual error, the better the model fits the data.

ii) As for the R^2 matrix it measures the proportion of variability in the target that can be explained using a future X. So, assuming the linear relationship if feature X can predict the target, then the proportion is high and R^2 value will be close to 1. If the opposite is true, then R^2 value will be closer to 0.

5] What is best line?

i) The best line is a straight line that represents the data on the scatter plot
ii] This line may pass through all the points or none or at some of points.
iii] Best line is determined from the matric such as least squares.

Algorithm — 1. Calculate average of X variable.
2. Calculate difference between each X and average X.
3. Square the differences & add it all up. SSxx
4. Calculate average of Y variable.
5. Multiply difference of X & Y from respective averages and add them. This SSxy
6. Using SSxx and SSxy, calculate intercept by subtracting SSxy / SSxx * Avg (x) from Avg (y).

Linear regression equation: $4.58x + 12.58$

Conclusion: Thus, we learned to find the trend of data using X & Y variable using linear regression.