Data Science Interview Question and Solution

03/01/2024

Question

Problem Statement: Suppose You are given a dataset containing two variables, X and Y. Your task is to find the linear regression equation to model the relationship between X and Y. Explain the mathematical steps you would take to achieve this.

Solution

Step 1: Formulate the Hypothesis

In linear regression, we assume that the relationship between the independent variable X and the dependent variable Y is linear and can be represented as:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

where:

- Y is the dependent variable (response),
- X is the independent variable (predictor),
- β_0 is the y-intercept,
- β_1 is the slope,
- ϵ is the error term.

Step 2: Define the Objective Function

The objective is to minimize the sum of squared differences between the predicted values and the actual values (Ordinary Least Squares - OLS):

$$\text{Minimize} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 \cdot X_i))^2$$

where n is the number of data points.

Step 3: Partial Derivatives and Gradients

Compute the partial derivatives of the objective function with respect to β_0 and β_1 and set them equal to zero to find the critical points.

$$\frac{\partial}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 \cdot X_i)) = 0$$

$$\frac{\partial}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - (\beta_0 + \beta_1 \cdot X_i)) = 0$$

Step 4: Solve for Coefficients

Solve the system of equations to find the values of β_0 and β_1:

$$\beta_1 = \frac{n(\sum_{i=1}^{n} X_i Y_i) - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n(\sum_{i=1}^{n} X_i^2) - (\sum_{i=1}^{n} X_i)^2}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} Y_i - \beta_1(\sum_{i=1}^{n} X_i)}{n}$$

Step 5: Interpretation

Interpret the values of β_0 and β_1 in the context of the problem. β_0 represents the y-intercept, and β_1 represents the slope of the regression line.

This is a simplified explanation of the mathematical steps involved in simple linear regression. The derivation becomes more involved in multiple linear regression with multiple predictors.

Linear Regression: Mathematical Explanation with Sample Values

In the Problem statement there are given a dataset with two variables, X and Y, where:

$$X = [1, 2, 3, 4, 5]$$
$$Y = [2, 3, 4, 5, 6]$$

The goal is to find the linear regression equation to model the relationship between X and Y.

0.1 Hypothesis

In linear regression, we assume the relationship is given by:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

0.2 Objective Function

Minimize the sum of squared differences (OLS):

$$\text{Minimize } \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 \cdot X_i))^2$$

0.3 Partial Derivatives and Gradients

$$\frac{\partial}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 \cdot X_i)) = 0$$

$$\frac{\partial}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - (\beta_0 + \beta_1 \cdot X_i)) = 0$$

0.4 Solve for Coefficients

$$\beta_1 = \frac{n(\sum_{i=1}^{n} X_i Y_i) - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n(\sum_{i=1}^{n} X_i^2) - (\sum_{i=1}^{n} X_i)^2}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} Y_i - \beta_1(\sum_{i=1}^{n} X_i)}{n}$$

0.5 Substitute Sample Values

$$n = 5$$
$$\sum_{i=1}^{n} X_i = 15$$
$$\sum_{i=1}^{n} Y_i = 20$$
$$\sum_{i=1}^{n} X_i^2 = 55$$
$$\sum_{i=1}^{n} X_i Y_i = 70$$

0.6 Solving for Coefficients

$$\beta_1 = \frac{5(70) - (15)(20)}{5(55) - (15)^2}$$
$$\beta_0 = \frac{20 - \beta_1(15)}{5}$$

Now, substitute these values into the equations to find β_0 and β_1:

$$\begin{aligned} \beta_1 &= \frac{5(70) - (15)(20)}{5(55) - (15)^2} \\ &= \frac{350 - 300}{275 - 225} \\ &= \frac{50}{50} \\ &= 1 \end{aligned}$$

Now that we have $\beta_1 = 1$, substitute it into the equation for β_0:

$$\begin{aligned} \beta_0 &= \frac{20 - \beta_1(15)}{5} \\ &= \frac{20 - 15}{5} \\ &= 1 \end{aligned}$$

So, the values of β_0 and β_1 for the linear regression equation are $\beta_0 = 1$ and $\beta_1 = 1$.