Sushant University

upGrad

Linear Regression
Subjective Assignment

Submitted By: Abhijit Mandal

Submitted on: 28/09/2023

Enrollment ID: 210BTCSEAM024

B.TECH CSE AI/ML

Semester: 5

Q1. Explain the linear regression algorithm in detail.

A1. Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors or features). The goal is to find a linear equation that best describes how changes in the predictor variables affect the target variable. This equation is typically represented as:

$$Y = \beta0 + \beta1 * X1 + \beta2 * X2 + \ldots + \beta n * Xn + \varepsilon$$

Where

- Y represents the target variable.
- X1, X2, ..., Xn are the independent variables.
- β0 is the intercept, and β1, β2, ..., βn are the coefficients of the independent variables.
- ε represents the error term, which accounts for the variability not explained by the model.

The linear regression algorithm aims to estimate the coefficients (β0, β1, β2, ...) that minimize the sum of squared differences between the predicted values and the actual values in the training data.

Q2. What are the assumptions of linear regression regarding residuals?
A2. Linear regression makes several assumptions regarding the residuals (the differences between observed and predicted values):

- **Linearity**: The relationship between the predictors and the target variable is assumed to be linear.
- **Independence**: Residuals should be independent of each other, meaning that the error for one data point does not depend on the errors of other data points.
- **Homoscedasticity**: The variance of residuals should be constant across all levels of the predictors (i.e., the spread of residuals should be roughly the same).
- **Normality**: Residuals are assumed to be normally distributed.

Q3. What is the coefficient of correlation and the coefficient of determination?

A2. **Coefficient of correlation**

- The coefficient of correlation (Pearson's correlation coefficient, denoted as "r") measures the strength and direction of a linear relationship between two variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.

Coefficient of determination

- The coefficient of determination (R-squared, denoted as "R^2") represents the proportion of the variance in the dependent variable (Y) that can be explained by the independent variables (X). It ranges from 0 to 1, with higher values indicating a better fit of the model to the data.

Q4. Explain the Anscombe's quartet in detail.

A4. Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but differ significantly when graphically plotted. It was created by statistician Francis Anscombe to emphasize the importance of data visualization in understanding relationships. The quartet highlights that even when summary statistics look similar, the underlying data patterns can be quite different.

Q5. What is Pearson's R?

A5. Pearson's correlation coefficient (often denoted as "r") is a measure of the linear correlation between two continuous variables. It quantifies the strength and direction of the linear relationship between the variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.

Q6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A6. Scaling :

- Scaling is the process of transforming the values of variables to a standard range or distribution. It's done to ensure that all variables have similar scales, which can be important for certain machine learning algorithms.
- Normalized scaling typically refers to scaling the values of a variable to the range [0, 1]. It's done by subtracting the minimum value and dividing by the range (max - min). This is useful when the variable doesn't follow a normal distribution.
- Standardized scaling (standardization) transforms variable values to have a mean of 0 and a standard deviation of 1. It's done by subtracting the mean and dividing by the standard deviation. This is useful when variables are normally distributed and have different units.

Q7. Why does the VIF sometimes have an infinite value?

A7. The Variance Inflation Factor (VIF) measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity (high correlation) between independent variables. VIF can become infinite when perfect multicollinearity exists, meaning that one predictor variable can be exactly predicted from a linear combination of others. This makes it impossible to estimate the coefficient for the perfectly correlated variable, resulting in an infinite VIF.

Q8. What is the Gauss-Markov theorem?

A8. The Gauss-Markov theorem is a fundamental result in statistics related to the properties of linear regression estimators. It states that in a linear regression model where the errors (residuals) have a mean of zero, are uncorrelated, have constant variance (homoscedasticity), and are normally distributed, the ordinary least squares (OLS) estimator of the regression coefficients is the Best Linear Unbiased Estimator (BLUE). In simpler terms, OLS provides the most efficient and unbiased estimates of the coefficients under these assumptions.

Q9. Explain the gradient descent algorithm in detail.

A9. Gradient descent is an optimization algorithm used to minimize the cost or loss function in machine learning and linear regression. It iteratively updates the model's parameters (coefficients) to find the values that minimize the cost. Here are the main steps:

- Initialize the coefficients randomly or with some initial guess.
- Compute the gradient (derivative) of the cost function with respect to each coefficient.
- Update the coefficients by moving in the opposite direction of the gradient.
- Repeat the above steps until convergence (the cost stops decreasing or changes very slowly).
- The learning rate determines the step size in each iteration.

Q10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

A10. A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, usually the normal distribution. It compares the quantiles (ordered values) of the dataset to the quantiles expected from the theoretical distribution.

The Q-Q plot is important in linear regression and statistics for several reasons:

- It helps check the assumption of normality in residuals. If the points on the Q-Q plot approximately follow a straight line, it suggests that the residuals are normally distributed.
- Deviations from a straight line can indicate departures from normality, which may affect the validity of statistical tests and confidence intervals.
- It allows you to identify potential outliers or data points that do not conform to the assumed distribution.

In linear regression, validating the assumption of normality in residuals is crucial for making valid statistical inferences and ensuring the reliability of the model's predictions.