Linear Regression Assignment

Pratik Hatwalne

General Knowledge Subjective Questions

Q1. Explain the Linear Regression Algorithm in detail.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent variables. It is used to predict the value of the dependent variable based on the combination of several independent variables. The algorithm is capable of finding the best possible linear equation that can provide a relationship between the dependent and the independent variables and also can predict the value of the dependent variables given the values of the independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable. It can be used in many different domains like finance, economics and even psychology.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of four datasets that were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. The datasets have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset contains 11 (X, y) points. The quartet is used to illustrate the importance of visualizing data and to show that the summary statistics alone can be misleading. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics such as the mean and variance, x and y correlation coefficient and the linear regression line.

Q3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that is used to determine the strength and direction of the linear relationship between two continuous variables. It is a number between -1 and 1, where a value of 1 indicates a perfect positive correlation and the value of -1 indicates a perfect negative correlation and the value of 0 indicates no correlation between the variables. The Pearson correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is commonly used in fields such as psychology, economics, and social sciences to measure the relationship between two variables.

Q4. What is scaling? Why is scaling performed? What is the difference between standardized scaling and normalized scaling?

Scaling is a method used to normalize the range of independent variables or features of data. It is performed during the data preprocessing step and is used to bring all features in the same standing. The main purpose of scaling is to improve the performance of the machine learning algorithms by

ensuring that all variables have comparable ranges and magnitudes. Scaling is important because real world datasets often contain features that are varying in degrees of magnitude, range and units. If scaling is not performed, Machine learning algorithms tend to weight greater values high and consider smaller values as lower values, regardless of the unit of those values.

Standardized scaling involves making the values of each feature in the data to have zero mean and unit variance, whereas in normalized scaling, we scale the values of a feature to range between 0 and 1.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The value of the variance inflation factor (VIF) is sometimes shown as infinity when there is perfect correlation between the independent variables. This means that one of the independent variables can be expressed as a linear combination of the other independent variables and in such cases the value of the VIF tends to infinity. In other words, VIF becomes infinite when one of the independent variables can be perfectly predicted from the other independent variables. This indicates a strong case of multicollinearity, which is a condition where two or more independent variables in a linear regression model are highly correlated. In such cases, the VIF becomes infinite because the variance of the regression coefficient cannot be estimated. When VIF is infinite, it indicates that the independent variables are not providing unique information to the model and that the model itself may not be reliable.

Q6. What is a Q-Q plot? Explain the use and the importance of a Q-Q plot in linear regression.

A Q-Q plot or a quantile-quantile plot is a graphical tool used to assess if a set of data plausibly came from some theoretical distribution, such as a normal one. It is a scatter plot created by plotting two different quantiles against each other. The first quantile is that of the variable being tested and the second one is the actual distribution that is being tested against. The Q-Q plot is used to visually compare the distribution of a variable to a theoretical distribution.

In linear regression, the Q-Q plot is sued to check the normality assumption of the residuals. The residuals are the differences between the predicted and the actual values of the dependent variable. If the residuals are normally distributed the Q-Q plot will show a straight line, if the Q-Q plot deviates from a straight line, it indicates that the residuals are not normally distributed and the normality assumptions of the linear regression model may be violated. It is an important tool in linear regression because it helps to identify any issues with the normality assumption of the residuals, which can affect the validity of the regression analysis.

Assignment based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- Fall and Summer come with the highest demand for motorcycles, Spring might be termed as 'Off-Season'

- There seems to be a significant growth in demand for 2018 and 2019

- June, July and August can be termed as the peak season since the median is also high and the spread is low. September October also see a high demand on some days, but could not be considered as the peak seasons due to the median being lower and spread being higher

Q2. Why is it important to use drop_first=True during dummy variable creation?

While it is always recommended to use the drop first = True option, it is not mandatory. It is done only to reduce the number of dummy variables in the data, thus making data processing faster. Suppose we have people from 3 categories, and we have to create dummy variables for them, we can create 0 1 and 2 levels, but creating two levels can help process the same amount of information but using less variables, ie if a person is not from category A as well as not from category B, he is bound to be from category C. We do not need to have a new level or variable that tells us that the person is from the third category. The same thing is obtained using the drop first = True option.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The ‘temp’ variable has the highest correlation with the target variable ‘cnt’

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The following checks were made:

1. Normality of error terms
2. Multicollinearity Check
3. Linear Relation Check
4. Homoskedasticity
5. Independence of Residuals

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top 3 features that contribute significantly towards the demand of the shared bikes are:

1. Bad Weather Situation
2. Temperature
3. Year

4. Spring Season