Linear Regression Interview Questions

We believe that you have learned both theoritical and practical knowledge on Naive Bayes classification algorithm through your assignment.

So let's test your knowledge here. This will help you to be prepared for interviews too!

Best with Quest

1. What are the important assumptions of Linear regression?

```
1. A linear relationship
2. Restricted Multi-collinearity value
3. Homoscedasticity

Firstly, there has to be a linear relationship between the dependent and the independent
variables. To check this relationship, a scatter plot proves to be useful.

Secondly, there must no or very little multi-collinearity between the independent
variables in the dataset. The value needs to be restricted, which depends on the domain
requirement.

The third is the homoscedasticity. It is one of the most important assumptions which
states that the errors are equally distributed.
```

2. What is heteroscedasticity?

Heteroscedasticity is exactly the opposite of homoscedasticity, which means that the error terms are not equally distributed. To correct this phenomenon, usually, a log function is used.

3. What is linear regression?

```
n simple terms, linear regression is a method of finding the best straight line fitting to
the given data, i.e. finding the best linear relationship between the independent and
dependent variables.
In technical terms, linear regression is a machine learning algorithm that finds the best
linear-fit relationship on any given data, between independent and dependent variables. It
is mostly done by the Sum of Squared Residuals Method.
```

4. Can we use linear regression for time series analysis?

One can use linear regression for time series analysis, but the results are not promising. So, it is generally not advisable to do so. The reasons behind this are : —

Time series data is mostly used for the prediction of the future, but linear regression seldom gives good results for future prediction as it is not meant for extrapolation. Mostly, time series data have a pattern, such as during peak hours, festive seasons, etc., which would most likely be treated as outliers in the linear regression

analysis.

5. What value is the sum of the residuals of a linear regression close to? Justify.

The sum of the residuals of a linear regression is 0. Linear regression works on the assumption that the errors (residuals) are normally distributed with a mean of 0, i.e.

$Y = \beta^T X + \varepsilon$

Here, Y is the target or dependent variable, β is the vector of the regression coefficient, X is the feature matrix containing all the features as the columns, ε is the residual term such that $\varepsilon \sim N(0,\sigma 2)$. So, the sum of all the residuals is the expected value of the residuals times the total number of data points. Since the expectation of residuals is 0, the sum of all the residual terms is zero. Note: $N(\mu,\sigma 2)$ is the standard notation for a normal distribution having mean μ and standard deviation σ2.

6. How does multicollinearity affect the linear regression?

Multicollinearity occurs when some of the independent variables are highly correlated (positively or negatively) with each other. This multicollinearity causes a problem as it is against the basic assumption of linear regression. The presence of multicollinearity does not affect the predictive capability of the model. So, if you just want predictions, the presence of multicollinearity does not affect your output. However, if you want to draw some insights from the model and apply them in, let's say, some business model, it may cause problems. One of the major problems caused by multicollinearity is that it leads to incorrect interpretations and provides wrong insights. The coefficients of linear regression suggest the mean change in the target value if a feature is changed by one unit. So, if multicollinearity exists, this does not hold true as changing one feature will lead to changes in the correlated variable and consequent changes in the target variable. This leads to wrong insights and can produce hazardous results for a business. A highly effective way of dealing with multicollinearity is the use of VIF (Variance Inflation Factor). Higher the value of VIF for a feature, more linearly correlated is that feature. Simply remove the feature with very high VIF value and re-train the model on the remaining dataset.

7. What is the normal form (equation) of linear regression?

The normal equation for linear regression is —

$\beta=(XTX)-1.XTY$

Here, $Y=\beta TX$ is the model for the linear regression, Y is the target or dependent variable, β is the vector of the regression coefficient, which is arrived at using the normal equation, X is the feature matrix containing all the features as the columns. Note here that the first column in the X matrix consists of all 1s. This is to incorporate the offset value for the regression line.

8. You run your regression on different subsets of your data, and in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

This case implies that the dataset is heterogeneous. So, to overcome this problem, the dataset should be clustered into different subsets, and then separate models should be built for each cluster. Another way to deal with this problem is to use non-parametric models, such as decision trees, which can deal with heterogeneous

with this problem is to use non-parametric models, such as decision trees, which can deal with heterogeneous data quite efficiently.

9. Your linear regression doesn't run and communicates that there is an infinite number of best estimates for the regression coefficients. What could be wrong?

This condition arises when there is a perfect correlation (positive or negative) between some variables. In this case, there is no unique value for the coefficients, and hence, the given condition arises.

10. How do you interpret the residual vs fitted value curve?

The residual vs fitted value plot is used to see whether the predicted values and residuals have a correlation or not. If the residuals are distributed normally, with a mean around the fitted value and a constant variance, our model is working fine; otherwise, there is some issue with the model. The most common problem that can be found when training the model over a large range of a dataset is heteroscedasticity(this is explained in the answer below). The presence of heteroscedasticity can be easily seen by plotting the residual vs fitted value curve.

Now Rest with this Quest :)
