Multiple 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 is robust regression?

A regression model should be robust in nature. This means that with changes in a few observations, the model should not change drastically. Also, it should not be much affected by the outliers.
A regression model with OLS (Ordinary Least Squares) is quite sensitive to the outliers. To overcome this problem, we can use the WLS (Weighted Least Squares) method to determine the estimators of the regression coefficients. Here, less weights are given to the outliers or high leverage points in the fitting, making these points less impactful.

2. What is the generalized linear model?

The generalized linear model is the derivative of the ordinary linear regression model. GLM is more flexible in terms of residuals and can be used where linear regression does not seem appropriate. GLM allows the distribution of residuals to be other than a normal distribution. It generalizes the linear regression by allowing the linear model to link to the target variable using the linking function. Model estimation is done using the method of maximum likelihood estimation.

3. How can learning curves help create a better model?

Learning curves give the indication of the presence of overfitting or underfitting.
In a learning curve, the training error and cross-validating error are plotted against the number of training data points. A

If the training error and true error (cross-validating error) converge to the same value and the corresponding value of the error is high, it indicates that the model is underfitting and is suffering from high bias.
If there is a significant gap between the converging values of the training and cross-validating errors, i.e. the cross-validating error is significantly higher than the training error, it suggests that the model is overfitting the training data and is suffering from a high variance.

4. What is VIF? How do you calculate it?

Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset. It is calculated as—

Here, VIFj is the value of VIF for the jth variable,

Rj2 is the R2 value of the model when that variable is regressed against all the other independent variables.
If the value of VIF is high for a variable, it implies that the R2 value of the corresponding model is high, i.e. other independent variables are able to explain that variable. In simple terms, the variable is linearly dependent on some other variables.

5. What do you mean by adjusted R2? How is it different from R2?

Adjusted R2, just like R2, is a representative of the number of points lying around the regression line. That is, it shows how well the model is fitting the training data. The formula for adjusted R2 is –

Here, n is the number of data points, and k is the number of features.
One drawback of R2 is that it will always increase with the addition of a new feature, whether the new feature is useful or not. The adjusted R2 overcomes this drawback. The value of the adjusted R2 increases only if the newly added feature plays a significant role in the model.

6. What are the disadvantages of the multiple linear model?

- Linear regression is sensitive to outliers which may affect the result.

- Over-fitting

- Under-fitting

7. Why do we square the error instead of using modulus?

It’s true that one could choose to use the absolute error instead of the squared error. In fact, the absolute error is often closer to what we want when making predictions from our model. But, we want to penalize those predicted values which is contributing the maximum error. Moreover looking a little deeper, the squared error is everywhere differentiable, while the absolute error is not (its derivative is undefined at 0). This makes the squared error more amenable to the techniques of mathematical optimization. To optimize the squared error, we can just set its derivative equal to 0 and solve. To optimize the absolute error often requires more complex techniques. Actually we find the Root Mean Squared Error so that the unit of RMSE and the dependent variable are equal.

8. Explain Ordinary Least Squares Regression in brief.

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable. The method estimates the relationship by minimizing the sum of the squares of the difference between the observed and predicted values of the dependent variable configured as a straight line. OLS regression is used in bivariate model, that is, a model in which there is only one independent variable (X) predicting a dependent variable (Y). However, the logic of OLS regression can also be used in multivariate model in which there are two or more independent variables.

9. Which evaluation technique should you prefer to use for data having a lot of outliers in it?

Mean Absolute Error(MAE) is preferable to use for data having too many outliers in it because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and starts penalizing the outliers by squaring the residuals.

10. What are the flaws in R-squared? Can R^2 be negative?

There are two major flaws:
Problem 1: R^2 increases with every predictor added to a model. As R^2 always increases and never decreases, it can appear to be a better fit with the more terms we add to the model. This can be completely misleading.

Problem 2: Similarly, if our model has too many terms and too many high-order polynomials we can run into the problem of over-fitting the data. When we over-fit data, a misleadingly high R^2 value can lead to misleading predictions.

Yes, R^2 can be negative
If the sum of squared error of the mean line(SSM) is greater than the regression line(SSR), R squared will be negative.

Now Rest with this Quest :)