

# General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

This is one of the supervised learning. Here, we find the most correlated attributes and then, we try to draw a linear line in between the datapoints. Based on some statistical calculations, we try to find out the best available line and then we try to predict the target result.

2. Explain the Anscombe's quartet in detail. (3 marks)

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets

3. What is Pearson's R? (3 marks)

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by  $r$ . Values for  $r$  between +1 and -1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range

## ***Normalization/Min-Max Scaling:***

- *It brings all of the data in the range of 0 and*

## ***Standardization Scaling:***

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean ( $\mu$ ) zero and standard deviation one ( $\sigma$ ).*

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If there is perfect correlation, then  $VIF = \infty$ . This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ , which leads to  $1/(1-R^2) = \infty$ . To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.