General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear Regression Algorithm is a machine learning algorithm based on supervised learning. We have covered supervised learning in our previous articles. Here we are going to focus on Linear regression. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

Regression analysis is used for three types of applications:

1. Finding out the effect of Input variables on Target variable.
2. Finding out the change in Target variable with respect to one or more input variable.
3. To find out upcoming trends.

Here are the types of regressions:

1. Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Polynomial Regression

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

2. Explain the Anscombe’s quartet in detail.

Answer: **Anscombe’s Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties.**There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.

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?

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Answer: Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

Suppose we have two features of weight and price, as in the below table. The “Weight” cannot have a meaningful comparison with the “Price.” So the assumption algorithm makes that since “Weight” > “Price,” thus “Weight,” is more important than “Price.”

Feature scaling is essential for machine learning algorithms that calculate **distances between data**. If not scale, the feature with a higher value range starts dominating when calculating distances.

The ML algorithm is sensitive to the "**relative scales of features,**" which usually happens when it uses the numeric values of the features rather than say their rank.

In many algorithms, when we desire **faster convergence**, scaling is a MUST like in Neural Network.

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

Answer: If there is perfect correlation, then **VIF** = **infinity**. A large **value of VIF** indicates that there is a correlation between the variables. If the **VIF** is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.