1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical algorithm used for modeling the relationship between a dependent variable (also known as the response or outcome variable) and one or more independent variables (also known as predictors or explanatory variables) that are assumed to have a linear relationship with the dependent variable.

The goal of linear regression is to find the line of best fit that summarizes the relationship between the dependent variable and the independent variables. This line of best fit is represented by an equation of the form:

y = b0 + b1*x1 + b2*x2 + ... + bn*xn

where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0 is the y-intercept (the point where the line intercepts the y-axis), and b1, b2, ..., bn are the coefficients (slopes) of the line for each independent variable.

The linear regression algorithm involves two main steps: training and testing.

Training:

1. Data collection: Collect the data for the dependent variable and independent variables.
2. Data preprocessing: Perform data cleaning, data transformation and feature selection techniques.
3. Model building: Build the linear regression model by fitting the data to the equation using an optimization technique such as least squares, where the goal is to minimize the sum of squared errors between the predicted and actual values of the dependent variable.

Testing:

1. Splitting the dataset: Split the dataset into a training set and a testing set.
2. Model evaluation: Evaluate the performance of the model on the testing set using metrics such as R-squared, mean squared error, root mean squared error, and adjusted R-squared.
3. Model refinement: Refine the model by tweaking the values of the coefficients and testing again until the desired level of performance is achieved.

Once the model is built and tested, it can be used for making predictions on new data by simply plugging in the values of the independent variables into the equation to obtain a predicted value for the dependent variable.

2. What are the assumptions of linear regression regarding residuals?

Answer:-
a) ***Normality assumption***: It is assumed that the error terms, ε(i), are normally distributed. If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data
b) ***Zero mean assumption***: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
c) Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2 . This assumption is also known as the assumption of ***homogeneity*** *or* ***homoscedasticity.***

d) ***Independent error assumption***: It is assumed that the residual terms are independent of each other, i.e. their pair-wise co-variance is zero. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves.
If some correlation is present, it implies that there is some relation that the regression model is not able to identify
If the independent variables are not linearly independent of each other, the uniqueness of the least square's solution (or normal equation solution) is lost.

3. What is the coefficient of correlation and the coefficient of determination?

Answer:

The **coefficient of determination** or R squared method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.

- The coefficient of determination is the square of the correlation(r), thus it ranges from 0 to 1.
- With linear regression, the coefficient of determination is equal to the square of the correlation between the x and y variables.
- If R^2 is equal to 0, then the dependent variable cannot be predicted from the independent variable.
- If R^2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error.
- If R^2 is between 0 and 1, then it indicates the extent that the dependent variable can be predictable. If R^2 of 0.10 means, it is 10 percent of the variance in the y variable is predicted from the x variable. If 0.20 means, 20 percent of the variance in the y variable is predicted from the x variable, and so on.

The value of R^2 shows whether the model would be a good fit for the given data set. In the context of analysis, for any given per cent of the variation, it(good fit) would be different. For instance, in a few fields like rocket science, R^2 is expected to be nearer to 100 %. But R^2 = 0(minimum theoretical value), which might not be true as R^2 is always greater than 0(by Linear Regression).

4. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

5. What is Pearson's R?

Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

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

Answer:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

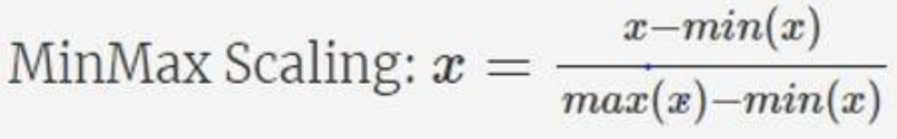
Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

*It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*

Normalization/Min-Max Scaling:

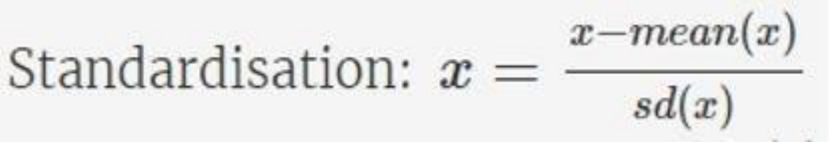
- *It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.*



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

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



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- ***sklearn.preprocessing.scale** helps to implement standardization in python.*

- *One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.*

7. 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. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

8. What is the Gauss-Markov theorem?

   Answer:

   The Gauss Markov theorem says that, under certain conditions, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (BLUE), that is, the estimator that has the smallest variance among those that are unbiased and linear in the observed output variables.

Assumptions

The regression model is where:

- is an vector of observations of the output variable (is the sample size);
- is an matrix of inputs (is the number of inputs for each observation);
- is a vector of regression coefficients;
- is an vector of errors.

The OLS estimator of is

We assume that:

1. has full-rank (as a consequence, is invertible, and is well-defined);
2. ;
3. , where is the identity matrix and is a positive constant.

OLS is linear and unbiased

First of all, note that is linear in . In fact, is the product between the matrix and , and matrix multiplication is a linear operation.

It can easily be proved that is unbiased, both conditional on , and unconditionally, that is,

9. Explain the gradient descent algorithm in detail.
   Answer:

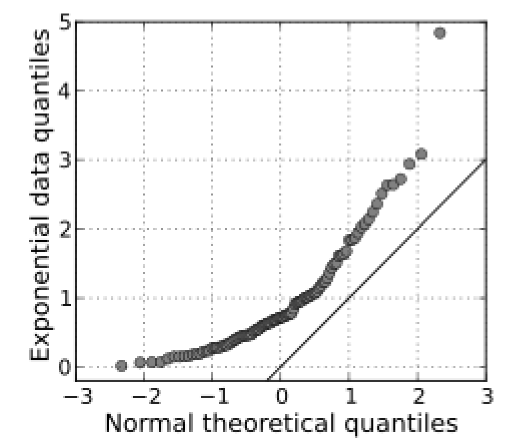
Gradient descent is an optimization algorithm which is commonly-used to train machine learning models and neural networks. Training data helps these models learn over time, and the cost function within gradient descent specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates. Until the function is close to or equal to zero, the model will continue to adjust its parameters to yield the smallest possible error. Once machine learning models are optimized for accuracy, they can be powerful tools for artificial intelligence (AI) and computer science applications.

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

Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.