1. Explain the linear regression algorithm in detail.

Explanation:-

Regression is the most commonly used predictive analysis model and is used to predict a continuous variable. Linear Regression can be classified into two types – Simple and Multiple Linear regression.

Simple Linear Regression is the most basic type of linear regression and explains the relationship between a dependent and one independent variable using a straight line.

Mathematically we can write this linear relationship as

Y0 = B0 + B1*X

Where Y is the continuous variable to be predicted and X is the independent variable that has a linear relationship with the dependent variable X.

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables.

The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Mathematically we can write this linear relationship as

Y0 = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4…. + Bn*Xn .

In practice the B -coefficients are unknown, to make predictions we use the data to estimate the B coefficients.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. In other words, we try to find the B coefficient's such that the resulting line is as close as possible to the n data points.

Error is the distance between the points to the regression line and can be represented as below

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

The strength of the linear regression model can be assessed using 2 metrics:

- R 2 or Coefficient of Determination
- Residual Standard Error (RSE)

R 2 or Coefficient of Determination

R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual

outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as: R 2 = 1 - (RSS / TSS)

RSS = Residual sum of Squares

TSS = Sum of errors from the data

Ordinary Least Squares

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.

The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations.

It is unusual to implement the Ordinary Least Squares procedure yourself unless as an exercise in linear algebra. It is more likely that you will call a procedure in a linear algebra library. This procedure is very fast to calculate.

Gradient Descent

When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

This operation is called Gradient Descent and works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

When using this method, you must select a learning rate (alpha) parameter that determines the size of the improvement step to take on each iteration of the procedure.

Gradient descent is often taught using a linear regression model because it is relatively straightforward to understand. In practice, it is useful when you have a very large dataset either in the number of rows or the number of columns that may not fit into memory .

Below are the assumptions of Linear Regression

1. Linear Relationship between the features and target:
2. Little or no Multicollinearity between the features:
3. Homoscedasticity Assumption

4. Normal distribution of error terms
5. No autocorrelation of residuals

Steps for Linear Regression:

1. Importing the dataset and understanding the data using EDA.
2. Data Manipulation for Modelling using Pandas or any other suitable library.
3. Splitting data into train and test
4. Rescaling the features using sklearn library.
5. Preparing X and Y
6. Feature selection using RFE.
7. Building Model using statsmodel or sklearn libraries
8. Finalizing the model
9. Making predictions
10. Model evaluation (Plot Actual vs Predicted /Plotting Error terms)

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

   Explanation: -
   Below are the assumptions of linear regression regarding residuals

   Homoscedasticity: - Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a goodway to check for homoscedasticity. There should be no clear pattern in the distribution

   Normal distribution of error terms: - Error terms should be normally distributed with mean approximately equal to 0.

   No autocorrelation in the residuals:
   Autocorrelation occurs when the residual errors are dependent on each other. The presence of correlation in error terms drastically reduces model's accuracy.

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

   Coefficient of correlation: - It is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way.

Coefficient of determination: - R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

4. Explain the Anscombe's quartet in detail.
   Explanation: -

   Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics.

   But things change completely, **,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

5. What is Pearson's R?

   Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

   The Pearson's correlation coefficient varies between -1 and +1 where:
   r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
   r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
   r = 0 means there is no linear association

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

   Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Min-Max Normalization (Normalized scaling): This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

   VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

   VIF is given by (1/1-R^2) where R is the coeff of correlation and indicates the relationship between two variables say x and y . . It can go between -1 and 1. 1 indicates that the two variables are moving in unison.

   If the two variables are perfectly correlated the R value comes out to be 1 substituting the value for R in such case in the formula mentioned above we see that the denominator comes out to 0 which leads the VIF value to be formulated as infinite.

8. What is the Gauss-Markov theorem?
   Gauss Markov theorem states that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives the best linear unbiased estimate (BLUE) possible.

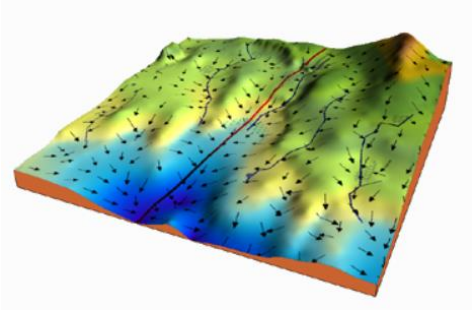
   Gauss Markov Assumptions
   There are five Gauss Markov assumptions (also called conditions):

   1. Linearity: the parameters we are estimating using the OLS method must be themselves linear.
   2. Random: our data must have been randomly sampled from the population.

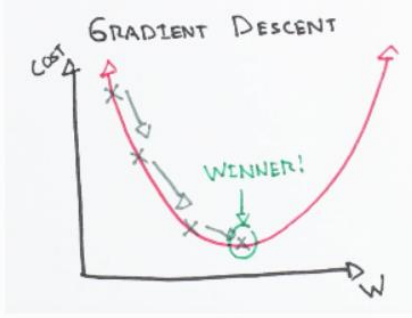
3. Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
4. Exogeneity: the regressors aren't correlated with the error term.
5. Homoscedasticity: the error of the variance is constant.

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.
In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression.
Consider the 3-dimensional graph below in the context of a cost function. Our goal is to move from the mountain in the top right corner (high cost) to the dark blue sea in the bottom left (low cost). The arrows represent the direction of steepest descent (negative gradient) from any given point–the direction that decreases the cost function as quickly as possible.



Starting at the top of the mountain, we take our first step downhill in the direction specified by the negative gradient. Next we recalculate the negative gradient (passing in the coordinates of our new point) and take another step in the direction it specifies. We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum.



Learning rate

The size of these steps is called the *learning rate*. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

Cost function

A Loss Functions tells us "how good" our model is at making predictions for a given set of parameters. The cost function has its own curve and its own gradients. The slope of this curve tells us how to update our parameters to make the model more accurate.

Given the cost function below

$$\boxed{f(m,b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2}$$

The Gradient can be calculated as

$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

To solve for the gradient, we iterate through our data points using our new m and b values and compute the partial derivatives. This new gradient tells us the slope of our cost function at our current position (current parameter values) and the direction we should move to update our parameters. The size of our update is controlled by the learning rate.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

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.