Shubhangi_Dhikale_35

1.What is Linear regression?

In [ ]:
```
1.Linear regression is a supervised machine learning algorithm.
2.It is predictive model which is used to predict the linear relation
between independent and depedent variable.
3.It will find regression related problem.
4.Basically there are two types of linear regression,
    1) simple linear regression
    2) multiple linear regression
In simple linear regression there are only one independent variable.
Equation of simple linear equation is y=mx+c
where, y is dependent variable
x is independent variable
m is slope
c is intercept
In multiple linear regression there are two or more independent variable.
Equation of multiple linear equation is y=m1x1+m2x2+m3x3+....+mnxn+c
where,y is dependent variable
x1,x2,x3....xn are independent variables
m1,m2.....mn are slopes
```

2.How do you represent a simple linear regression?

```
1.In multiple linear regression their is one is the dependent and one
independent variable.
2.simple linear regression is represented by equation y=mx+c where,
y=dependent variable m=slope
of line x=independent variable c=intercept
3.The relationship in simple Linear Regression equation
is linear or a sloped is straight line then it is called Simple Linear Regression.
4.In simple linear regression the dependent values must be continuous or real values.
5.Consider the following graph it is simple linear regression graph.
In this example salary and year of experience are two variables.
Salary is dependent variable and year of experience is independent variable.
As experience of work increases the salary also increases.
Hence we predict the next year of experience salary with the help of regression.
```

3.What is multiple linear regression

In [ ]:
```
Multiple linear regression is a statistical technique that uses several
explanatory/input variables to predict the
outcome of a response/target variable.
The equation of multiple linear regression is
y=m1x1+m2x2+m3x3+.....+mnxn+c
where,y is dependent variable
x1,x2,x3...xn are input variables
m1,m2....mn are slopes
c is intercept
```

4.What are the assumptions made in the Linear regression model?

```
1.linearity:In linear regression their is must the relationship between the independent
and dependent variables to be linear.
It will check by the scatter plot.with the help of correlation(pearson correleation coefficient)
we can find the direction of the linear variables.
formula: corr=covariance/(prod of std)
2.Independence:Input variable should not be dependent to each other.
3.No multicolinearity:Input variable should not be highly correlated to each other.
4.Normality:normal distribution of the residual or error.
Normality means variables are normally distributed.
we can check normality with Shapiro-Wilk test,
Kolmogorov- Smirnov test, skewness, kurtosis, histogram,
box plot, P-P Plot, Q-Q Plot,and mean with SD.
5.Homoscedasticity:Homoscedasticity in a model means that the error is
constant along the values of the dependent variable.
The best way for checking homoscedasticity is to make a scatterplot with
the residuals against the dependent variable.
```

5.What if these assumptions get violated?

In [ ]:
```
Linearity: If you fit a linear model to a non-linear, non-additive data set,
the regression algorithm would fail to capture the trendmathematically,
thus resulting in an inefficient model.

Normal Distribution of error terms: If the error terms are non- normally distributed,
confidence intervals
may become too wide ornarrow. Once confidence interval becomes unstable,
it leads to difficulty in estimating
coefficients based on minimization of least squares.
Presence of non – normal distribution suggests that there are a few unusual data points
which must be studied closely to make a better model.

Multicollinearity: This phenomenon exists when the independent variables
are found to be moderately or highly correlated.
In amodel with correlated variables, it becomes a tough task to
figure out the true relationship of a predictors with response variable.

Heteroskedasticity: The presence of non-constant variance in the
error terms results in heteroskedasticity.
Generally, non-constant variance arises in presence of outliers or extreme leverage values.
Look like, these values get too much weight,
thereby disproportionately influences the model's performance.
When this phenomenon occurs, the confidence interval for out of
sample prediction tends to be unrealistically wide or narrow.
```

6.What is the assumption of homoscedasticity?

```
Homoscedasticity: Residual should follow the homoscedasticity behaviour.
Residual should be in constant manner then this is homoscedasticity.
```

7.What is the assumption of normality

In [ ]:
```
Normality:
It is the Normal distribution of the residual or error.
If the residuals are not skewed,that means that the assumption is satisfied.
we draw a histogram of the residuals and then examine the normality of the residuals.
This are depend on standard deviation,
if some of the data points are closed to the mean point then we can say that it is low standard deviation.
if some of the data points are far away from the mean point then we can say that it is high standard deviation.
```

8.How to prevent heteroscedasticity?

In [ ]:
```
There are three common ways to fix heteroscedasticity:
1.Transform the dependent variable
One way to fix heteroscedasticity is to transform the dependent variable in some way.
One common transformation is to simply take the log of the dependent variable.
For example, if we are using population size (independent variable) to predict the number
of flower shops in a city (dependent variable),
we may instead try to use population size to predict the log of the number of flower shops in a city.
2.Redefine the dependent variable Another way to fix heteroscedasticity is to redefine the dependent variable.
One common way to do so is to use a rate for the dependent variable, rather than the raw value.
For example, instead of using the population size to predict the number of flower shops
in a city, we may instead use population size to predict the number of flower shops per capita.
3.Use weighted regression Another way to fix heteroscedasticity is to use weighted regression.
This type of regression assigns a weight to each data point based on the variance of its fitted value.
Essentially, this gives small weights to data points that have higher variances,
which shrinks their squared residuals. When the proper weights are used, this can eliminate the problem of heteroscedasticity.
```

9.What does multicollinearity mean?

In [ ]:
```
Multicollinearity means the high correlation between the independent variables.
```

10.What are feature selection and feature scaling?

In [ ]:
```
feature scaling:
Feature scaling is a method used to standardize the range of independent variables or features of data.
or Feature scaling is the method through which we can scale down the numeric features in the same scale.

Feature selection:
Feature selection one of the main components of feature engineering,is the process of selecting the most important
features to input in machine learning algorithms.
```

11.How to find the best fit line in a linear regression model?

In [ ]:
```
We can find the Best Fit Line with the help of gradient descent algorithm in linear regression model.
It uses the babysteps having minimal learning rate which is 0.001 to achieve global minima.
So we can get the best m and c values and reduces the cost function to get best fit line or linear line or regression line.
```

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

In [ ]:
```
we will doing square the error instead of using modulus because the squared function is differentiable evrywhere,
while the absolute error is not differentiable at all the points in its domain.This makes the squared error
more preferable to the techniques of mathematical optimization.
```

13.What are techniques adopted to find the slope and the intercept of the linear regression line which best fits the model?

In [ ]:
```
There are two methods:
1) Ordinary Least Sqaures
2) Greadient Descent
```

14.What is cost Function in Linear Regression?

In [ ]:
```
For the Linear regression model, the cost function will be the minimum of the Root Mean Squared Error of the model,
obtained by subtracting the predicted values from actual values.
The cost function will be the minimum of these error values.

cost function = sum[(Ya - Yp)^2]/N
where,
Ya : Actual Datapoint
Yp : Predictive Datapoint
N : Total Samples or Datapoints
```

15.briefly explain gradient descent algorithm?

In [ ]:
```
As we know if we have maximum error rates in the data, then accuracy might get impacted.so, this will help
to reduce the cost function or mean squared error.
by using partial derivative it will also useful to find best m and c values.
for this gradient descent algorithm do follow baby steps so that we reach to global minima.
At that point mean squared error is reduce and we get best m and c values.
For getting m and c values we have getting new m and c values by using formula,
M new=M old-L* dMSE/dM
C new=c old-L*dMSE/dC
where L is learning rate,It should be 0.001.
If L is low more time required and L is large model will be overfitted.
It uses minimal learning rate to acheive the global minima so where we can get M and C values and lower cost function
or error rates.
```

16.How to evaluate regression models?

In [ ]:
```
To eveluate the model we have some model evaluation parameters:
1)Mean Squared Error :
MSE = sum[(Ya - Yp)'2]/N
The mean squared error (MSE) tells you how close a regression line is to a set of datapoints.
2)Mean Absolute Error : MSE = sum[(Ya - Yp)]/N
3)Root mean Squared Error : RMSE = np.sqrt(MSE)
4)R2 Score :
It helps to decide the final model.It increases the R2 Score of the model if you add the
correalated features or non-correlated features.This is drawback of R2 score.
Though we use R2 score only because it increase very less value like 0.0012 like this.So this is considerable.

R2 Score = 1 - [SSE/SST]
where ,
SSE = sum[(Ya - Yp)'2] = Sum of Squared Error SST = sum[(Ya - Ymean)'2] = Total Error
SST = SSE + SSR
SSR = sum[(Yp - Ymean)'2]

5)Adjucent R2 Score :
Adjucent R2 Score = R2 score-[(k-1)/(n-k)](1-R2 score) where,
n : No. of samples
k : No. of attributes
The Drawback of R2 score is overcome by Adjucent R2 Score.
```

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

In [ ]:
```
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.
```

18.What is residual? How is it computed?

In [ ]:
```
Residual is also called Error. It is the difference between the predicted y value and the actual y value.
Residual = Actual y value - Predicted y value.
```

19.What are SSE, SSR, and SST? and What is the relationship between them?

In [ ]:
```
1) SSE:
It is sum of Squared difference between y actual and y predicted values.
It is also calles as residual sum of squares(RSS).
SSE=sum(Y actual - Y predicted)^2
2) SSR:
It is sum of squared difference between y predicted and y mean values.
SSR=sum(Y pred-Ymean)^2
3) SST:
It is sum of squared difference between y actual and y mean values.
SST=SSE+SSR
```

20.What's the intuition behind R-Squared?

In [ ]:
```
R2 score is used to evaluate the performance of a linear regression model.We can say that it hepls to decide the final model.
Higher value of R2 score is preferable.
R2 score is also called the coefficient of determination.
R2 score tells us what percent of the variability in the y variable is accounted for by the regression on the x variable.
The value of R2 score varies from 0 to 1.
Ideal value for R2 score is '+1', which gives best model. If R2 score is '0' then that is worst model.
```

21.What does the coefficient of determination explain?

In [ ]:
```
The coefficient of determination is the measure of the variance in the response variable 'Y'
that can be predicted using the predictor variable 'x'.
or The coefficient of determination is a measurement used to explain how much variability of one factor
can be caused by its relationship to another related factor.
It is also called as r2score
r2 score=1-SSE/SST
if SSE>SST then r2score is positive
if SSE<SST then r2score is negative
if r2 score=1 then it is best value
if r2 score=0 then it is worst model.
```

22.Can R^2 be negative?

In [ ]:
```
The formula for R2 score is as follows: R2 Score = 1 - [SSE/SST]
If SSE > SST then R2 Score is negative.
yes,it can be negative.
```

23.What are the flaws in R-squared?

In [ ]:
```
If we have 5 features and adding one correlation feature then R2 value increases again
if one non-correlated feature is added then also R2 score value increases which is not desired condition.
This is the drawback of R2 score.
```

24.What is adjusted R^2?

In [ ]:
```
The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not
significant in a regression model.In other words, the adjusted R-squared shows whether adding additional
predictors improve a regression model or not.
formula for adjusted R^2:
adjusted R^2=R^2-(k-1)/(n-k)*(1-R^2)
where,
n=total sample size
k=Number of features or attributes
```

25.What is the Coefficient of Correlation: Definition, Formula

In [ ]:
```
Linear relation between two variables that is input variables and output variables.
It is denoted as R.
R value lies between -1 to 1
formula :
R=sum(xi-xbar)*(yi-ybar)/sqrt[(sum(xi-xbar)^2)*(sum(yi-ybar)^2)]
```

26.What is difference between Correlation and covariance?

In [ ]:
```
1.covariance illustrates the degree to which two variables vary with respect to each other, while correlation determines the
strength and direction of this relationship.
2.Covariance and correlation are interlinked with each other.
In simple terms, correlation refers to the scaled version of covariance.
3.Covariance deals with the linear relationship of only two variables in the dataset,
whereas correlation can involve two or multiple
variables or data sets and their linear relationships.
4.Correlation coefficients are standardized, therefore displaying an absolute value within a definite
range from -1 to 1. On the other hand, covariance values are not standardized and use an indefinite range from -= to + .
5.Correlation is dimensionless where covariance is in units.
6.Covariance is affected by the change in scale.Conversely, correlation is not affected by the change in scale.
```

27.What is the relationship between R-Squared and Adjusted R-Squared?

In [ ]:
```
Adjusted R2 score will always be less than or equal to R2 score.
If we have 5 features and adding one correlation feature then R2 value
increases again if one non-correlated feature is added then also R2 score
value increases which is not desired condition.This is the drawback of R2 score.
This Drawback can be overcome by Adjucent R2 score.
Formula:
Adjucent R2 Score = R2 score-[(k-1)/(n-k)](1-R2 score)
```

28.What is the difference between overfitting and underfitting?

In [ ]:
```
When model neither learns from the training dataset and not from testing dataset. That is nothing but underfitting.
On other hand when model learns well from training dataset and performs well on unseen dataset or testing dataset,
that is overfitting.
In overfitting the difference between the trainning accuaracy and testing accuracy is less but that
difference is more or large in underfitting.
```

29.How to identify if the model is overfitting or underfitting?

In [ ]:
```
When the difference between the trainning accuaracy and testing accuracy is less
then this is overfitting and When the difference between the trainning accuaracy and testing accuracy is more or large ,
that is underfitting.
```

30.How to interpret a Q-Q plot in a Linear regression model

In [ ]:
```
Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability
distributions by plotting their quantiles against each other. If the two distributions which we are
comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.
Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line.
We also call it the 45-degree line in statistics. It entails that each of our distributions
has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be
skewed when compared to the other.
```