Interview Questions — Oct-15, 2022

① what is linear regression?

A. It is a statistical method that is used for predictive analysis. Linear algorithm shows a linear relations between a dependent (y) and one or more independent (x) variables, hence called as linear regression.

② How we can Calculate error in linear regression?

A. Difference between predicted value and real (Actual) value.

③ Difference between loss and cost function?

A. The loss function capture the difference between the actual and predicted values for a single record whereas cost functions aggregate the difference for the entire training dataset.

④ Difference between MAE, MSE and RMSE?

A. MAE :- Mean Absolute Error

Mean : average
Absolute : without direction, get rid of any negative signs

MSE :- Mean Squared Error

Mean : average
Squared : Square the errors

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x)^i - y^i)^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|h_\theta(x)^i - y^i|$$

RMSE :- Root Mean Squared Error

- Just square root of MSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h_\theta(x)^i - y^i)^2}$$

* MSE & RMSE are very useful when we wanna to see if the outliers are messing with our predictions.

⑤ How Gradient Descent work? in Linear Regression?

A. Linear Regression ⇒ ① Calculation → $y = mx + c$

② Loss / Error → Cost function [MSE, MAE, RMSE]

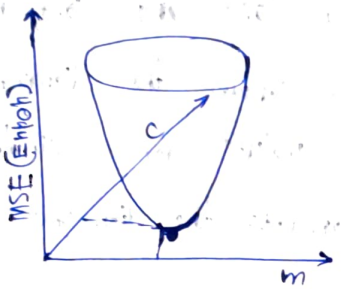
③ Optimize (m, c) → To obtain minimum loss

A. we have to find out the best fit line which gets minimum loss for all the predicted values and Actual values.

for finding the best fit line we have to get the Correct 'm' and 'c' values but we are not trying all the permutations and Combinations of 'm' & 'c'.

for that we use "Gradient Descent Algorithm" that finds the best fit line for a given training dataset in Smaller number of iterations.

If we plot 'm' and 'c' against MSE, it will acquire a bowl shape.



for some Combination of 'm' and 'c', we will get the least Error (MSE). That Combination of 'm' and 'c' will give us the best fit line.

⑥ Explain what is the intercept term?

A. In the best fit line when $x=0$ where this line meets the y-axis that is called intercept (c).

⑦ write all the assumption for Linear Regression?

A. There are 4 assumptions associated with linear regression model:-

1. Linearity - The relationship between X and the mean of y is linear
2. Homoscedasticity - The variance of residual is the same for any value of X.
3. Independence - Observations are independent of each other.
4. Normality - for any fixed value of X, Y is normally distributed.

⑧ How is hypothesis testing using in linear regression?

A. Hypothesis testing is used to Confirm if our beta Coefficients are significant in a linear regression model. Every time we run the linear regression model, we test if the line is significant or not by checking if the Coefficient is significant.

⑨ How would you decide the importance of variable for the multi-variate regression?

A. Based on correlation of the variable with target feature. We can decide the importance of the variable.

⑩ R vs. Adjusted R^2?

A. Once we have predicted the values by using linear regression algorithm. what next? So the next step was to evaluate its performance.

So, choosing the most appropriate Evaluation metric is a crucial. Then we came across two important metrics:

R-squared
Adjusted R-squared } a part from MAE/MSE/RMSE

Difference between these two metrics:

R-Squared: It measures the proportion of the variation in our dependent variable explained by all of our independent variables in the model.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$\{SS_{res}\}$ → Sum of squares of residuals

$\{SS_{tot}\}$ → total sum of squares

$SS_{res} + SS_{reg} = SS_{tot}$

SS_{res} = measures unexplained variation

SS_{reg} = measures explained variation

- R-Squared is also called coefficient of determination.
- It lies blw 0% - 100%.
- A R-Squared value of 100% means the model explains all the variation of the target variable.
- A value of 0% measures zero predictive power of the model.

Higher R-Squared value ∝ better the model

Adjusted R-squared: It measures the proportion of variation explained by only those independent variable that really help in explaining the dependent variable.

df = degree of freedom
df = n-1

$$\bar{R}^2 = 1 - \frac{SS_{res}/df_e}{SS_{tot}/df_t}$$

The only difference blw R-Squared & Adjusted R-Squared equation is degree of freedom.

$df_e = N-1$ [estimate the population variance of the dependent variable]

$df_t = N-P-1$ [estimate of the underlying population error variance]

So, $\bar{R}^2 = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$

where, R^2 = Sample R-square
P = Number of predicators
N = Total Sample Size

Difference between R-Squared & Adjusted R-squared:

Every time we add a independent variable to a model, the R-squared increases, even the independent variable is insignificant. It never declines. whereas Adjusted R-Squared increases only when independent variable is significant and affects dependent variable.