Machine Learning Algorithms

To crack any Machine Learning Interview.
-Er.Rehan Sagar

- Classification - Only Discrete value is available as a Target column.
  — In target column, value could be 0/1, True/False, Binary/Multiple No.

  For eg. 1.Cancer Detection - The target column will be → 0(Not cancerous) or 1(Cancerous)
  2.Wine Quality - The target column will be in the range of 0 - 5 by suggesting the quality of a wine.

This type of examples or problem can be solved with the help of classification algorithm.

- Regression - Continuous value is available as a Target column.
  — In target column, value could be anything.Some value can be discrete,some can be decimal.

  For eg. 1. House Price Prediction - The target column will be → 12.5cr, 0.75cr, 10lakh, 75.46 lakh, 2.33cr.
  2. Salary Prediction - The target column will be → 12.75k, 7.5k etc

This type of problems can be solved with the help of Regression algorithm.

Linear Regression

— By the name of linear regression, we can easily understood the problem come in this chapter would be contain continuous values and therefore we follow the Regression Algorithm.
So first algorithm,we put light on it is “Linear Regression”.

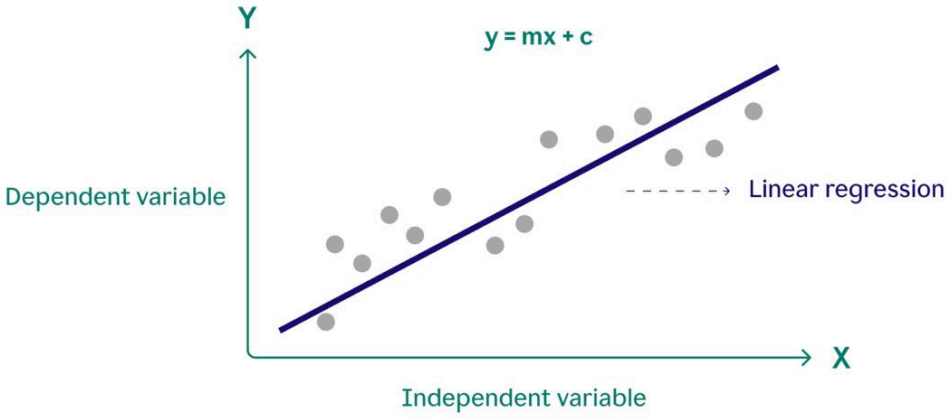
Height(x)	Weight(y)
5.5	50
6.0	65
4.9	45
5.2	53
5.4	58
5.7	??

If x = 5.5 then y = 50, x = 6.0 then y = 65, so if x = 5.7 then y can be ?

Linear Regression is a method to predict a number using a straight line called best fit line. **(R)**

A supervised learning algorithm used to predict continuous values by fitting a straight line. **(I)**

Q.1 What is Best fit line?
In linear regression, the best-fit line is the straight line that most accurately represents the relationship between the independent variable - x (input) and the dependent variable - y (output). It is the line that minimizes the difference between the actual data points and the predicted values from the model.



Mathematical equation for a straight line : y = mx + c

Where, y → Dependent variable
x → Independent variable
m → Slope
c → Intercept

The value of x is 5.5 and y is 50 but $\hat{y}$ is 54.

Therefore, Error = | $\hat{y}$ - y | where, $\hat{y}$ → Predicted Value
y → Actual Value

It's mean, Error = 54 - 50 = 4

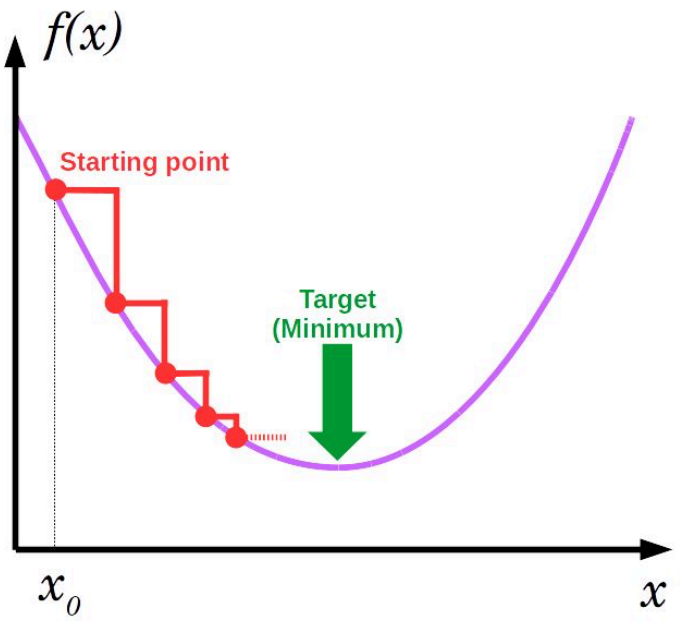
Now our first target is to minimize the error (Closure to zero).
To reduce the error we need to understand the concept of Gradient Descent.

Gradient Descent

— Gradient descent is an optimization technique used to train a linear regression model by minimizing the prediction error. It works by starting with random model parameters and repeatedly adjusting them to reduce the difference between predicted and actual values.

Q.1 How it works:

1. Start with random values for slope and intercept.
2. Calculate the error between predicted and actual values.
3. Find how much each parameter contributes to the error (gradient).
4. Update the parameters in the direction that reduces the error.
5. Repeat until the error is as small as possible.



Interview question asked in MNC's and Big tech Giant.

Q.1 What happens if learning rate is too high?

→ Model may overshoot meant to be if it start with the random parameter and jumping high values of slope and intercept then are lots of chances that model can skip the actual correct parameter and missed the actual global minima. **(RAW ANSWER)**

Or

→ **If the learning rate is too high, gradient descent takes very large steps and may overshoot the global minimum, causing the loss to oscillate or diverge instead of converging.(INTERVIEW)**

Q.2 When should you NOT use Linear Regression?

- Non linear relationship
- Too much outlier
- Categorical output

Q.3 Explain Residuals.

→ Residual = $y_{actual} - y_{predicted}$

Residuals tell us how wrong the model is for one data point.

House price prediction

House	Actual Price	Predicted Price	Residual
A	50 L	48 L	+2
B	60 L	63 L	−3

Positive Residual → Model is under - predicted

Negative Residual → Model is over - predicted

Note: Random Residual indicate model captures the true relationship and perform well.

Q.4 Assumptions of Linear Regression.

- Non Linear relationship - Linear regression works when the relationship between features and target is linear, meaning the rate of change of y with respect to x is constant, regardless of whether the slope is positive or negative.

- Independence - Independence refers to no data points should be correlate with any other data or we can say no data point can be depend on the previous data point.**(R)**

  Or

  Independence means that each observation and its error are not influenced by any other observation, especially previous ones in time-ordered data.

- Normality - Normality refers to residuals forming a normal distribution, which indicates that most errors are very small and large errors are very rare.**(R)**

Or

The normality assumption states that regression residuals follow a normal distribution, meaning most prediction errors are small and only a few are large.(I)

- Homoscedacity - Homoscedasticity is the assumption that the variance of residuals remains constant across all values of the independent variables.

Steps to analyze the dataset -
1. Target column is available or Not
2. If available which type of value is present?Is it discrete or continuous?
3.If it is discrete then go for Classification Algorithm OR If it is continuous value then go for Regression Algorithm.