ML:

1. What Are the Different Types of Machine Learning?

**Supervised Learning**

In, supervised machine learning, a model makes predictions or decisions based on past or labeled data. Labeled data refers to sets of data that are given tags or labels, and thus made more meaningful.

**Unsupervised Learning**

In unsupervised learning, we don't have labeled data. A model can identify patterns, anomalies, and relationships in the input data.

**Reinforcement Learning**

Using reinforcement learning, the model can learn based on the rewards it received for its previous action

**Linear Regression:**

A supervised machine learning algorithm finds the best linear-fit relationship on the given dataset, between independent and dependent variables. It is mostly done with the help of the **Sum of Squared Residuals Method**, known as the **Ordinary least squares (OLS) method**.

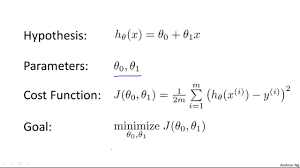
(or)

**Linear Regression** is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. (or) dependent variables has a constant relationship with independent variables.

Types: simple linear(y=mx+c) and multi-variable (y=w1x1+w2x2+w3x3+……..) , where=weights/coeff.

🡺main aim is to find a best fit line which passes/close to many no. of data points.

🡺draw a line and keep on changing the coefficient/weights to get best fit line.



Ѳ0 🡺intercept, Ѳ1🡺slope/coeff/weight

Here we need to keep on changing Ѳ1 to minimize he error/cost function

Cost function: A **cost function** in machine learning is a mechanism that returns the error between predicted outcomes and the actual outcomes.

**Gradient descent** is a first-order optimization algorithm.

In linear regression, this algorithm is used to optimize the cost function to find the values of the **βs (estimators)** corresponding to the optimized value of the cost function.

## **What are the basic assumptions of the Linear Regression Algorithm?**

* **Linearity:**There should be a relationship between the dependent and independent variables.
* **Homoscedasticity:** The error term has a constant variance.
* **Multicollinearity:**There is no multicollinearity between the features.
* **Independence:** Observations are independent of each other.
* **Normality:**The error(residuals) follows a normal distribution.

## **Explain the difference between Correlation and Regression**.

**Correlation:** It measures the strength or degree of relationship between two variables. It doesn’t capture causality. It is visualized by a single point.

**Regression:** It measures how one variable affects another variable. Regression is all about model fitting. It tries to capture the causality and describes the cause and the effect. It is visualized by a regression line.

## **Justify the cases where the linear regression algorithm is suitable for a given dataset.**

Scatter plot is used and in case of multi-variable LR, two-dimensional pairwise scatter plots, rotating plots, and dynamic graphs

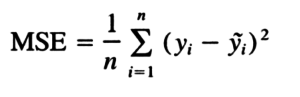
## **List down some of the metrics used to evaluate a Regression Model.**

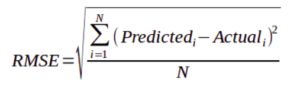
Mainly, there are five metrics that are commonly used to evaluate the regression models:

* Mean Absolute Error(MAE)
* Mean Squared Error(MSE)
* Root Mean Squared Error(RMSE)
* R-Squared(Coefficient of Determination)
* Adjusted R-Squared

**NOTE:** The sum of the residuals in a linear regression model is 0 since it assumes that the errors (residuals) are normally distributed with an expected value or mean equal to 0

**MSE (Mean Squared Error)**is defined as the average of all the squared errors(residuals) for all data points



**RMSE (Root Mean Squared Error)** is the square root of the average of squared differences between predicted and actual values. 

**NOTE**: Increment in RMSE is larger than MAE as the test sample size increases. In general, as the variance of error magnitudes increase, MAE remains steady but RMSE increases.

**MAE** is an average of absolute or positive differences between predicted values and the actual values.

(**MAE)** is preferred when we have too many outliers present in the dataset because MAE is robust to outliers

**Mean Absolute Percent Error**, which calculates the average absolute error in percentage terms

**\*\*\*NOTE**: the squared function is differentiable everywhere, while the absolute error is not differentiable at all the points in its domain(its derivative is undefined at 0). This makes the squared error more preferable to the techniques of mathematical optimization.

## **When should it be preferred to the Gradient Descent method instead of the Normal Equation in Linear Regression Algorithm?**

### Gradient descent:

* Needs hyper-parameter tuning for alpha (learning parameter).
* It is an iterative process.
* Time complexity- O(kn2)
* Preferred when n is extremely large.

### Normal Equation:

* No such need for any hyperparameter.
* It is a non-iterative process.
* Time complexity- O(n3) due to evaluation of XTX.
* Becomes quite slow for large values of n.

**where,**

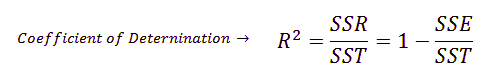
**‘k’**represents the maximum number of iterations used for the gradient descent algorithm, and

**‘n’** is the total number of observations present in the training dataset.

**Clearly**, if we have large training data, a normal equation is not preferred for use due to very high time complexity but for small values of ‘n’, the normal equation is faster than gradient descent.

## What are R-squared and Adjusted R-squared?

**R-square (R2)**, also known as the **coefficient of determination** measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model.



Adjusted-R2 square comes into the picture by penalizing those adding independent variables that do not improve your existing model.

## **What is Multicollinearity?**

### It is a phenomenon where two or more independent variables(predictors) are highly correlated with each other

### Reasons for Multicollinearity:

* Inaccurate use of dummy variables.
* Due to a variable that can be computed from the other variable in the dataset.

**Impacts of Multicollinearity:**

* Impacts regression coefficients i.e, coefficients become indeterminate.
* Causes high standard errors.

**Detecting Multicollinearity:**

* By using the correlation coefficient.
* With the help of Variance inflation factor (VIF), and Eigenvalues.

## **What is Heteroscedasticity? How to detect it?**

It refers to the situation where the variations in a particular independent variable are unequal across the range of values of a second variable that tries to predict it.

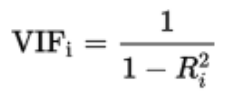
to detect heteroscedasticity, we can use graphs or statistical tests such as the **Breush-Pagan test** and **NCV test**, etc.

## What are the disadvantages of the linear regression Algorithm?

* **Assumption of linearity:**  not able to fit the complex problems with the help of a linear regression algorithm.
* **Outliers:**It is sensitive to noise and outliers.
* **Multicollinearity:** It gets affected by multicollinearity.

## **What is VIF? How do you calculate it?**

**VIF** stands for **Variance inflation factor**, which measures how much variance of an estimated regression coefficient is increased due to the presence of collinearity between the variables. It also determines how much multicollinearity exists in a particular regression model.



## **How is Hypothesis testing used in Linear Regression Algorithm?**

**1.**To check whether an independent variable (predictor) is significant or not for the prediction of the target variable. Two common methods for this are —

### By the use of p-values:

If the p-value of a particular independent variable is greater than a certain threshold (usually 0.05), then that independent variable is insignificant for the prediction of the target variable.

### By checking the values of the regression coefficient:

If the value of the regression coefficient corresponding to a particular independent variable is zero, then that variable is insignificant for the predictions of the dependent variable and has no linear relationship with it.

1. To verify whether the calculated regression coefficients i.e, with the help of linear regression algorithm, are good estimators or not of the actual coefficients.

## **Is it possible to apply Linear Regression for Time Series Analysis?**

**Yes**, we can apply a linear regression algorithm for doing analysis on time series data, but the results are not promising and hence is not advisable to do so.

time-series data have a pattern, such as during **peak hours**, **festive seasons**, etc., which would most likely be treated as outliers in the linear regression analysis.