Machine Learning Interview questions by Jeevan Raj

1. What is Machine Learning

Ans: Machine Learning is a Subset of Artificial Intelligence, In Simple words Machine learning is nothing but making machine to learn based on its past experience and machine learning algorithms use patterns and insights found in large datasets to make predictions and decisions.

2. Type of Machine Learning

Ans: There are 3 types of Machine Learning

I. Supervised Learning
II. Unsupervised Learning
III. Reinforcement Learning

Criteria	**Supervised Learning**	**Unsupervised Learning**	**Reinforcement Learning**
Definition	Learning by using labelled data	Trained by using Unlabeled data without any guidance	Works on Interacting with the environment
Type of Data	Labelled Data	Unlabeled Data	No-Predefined Data
Type of Problems	Regression & Classification	Association & Clustering	Exploitation & Exploration
Supervision	Extra Supervision	No-Supervision	No-Supervision
Algorithms	Linear Regression, Logistic Regression, KNN, SVM, Decision Tree, Random Forest	K-Means, Hierarchical Clustering, Apriori Clustering	Q-Learning, SARSA
Applications	Face Detection, Signature recognition, Forecast Sales	Product Segmentation, Customer Segmentation, Recommendation System	Self-Driving Car, Gamming

Linear Regression

3. Explain How Linear Regression Works

Linear Regression is a Supervised Machine Learning Algorithm to Predict Continuous Variable based on certain features that we have (So we use Linear Regression if the Target Variable is Continuous). If we take a Unidimensional Case for instance, Where Y is dependent on specific value of X. Linear Regression tries to find Best Fit Line, that passes through all the points that we have as closely as possible & the way it does is by trying to minimize the Residuals of the Points from the line.

4. What is the difference between Cost Function & Loss Function

Cost Function: The Cost Function is the Average Error of number of samples in the data.

Loss Function: The Loss Function is the error for individual data point

The Cost Function of Linear Regression is MSE or RMSE

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

5. What is Gradient Descent

Gradient Descent is an iterative optimization algorithm to find the global minimum of a function,

Here that function is our Loss Function

6. What is the Difference Between Gradient Descent, Stochastic Gradient Descent, Mini Batch Gradient Descent?

The Only difference comes while Iterating

- In Gradient Descent we consider all the points in the dataset while calculating loss & Derivative
- In Stochastic Gradient Descent we use single point for calculating loss & Derivative randomly
- In Mini Batch Stochastic Gradient Descent we consider Batch of data points
  Ex:- We takes some sample while calculating loss & Derivative

Formula for finding Loss Function are:-

Gradient Descent

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Mini Batch Stochastic Gradient Descent

$$MSE = \frac{1}{K}\sum_{i=1}^{K}(y_i - \hat{y})^2$$

Stochastic Gradient Descent

$$\mathrm{MSE} = \boxed{\quad}(Y_i - \hat{Y}_i)^2.$$

Always K < n

7. What are the Basic Assumptions of the Linear Regression Algorithm.

There are 5 Most important Assumptions of Linear Regression Algorithm.

1. Linearity:- There should be Linear Relationship between the Features & Target Variable

2. Homoscedasticity:- The Error term has a constant variance

3. Multi-collinearity:- There should be no multicollinearity between the features

4. Independence:- Observations are Independence of each other

5. Normality:- The Error(Residuals) fallows the Normal Distribution

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

Mainly, There are Five Metrics that are commonly used to evaluate regression models:-

I. Mean Absolute Error [MAE]
II. Mean Squared Error [MSE]
III. Root Mean Squared Error [RMSE]
IV. R-Squared (Co-efficient of Determination)
V. Adjusted R-Squared

9. Explain the difference between correlation & Regression:-

Correlation:- It Measures the strength or degree or 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 & describe the cause and effect. It is visualized by a Regression Line.

10. Justify the case where the Linear Regression algorithm is suitable for a given dataset.

Generally a Scatter plot is used to see if Linear Regression is suitable for any given data.

So we can go for a linear model if the relationship looks somewhat linear.

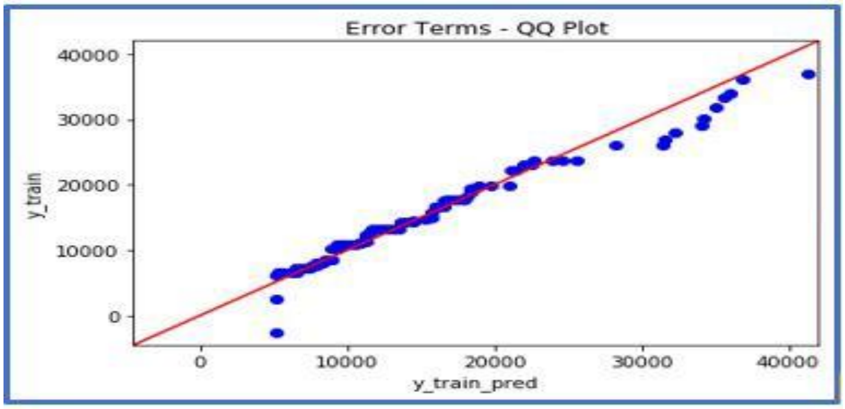
Plotting the scatter plot is easy in case of simple or univariate linear regression, but if we have more than one independent variable ie., In the case of Multivariate linear regression, Then 2 dimensional pairwise scatter plot are used.

11. For a Linear Regression Model, How do we interpret a QQ Plot?

A QQ-Plot is a Graphical Representation of plotting the quantiles of 2-distribution with respect to each other.

In Simple words., We plot quantiles against quantiles in the QQ-Plot with is used to check the normality of error

Whenever we interpret a QQ-Plot, We should concentrate on the Y=X line, Which corresponds to a normal distribution. This is also called as 45 degree line



12. In Linear Regression what is the value of the sum of Residuals for a given Dataset

The Sum of residuals in a linear regression model is 0, Since it assume that the error (residuals) are Normally Distributed with expected value of mean equal to 0

13. What are RMSE & MSE? How to calculate it ?

MSE[Mean Squared Error] :- In simple words, we can say, It is an average of squared difference between actual & predicted value

RMSE[Root Mean Squared Error] :- It is the Square root of the Average squared difference between Actual & Predicted value

Conclusion:-

In general, As the variance of error magnitude increases,

MAE remains steady but RMSE increases.

14. What is OLS?

OLS stands for Ordinary Least Squares. The Main objectives of Linear Regression algorithm is to find co-efficient or estimates by minimizing the error term.

This method finds the best fit line, also known as Regression Line by minimizing the sum of squared difference between the observed & predict value.

15. What are R-Squared & Adjusted R-Squared?

R-Squared also known as the coefficient of determination, It measures the proportion of variation in your dependent variable (Y), Explained by your independent variable (X)

The main problem with R-Squared is that it will always remains the same or increases as we are adding more independent variable, Therefore to overcome this problem,

An Adjusted R-Squared came into picture by penalizing those adding independent variable that do not improve your existing model.

Ridge (L2) & Lasso Regression(L1) [Regularization Technique]

Pre-requisite before going for Ridge & Lasso Regression

I. Better understanding of Linear Regression
II. Bias & Variance

Bias & Variance

We use bias term for training & variance for testing result

If my model performs good in Training dataset & performs badly in the testing dataset, we call this as low bias & high variance [Overfitting condition]

If my model perform worst in both training & testing dataset, which is giving high error we call this as high bias & high variance [Underfitting condition]

If model perform good in both Training & Testing dataset then it is called as Low Bias & Low Variance [Generalized Model]

Note:- Mainly used regularization technique are:-

1) Ridge
2) Lasso

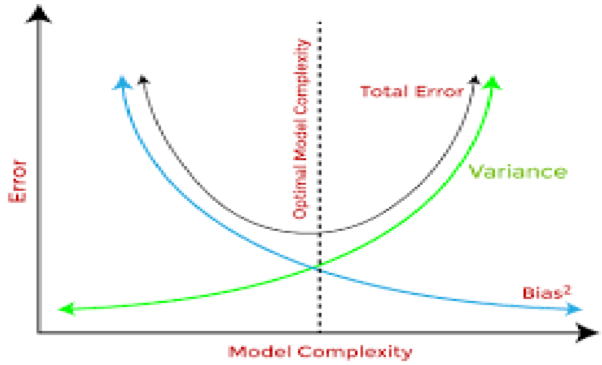
Machine Learning

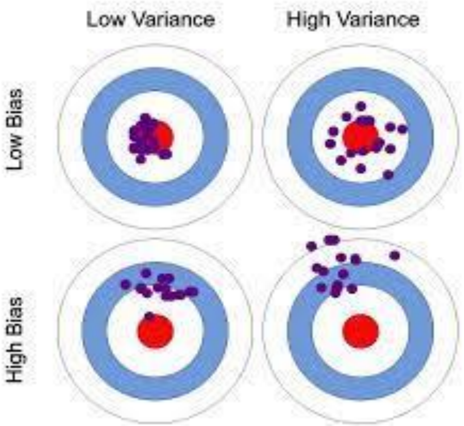
3) Drop out
4) Early Slope

Deep Learning

Machine Learning Interview questions by Jeevan Raj

Explain what is Bias-Variance Tradeoff





In a Statistics & Machine Learning, The Bias-variance Tradeoff is the property of a model.

That the variance of parameter estimated across sample can be reduced by increasing the bias in the estimated parameters.

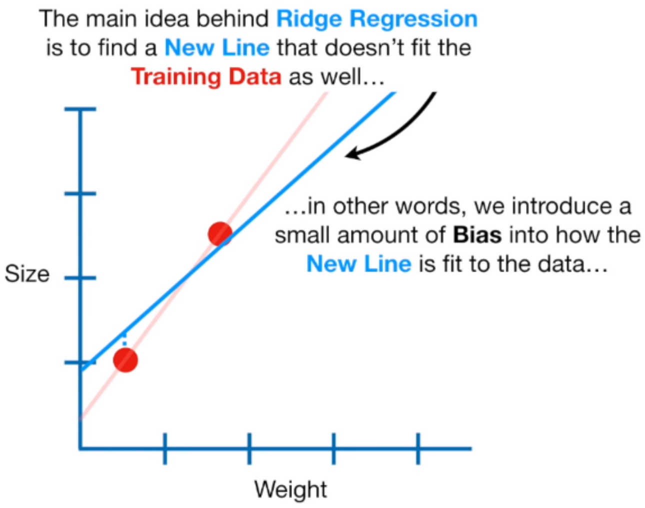
What is Ridge and Lasso Regression?

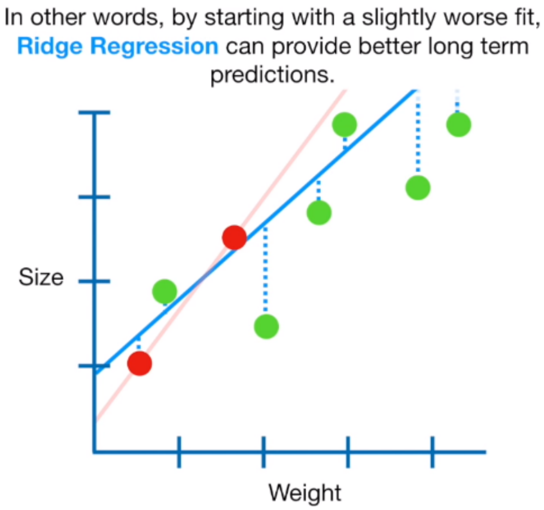
Ridge & Lasso Regression are types of Regularization technique.

Regularization technique are used to deal with overfitting & when the dataset is large. Ridge & Lasso Regression involve adding penalties to the cost function

Machine Learning Interview questions by Jeevan Raj

Ridge & Lasso Regression





Ridge & lasso Regression is a Regularization hyper tuning technique, we use Ridge & Lasso regression if the model is overfitting.

Lets take some example, where I have only 2 data-points to train my model & if I use the best fit line for this data, The best fit line passes through this 2 points & I will get 0 Error in training data & if I check for testing data I will get high error it is also called as overfitting where I have low bias or 0 bias & high variance. To overcome this will use Ridge & lasso Regression so that it adds penalty (λ*Slope^2) for cost function in Ridge Regression & adds penalty (λ*|Slope|) for cost function in Lasso Regression.

Now will understand the cost function for Ridge Regression

You can see the above graph where my best fit line slope has high steep, where 1 unit change in my weight, their will be 2 unit change in my size, but if you now impute this ridge regression line the steep slope will slightly reduce, Its because of we are adding penalty term with residuals

Where λ(lambda) can be in the range of 0 to any positive value

Note:- If I have multiple features our cost function for Ridge Regression is

Ypred=m1x1+m2x2+C

$$\textbf{Cost Function = (yi-ypred)}^{\mathbf{2}} + \underbrace{\lambda\ *[\mathrm{m1}^2+\mathrm{m2}^2]}$$

Note:- The Lambda (λ) is selected by cross validation & will check for which lambda value we are getting

We are using this penalty term to just penalize the steeper slope where it can convert over fitting model to Generalized model.

Lower cost function will consider that has our lambda value

When ever you are increasing the λ(lambda) value, The slope will move very close to zero

Now will understand the Lasso Regression

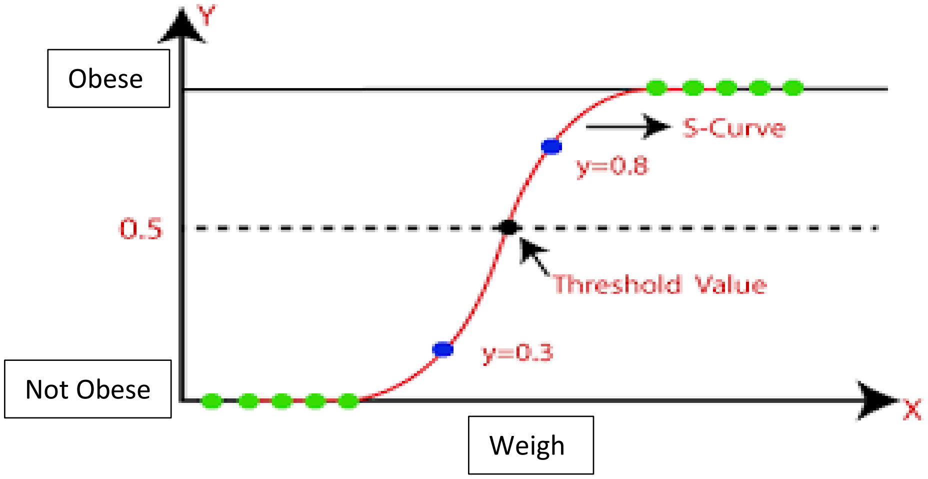
The cost function of lasso regression is **(yi-ypred)^2 +** λ *|Slope| where instead of slope square we use magnitude of slope because the magnitude of slope help us to do the feature selection.

Where if I have multiple feature like y=m1x1+m2x2+m3x3+m4x4+c

My cost function will looks like **(yi-ypred)^2 +** λ *[|m1+m2+m3+m4|]

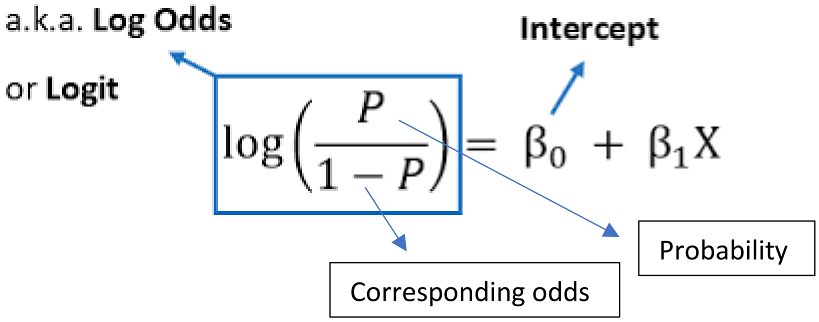
When my slope become ‘0’ then that features will get removed where we can say that those features are not important to predict the model & will keep rest of the feature to predict the target variable

Logistic Regression



Logistic Regression will always give you the probability of happening of the event

If we are using multiclass classification in logistic regression we use OVR[One v/s Rest]



Logistic Regression is used for binary classification where you want to predict the happening, it is used for classification problem statement.

In above problem statement based on weight I am categorizing obese or not obese, If I consider a person weight >=75, I want classify them as obese & whose ever weight is <75, I want to consider them as not obese, So I am going to write an equation when ever my value>=75 it is called obese by using a equation of Best Fit Line Y=mx+c (or) $Y=w^Tx+b$

When the datapoints have outliers the best fit line get deviated & it makes misclassification so to avoid the affect of outliers will use sigmoid. S-Curve & then regression line gives the value >1 & <0 in that situation to skews the line we use sigmoid curve.

16. What are the Metrics used for classification problem statement:-

Machine Learning Interview questions by Jeevan Raj

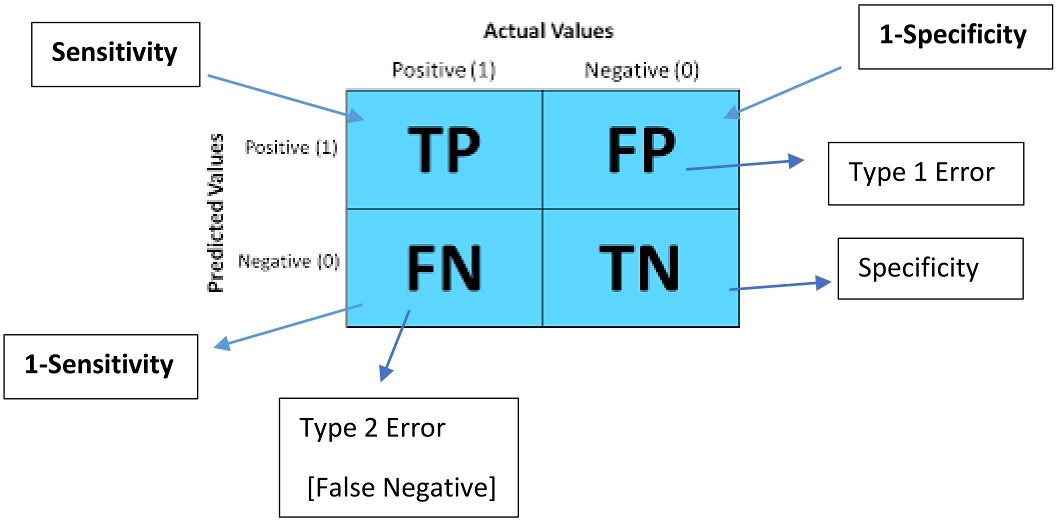
The Metrics used to evaluate how good your model is fitted, Some of metrics used for classification problem statement are:-

1) Confusion Matrix
2) FPR (False Positive Rate) (Type 1 Error)
3) FNR (False Negative Rage) (Type 2 Error)
4) Recall (TPR, Sensitivity)
5) Precession (+ve Prediction Value)
6) Accuracy
7) F Beta Score
8) Cohen kappa
9) ROC Curve, AUC Score [ROC (Receiver operating characteristic curve) & (AUC (Area under Curve)]
10) PR Curve

If we have balanced dataset the metrics we use is Accuracy. But if we have unbalanced dataset we should not use accuracy we should use recall, precision, F-Beta Score

1) **<u>Confusion Matrix :- [Error Matrix]</u>**

In case of Binary classification it is 2*2 Matrix



- <u>If the Dataset is balanced go for Accuracy</u>

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Machine Learning Interview questions by Jeevan Raj

Accuracy = Sum of correctly Predicted by total no of observation

If the dataset is Un Balanced:-

If I have unbalanced dataset ex:- yes = 900 and no = 100 ie., 90:10 ratio , I can't go for accuracy because if my model blindly say all are yes then my accuracy will be 90%. So in that case we should go for Recall, Precision, F-Beta Score

Our main focus is to reduce the Type1 and Type2 error

Precision:- Out of Total Predicted Positive Result how many results were actually positive

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

Ex:- In Spam detection we mainly focus on Precession.

Where the mail is not a spam but your model is predicted as spam, So this is a big affect so we should mainly focus on precision and we want to reduce false positive value

Recall:- Out of total positive actual values how much positive values did we correctly predicted

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

It is also called as True Positive Rate [sensitivity]

Ex:- If a Person is having cancer or not

We mainly focus on recall where the person is having cancer but your model is predicted as you don't have cancer where this may affect a human life so we should mainly focus on recall and we want to try to reduce the false negative.

F-Beta:-

In an imbalanced data both false positive and false negative both are important at that time we should consider both precision and recall at that time will use f-beta score.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

We are going to select optimum value of beta (**β)** based on problem statement

Where., **β = 1 → F1 Score, β = 0.5 → F0.5 Score, β = 2 → F2 Score**

F-1 Score :- The F-1 score is a popular binary classification metric representing a balance between **precision** and **recall**. It is the Harmonic mean of precision and recall.

When to choose β = 1 :- When False Positive and False Negative Both are equally important

$$F1 = 2 \cdot \frac{precison \cdot recall}{precision + \ recall}$$

When to choose β = 0.5 :- When False Positive is more important we reduce the β value

$$F_{0.5} = (1 + 0.5^2) \times \frac{\text{precision} \times \text{recall}}{(0.5^2) \times \text{precision} + \text{recall}}$$

When to choose β = 2 :- When False Negative is more important we increase the β value

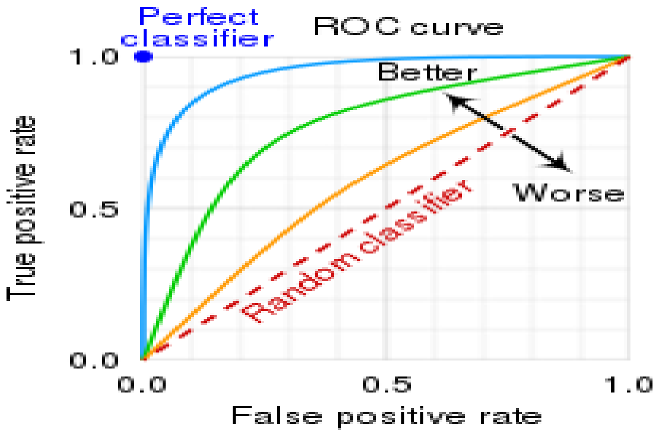
$$F_2 = (1 + 2^2) \times \frac{\text{precision} \times \text{recall}}{(2^2) \times \text{precision} + \text{recall}}$$

17. What is R.O.C & A.U.C Curve

The Receiver Operating Characteristic [R.O.C] is a evaluation matrix for binary classification problem.

It is a probability curve that plots the TPR against FPR at various thresholds values and Essentially separates the signal from the Noise.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



18. What is Cohen’s Kappa Score

Cohen’s kappa coefficient (k) is a statistical measure that is used to measure inter-rater reliability and also intra-rater reliability for qualitative data

1. What do you mean by Logistic Regression.

- Logistic Regression is a Supervised machine learning algorithm that is used when the target variable is of categorical in nature.
- The main objective behind logistic regression is to determine the relationship between feature and the probability of a particular outcome.

Ex:- Where we need to predict whether a student Pass or Fail in an Exam

Given the number of hrs spent for studying as a feature, where the target variable comprises 2-category ie., Pass and Fail

2. What are the different types of Logistic Regression?

Three different types of Logistic Regression are as follows:

- Binary Logistic Regression: In this, the target variable has only two 2 possible outcomes. For Example, 0 and 1, or pass and fail or true and false.
- Multinomial Logistic Regression: In this, the target variable can have three or more possible values without any order.
  For Example, Predicting preference of food i.e. Veg, Non-Veg, Vegan.
- Ordinal Logistic Regression: In this, the target variable can have three or more values with ordering.
  For Example, Movie rating from 1 to 5.

3. What are odds?

Odds are defined as Ratio of the Probability of an Event occurring to the Probability of the event not occurring.

Ex:- Lets assume that the probability of winning a game is 0.06, Then the probability of not winning is 1 – 0.06 = 0.94

4. What factors can attribute to the popular of Logistic Regression

- Logistic Regression is a Popular algorithm as it converts the value of the log of odds which can range from –inf to +inf that ranges between 0 & 1

- Since Logistic function outputs the probability of occurrence of an event. They can be applied to many real life scenarios.

5. What is the impact of outliers in Logistic Regression.

Logistic Regression model are not much impacted due to the presence of outlier, because the sigmoid function tapers the outliers, but the presence of extreme outliers may somehow affects the performance of the model.

6. What is the difference between the output of the logistic model and the logistic function:-

The Logistic Model outputs the logit ie, log of odds.

Where as the logistic function outputs the Probabilities

Logistic model =

$$\overset{\text{Dependent Variable}}{Y_i} = \underbrace{\overset{\text{Population Y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1}\overset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

Logistic Function =

$$y = \frac{1}{1 + e^{-x}}$$

7. What are the Assumptions of Logistic Regression

Some of the Assumptions of Logistic Regression are as fallows:-

- The Logistic Regression which has binary classification ie., 2 classes (It assumes that the target variable is binary, such as yes or no, Pass or Fail.
- There should be linear relationship between the logit of the outcome and each predictor variable [Independent variable]. The logistic function is described as

$$\text{logit}(p) = \log(\frac{p}{1-p})$$

- Independence of observations: Logistic regression assumes that the observations in the dataset are independent of each other
- No multi-collinearity: Logistic regression assumes that the independent variables are not highly correlated with each other
- Large sample size: Logistic regression performs well with moderate to large sample sizes.

8. Can we solve the multiclass classification problems using Logistic Regression? If Yes then How?

- Yes, in order to deal with multiclass classification using Logistic Regression, the most famous method is known as the one-vs-all approach. In this approach, a number of models are trained, which is equal to the number of classes. These models work in a specific way.
- For Example, the first model classifies the datapoint depending on whether it belongs to class 1 or some other class(not class 1); the second model classifies the datapoint into class 2 or some other class(not class 2) and so-on for all other classes.
- So, in this manner, each data point can be checked over all the classes.

9. Why is Logistic Regression termed as Regression and not classification?

The major difference between Regression and classification problem statements is that the target variable in the Regression is numerical (or continuous) whereas in classification it is categorical (or discrete).
Logistic Regression is basically a supervised classification algorithm. However, the Logistic Regression builds a model just like linear regression in order to predict the probability that a given data point belongs to the category numbered as "1" or "0"

10. What are the Advantages and disadvantages of Logistic Regression:-

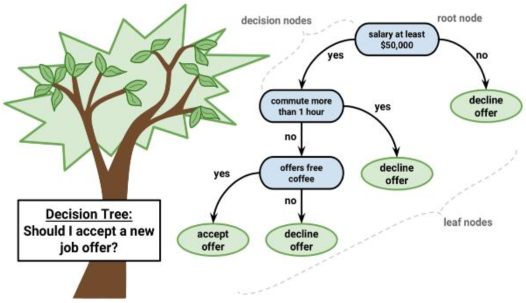
Advantages:-

- It performs well when the dataset is linearly separable.
- They are easier to implement, interpret and very efficient to train.

Disadvantages:-

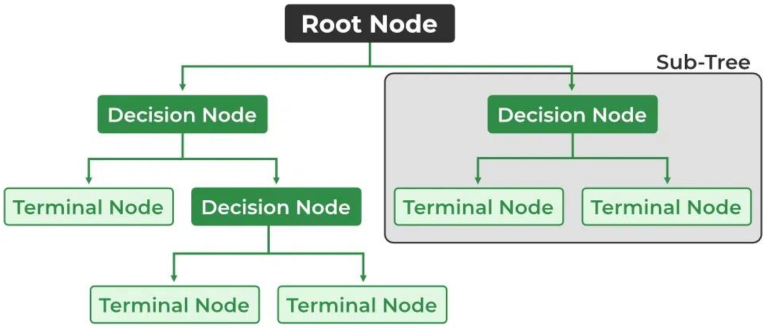
- It is quite sensitive to noise and overfitting
- Sometimes lot of feature engineering is required

Decision Tree



A Decision Tree is a supervised Machine learning algorithms used for both regression and classification problem statement.

A decision tree is a flowchart-like tree structure (flow control statement) it is built based on Top-Down approach where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. The final output of the decision tree is a Tree having decision nodes and leaf nodes.



Decision Tree Terminologies

Root Node: It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.

Decision/Internal Node: One or more decision nodes that result in the splitting of data into multiple data segments and our main goal is to have the children nodes with maximum homogeneity or purity.

Leaf/Terminal Node: This node represent the data section having the highest homogeneity.

Splitting: The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.

Impurity: A measurement of the target variable’s homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The Gini index and entropy are two commonly used impurity measurements in decision trees for classifications task

Information Gain: Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain, It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets

Pruning: The process of removing branches from the tree that do not provide any additional information or lead to overfitting.

1. **What is Decision Tree? (above page you can find the answer)**

2. **List down some popular algorithm used for deriving decision tree along with their attribute selection measure.**

   Some of the popular algorithms used to constructing trees are:-

   1. **ID3 [Iterative Dichotomizer] :-** It is a core algorithm for building a decision tree, it uses information gain as an attribute for selection measure to construct a classification decision tree, In case of regression it uses Standard deviation reduction.
   2. **C 4.5 (Successor of ID3) :-** It uses gain ratio as attribute for selection measure
   3. **CART [Classification And Regression Tree] :-** Uses Gini index as an attribute for selection measure

3. **Explain CART Algorithm for decision tree:-**

CART [Classification And Regression Tree] algorithm is used for building a decision tree based on Gini's impurity index as splitting criterion.

CART is a binary tree, it is built by splitting node into two child nodes.

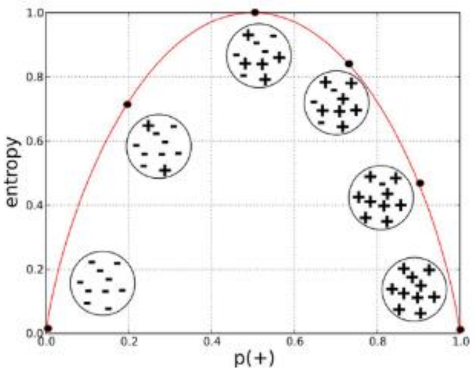
CART algorithm used to find the independent variable that creates the best homogeneous [same type] group when splitting the data.

4. **What is Entropy:-**

Entropy is used for checking the impurity or uncertainty present in the data.

The entropy is 0 when the dataset is completely homogeneous, meaning that each instance belongs to the same class. It is the lowest entropy indicating no uncertainty in the dataset sample.

When the dataset is equally divided between multiple classes, the entropy is at its maximum value. Therefore, entropy is highest when the distribution of class labels is even, indicating maximum uncertainty in the dataset sample.



$$E = -\sum_{i=1}^{n} p_i \, log_2(p_i)$$

Entropy is used to evaluate the quality of a split. When the Entropy is zero then the sample is completely homogenous and Entropy is One when the sample is equally divided between different classes.

5. **What is Information Gain:-**

Information Gain indicates how much information a particular feature/ variable give us about the final outcome.

It measures the reduction in Entropy before and after split on a subset, The more the information gain the better the model,

$$\text{Information Gain}(T,X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

Or

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{K} Entropy(j,\ after)$$

6. **Explain the difference between ID3 and CART Algorithm:-**

ID3 Algorithm	CART Algorithm
Splitting Criterion:- ID3 uses information gain as the splitting criterion, which measures the reduction in entropy (or increase in purity) achieved by splitting on a particular feature.	CART uses Gini impurity as the splitting criterion, which quantifies the probability of misclassifying a randomly chosen element in a subset if it were randomly labeled according to the class distribution in that subset.
Handling continuous variables:- ID3 is designed to handle categorical features and cannot directly handle continuous variables.	CART can naturally handle continuous variables by evaluating different splitting points based on the Gini impurity.
Tree structure: ID3 constructs a decision tree that can include multiple branches at each internal node, allowing for a more expressive tree structure	CART constructs binary decision trees, where each internal node has exactly two branches based on the binary splitting condition.
Output: ID3 generates decision trees that are primarily used for classification tasks	CART is more versatile and can be used for both classification and regression tasks.
Handling missing values: ID3 does not handle missing values in the dataset and typically requires preprocessing or imputation techniques to address them.	CART can handle missing values by considering alternative splits that separate the missing values into a separate branch.

Pruning: ID3 does not incorporate explicit pruning techniques, which can lead to overfitting.	CART includes a pruning mechanism, where the tree is initially grown to its full depth and then pruned back based on a validation set performance. This process helps to reduce overfitting and improve the generalization ability of the tree.

7. Which should be preferred among Gini impurity and Entropy?

Most of the time it does not make a big difference, they leads to almost similar.

- **Interpretability:** Entropy has a more intuitive interpretation in terms of information theory. It measures the level of impurity or unpredictability in a dataset. Gini impurity, on the other hand, measures the probability of misclassifying a randomly chosen element. If interpretability is a priority, entropy may be preferred.

- **Computational efficiency:** Gini impurity is generally faster to compute than entropy since it involves fewer calculations, as it does not require logarithmic operations. If computational efficiency is crucial, Gini impurity may be preferred, especially for large datasets.

- **Sensitivity to imbalanced classes:** Gini impurity tends to be slightly more sensitive to imbalanced class distributions compared to entropy. Gini impurity tends to favor splits that result in larger subsets, potentially leading to biased trees when dealing with imbalanced classes. In such cases, entropy may be more suitable as it considers the information gain for each class independently.

8. List down the different types of Nodes in decision Tree:-

- **Root Node:-** It is the Top-most node of the tree from where the tree starts.
- **Decision Node:**- One or more Decision nodes that result in the splitting of data into multiple data segments and our main goal is to have the children nodes with maximum homogeneity or purity.
- **Leaf Nodes:**- These Node represent the data section having the highest homogeneity

9. Do we require feature scaling for decision tree? Explain

Decision Tree are mainly intuitive, easy to interpret as well as require less data preparation.

The decision tree doesn’t require feature scaling or centering (standardization) at all. Such models are often called as **White – box Models.**

Decision Tree provides simple classification rule based on If & else statement.

10. Explain the steps in making a decision tree.

a) Take the Entire dataset as an input.
b) Calculate the Entropy of the target variable, As well as the predictor attributes
c) Calculate the information gain of all attributes.
d) Choose the attribute with the highest information gain as the Root Node
e) Repeat the same procedure on every branch until the decision node of each branch is finalized.