Q1. Explain the Decision Tree algorithm in detail.?

Ans.

1. Decision Tree algorithm is a supervised machine learning algorithm used for both classifications as well as regression. Preferred for classification problems.

2. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

Decision trees consist of three different elements:

1. **Root Node:**

The top-level node represents the ultimate objective or big decision you’re trying to make.

1. **Branches:**

Branches, which stem from the root, represent different options—or courses of action—that are available when making a particular decision.

1. **Leaf Node/ Pure Node:**

The leaf nodes—which are attached at the end of the branches—represent possible outcomes for each action.

Assumptions for Decision Tree Algorithm:

1. In the beginning, we consider the whole dataset as root.
2. Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model
3. On the basis of attribute values, records are distributed recursively.
4. We use statistical methods for ordering attributes as root or the internal node.

In Decision Tree, the major challenge is to the identification of the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

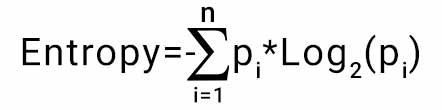
1. Information Gain :
2. Gini Index

Information Gain: Information gain is a measure of this change in entropy.

Entropy :it is the measurement of the impurity or randomness in the data points.

1. **Entropy is calculated between 0 and 1**

Entropy Formula :



Information gain values also lies between 0 and 1

Information Gain Formula:

**Information Gain = E(S) - sum [ weighted avg \* entropy of each attribute]**

Where

E(S) = Entropy of Target feature

Weighted avg = no of attributes / total sample of feature

Information gain is used to decide which feature to split on at each step in building the tree.

Attribute having highest Information Gain (IG) value will select for root node, and it will spilt the data points.

Gini Index/ Gini Impurity:

Gini Index or gini impurity is also a measure of impurity like entropy.

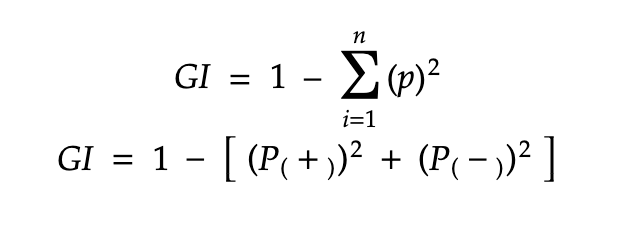
Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.

It means an attribute with lower Gini index should be preferred.

Gini index is calaculated between 0 to 0.5

Classification and Regression Tree [(CART) algorithm](https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/) deploys the method of the Gini Index to originate binary splits.

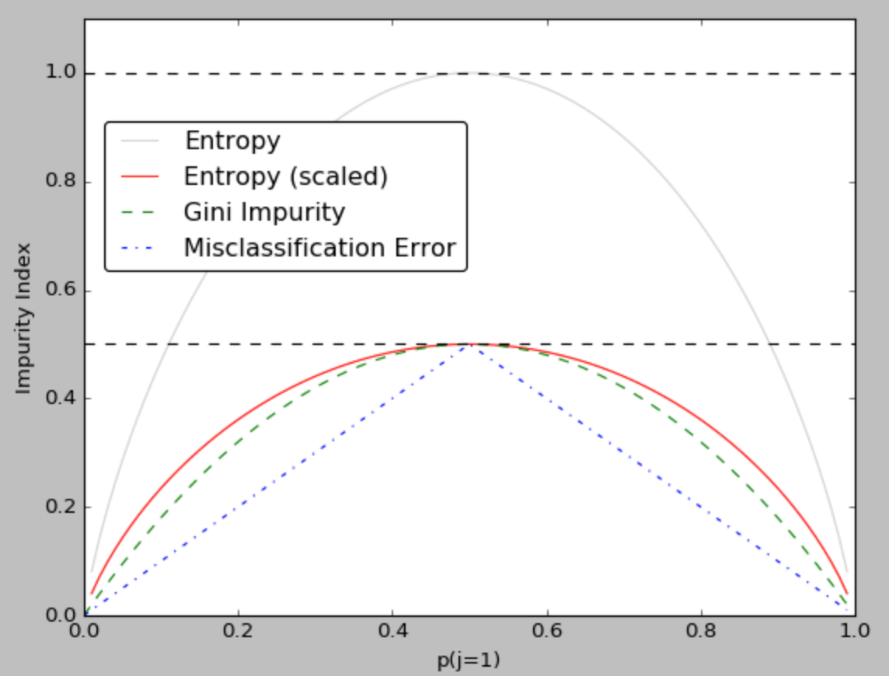
Gini Index Formula:



Gini vs Entropy:

Gini Index has values inside the interval [0, 0.5] whereas the interval of the Entropy is [0, 1]

Computationally, entropy is more complex since it makes use of logarithms and consequently, the calculation of the Gini Index will be faster.



Decision Tree algorithms

1. ID3 [Iterative Dichotomiser 3] : Use of Information gain & Entropy

This algorithm uses Information Gain to decide which attribute is to be used classify the current subset of the data. For each level of the tree, information gain is calculated for the remaining data recursively.

1. CART [Classification and Regression Tree]: Use of Gini index/ Gini impurity Or Entropy

It is a dynamic learning algorithm which can produce a regression tree as well as a classification tree depending upon the dependent variable.

Q2. What are the steps for making a decision tree?

Ans.

**Decision Trees (DTs)** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Steps for making decision Tree:

1. Select Algorithm : There are several algorithms are available to build decision tree. Mostly ID3 & CART is majorlyused.
2. If ID3 is selected for decision three
   1. Information Gain is calculated for every features with help of entropy.

Entropy is measureof impurity,

Entropy fromula

**E(S) = - sum[P(Y) \* log2 P(Y)]**

Information Gain formula:

**IG = E(S) – sum[ weighted avg \* E(S) of every attribute]**

* 1. Information gain values is between 0 to 1.
  2. Highest information gain selected for root node to be spilt.
  3. This precoess is repated recursively util we find the pure node.

1. CART (Classification and Regression Tree) :

CART is a DT algorithm thatroduces **binary** Classification or Regression Trees, depending on whether the dependent (or target) variable is categorical or numeric.

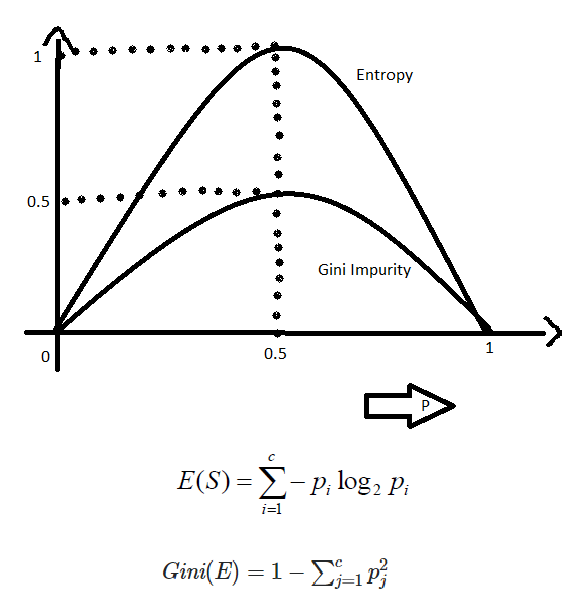
* 1. Gini Index or Gini ipurity is calcuated for evary feature

Gini index is the metric used in the CART algorithm to measure impurity.

Gini Index Formula:

**Gini Index = 1 – [P(+)2 + P(-)2]**

* 1. Gini Impurity range is between 0 to 0.5
  2. Feature with lowest score selected for root node.
  3. Again Gini index shall be calculated for reamaining features & tree will cretaed until we find the pure node.



The above graph shows the Gini index and entropy with respect to the probabilities of each feature.

Q3. What are the Algorithms used in the Decision Tree?

Ans.

There are several algorithms to build a decision tree.

1. CART-Classification and Regression Trees

Impurity Matrix = Gini Index/ Gini Impurity

1. ID3-Iterative Dichotomiser 3

Impurity Matrix = Information Gain (Using Entropy)

1. C4.5
2. CHAID-Chi-squared Automatic Interaction Detection

Majorly CART & ID3 algorithms are used.