Decision Trees

Aaryan

CO21BTECH11001

Let number of features of dataset = n

Let number of sets of features = m

Data consists of matrices X and y where i^{th} column of X represents the i^{th} feature of dataset and i^{th} element of y represents the value of variable dependent on set of features listed in i^{th} row of X.

Decision tree is a kind of classification as well as regression algorithm, which create yes/no questions and continually split the dataset until you isolate all data points belonging to each class.

The algorithm tries to completely separate the dataset such that all leaf nodes, i.e., the nodes that don't split the data further, belong to a single class.

To pick the best split, we have to introduce a loss function corresponding to a node known as Gini Impurity.

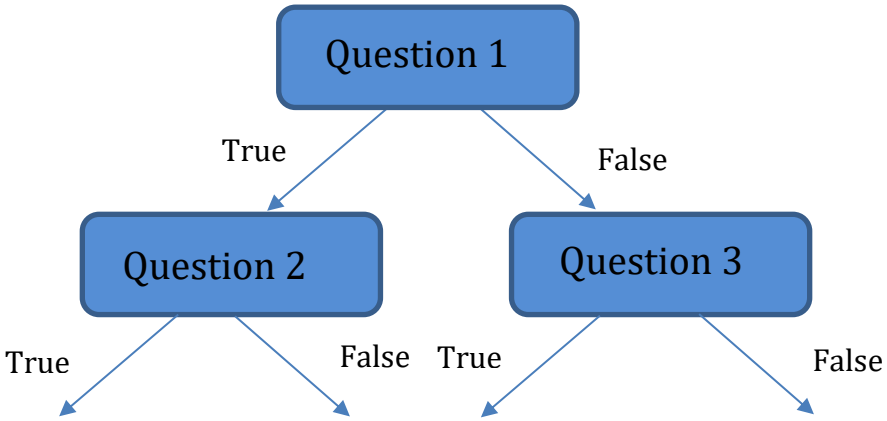
$$Gini(node) = \sum_{c} p_c(1 - p_c)$$

where p_c is the probability of picking a data point from class *c.*

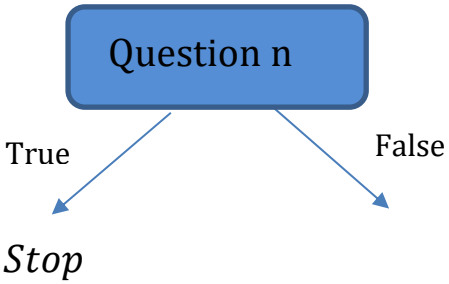
$$p_c = \frac{number\ of\ observations\ with\ class\ c}{total\ observations\ in\ node}$$

Objective – Pick the node with minimum Gini impurity for further splitting.

The procedure looks as follows –



and so on… until it reaches a *leaf* where no further splitting can be done i.e. Gini impurity of node becomes 0.



We can now have a test case which we can feed in the tree and after some n number of questions, it will make a decision.

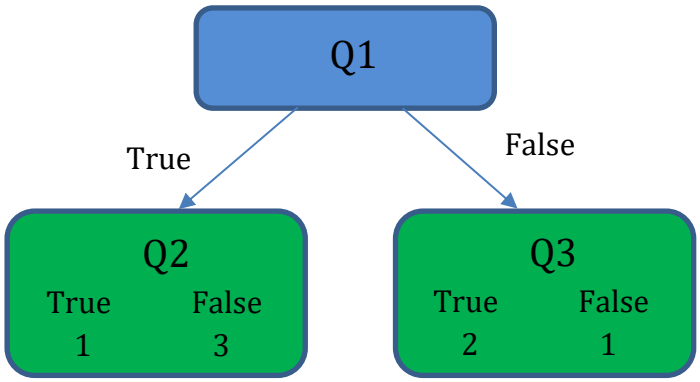
Questions –

1. What is the type of Decision Tree algorithm?

   Ans. It is a supervised machine learning algorithm that can be used for both Regression and Classification problems.

2. What are the advantages of Decision Tree algorithm?
   Ans. It requires less data cleaning, we can use multiple data types in this tree.
3. What is the meaning of a “Pure Node”?
   Ans. It means the Gini Impurity of the node is 0.

4. What is the Gini Impurity Q1 –



Ans. For Q2, $Gini(Q2) = \frac{1}{1+3}\left(1 - \frac{1}{1+3}\right) = \frac{3}{16}$

For Q3, $Gini(Q3) = \frac{1}{1+2}\left(1 - \frac{1}{1+2}\right) = \frac{2}{9}$

$Gini(Q1)$ =Weighted average of Gini Impurity of Q2 and Q3

$$= \frac{4}{4+3} \times \frac{3}{16} + \frac{3}{4+3} \times \frac{2}{9} = \frac{17}{84} = 0.2024$$

5. What are disadvantages of using Decision Trees?

Ans. It is prone to Overfitting .