

ASSIGNMENT – 07

TOPIC: DECISION TREE

Q.1. Write notes on Decision Tree?

Ans. Decision tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables.

Types of Decision Trees

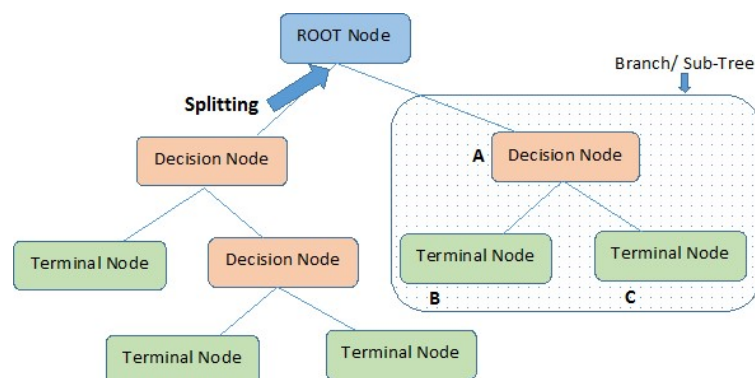
Types of decision tree is based on the type of target variable we have. It can be of two types:

1. **Categorical Variable Decision Tree:** Decision Tree which has categorical target variable then it called as categorical variable decision tree. Example: - In above scenario of student problem, where the target variable was “Student will play cricket or not” i.e. YES or NO.
2. **Continuous Variable Decision Tree:** Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Important Terminology related to Tree based Algorithms

Let’s look at the basic terminology used with Decision trees:

1. **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.

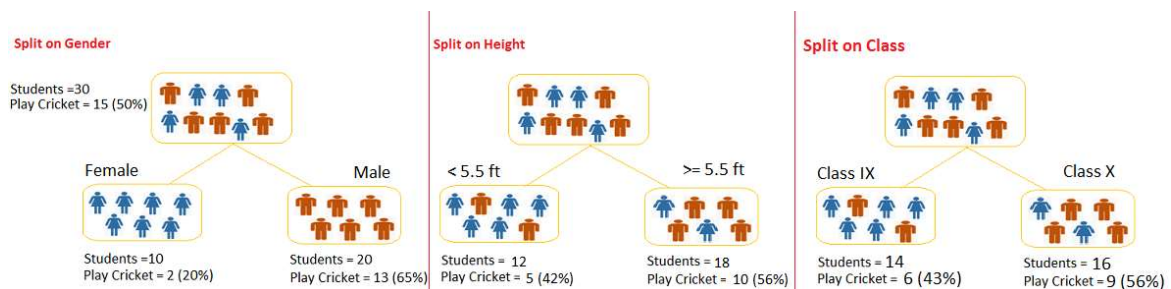


Note:- A is parent node of B and C.

Example:-

Let's say we have a sample of 30 students with three variables Gender (Boy/ Girl), Class (IX/ X) and Height (5 to 6 ft.). 15 out of these 30 play cricket in leisure time. Now, I want to create a model to predict who will play cricket during leisure period? In this problem, we need to segregate students who play cricket in their leisure time based on highly significant input variable among all three.

This is where decision tree helps, it will segregate the students based on all values of three variable and identify the variable, which creates the best homogeneous sets of students (which are heterogeneous to each other). In the snapshot below, you can see that variable Gender is able to identify best homogeneous sets compared to the other two variables.



As mentioned above, decision tree identifies the most significant variable and its value that gives best homogeneous sets of population. Now the question which arises is, how does it identify the variable and the split? To do this, decision tree uses various algorithms, which we will discuss in the following section.

How does a tree based algorithms decide where to split?

The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria is different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

Gini

Gini says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

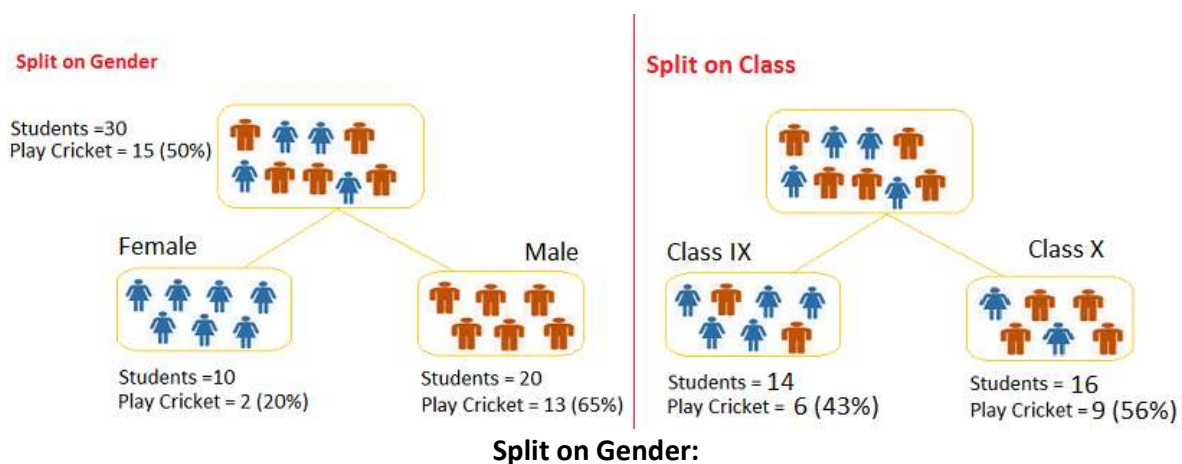
1. It works with categorical target variable "Success" or "Failure".
2. It performs only Binary splits

3. Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses Gini method to create binary splits.

Steps to Calculate Gini for a split

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure ($p^2 + q^2$).
2. Calculate Gini for split using weighted Gini score of each node of that split

Example: – Referring to example used above, where we want to segregate the students based on target variable (playing cricket or not). In the snapshot below, we split the population using two input variables Gender and Class. Now, I want to identify which split is producing more homogeneous sub-nodes using Gini.



1. Calculate, Gini for sub-node Female = $(0.2) * (0.2) + (0.8) * (0.8) = 0.68$
2. Gini for sub-node Male = $(0.65) * (0.65) + (0.35) * (0.35) = 0.55$
3. Calculate weighted Gini for Split Gender = $(10/30) * 0.68 + (20/30) * 0.55 = 0.59$

Similar for Split on Class:

1. Gini for sub-node Class IX = $(0.43) * (0.43) + (0.57) * (0.57) = 0.51$
2. Gini for sub-node Class X = $(0.56) * (0.56) + (0.44) * (0.44) = 0.51$
3. Calculate weighted Gini for Split Class = $(14/30) * 0.51 + (16/30) * 0.51 = 0.51$

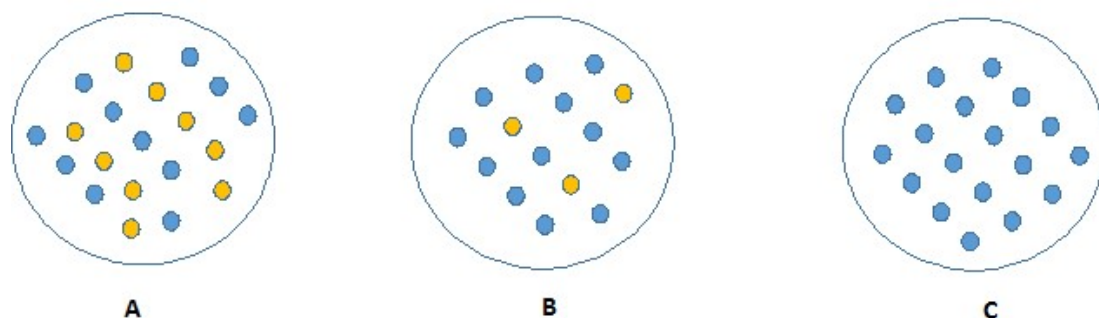
Above, you can see that Gini score for *Split on Gender* is higher than *Split on Class*, hence, the node split will take place on Gender.

You might often come across the term 'Gini Impurity' which is determined by subtracting the Gini value from 1. So mathematically we can say,

$$\text{Gini Impurity} = 1 - \text{Gini}$$

Information Gain:

Look at the image below and think which node can be described easily. I am sure, your answer is C because it requires less information as all values are similar. On the other hand, B requires more information to describe it and A requires the maximum information. In other words, we can say that C is a pure node, B is less Impure and A is more impure.



Now, we can build a conclusion that less impure node requires less information to describe it. And, more impure node requires more information. Information theory is a measure to define this degree of disorganization in a system known as Entropy. If the sample is completely homogeneous, then the entropy is zero and if the sample is an equally divided (50% – 50%), it has entropy of one.

Entropy can be calculated using formula:- $\text{Entropy} = -p \log_2 p - q \log_2 q$

Here p and q is probability of success and failure respectively in that node. Entropy is also used with categorical target variable. It chooses the split which has lowest entropy compared to parent node and other splits. The lesser the entropy, the better it is.

Steps to calculate entropy for a split:

1. Calculate entropy of parent node
2. Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.

Example: Let's use this method to identify best split for student example.

1. Entropy for parent node = $-(15/30) \log_2 (15/30) - (15/30) \log_2 (15/30) = 1$. Here 1 shows that it is an impure node.
2. Entropy for Female node = $-(2/10) \log_2 (2/10) - (8/10) \log_2 (8/10) = 0.72$ and for male node, $-(13/20) \log_2 (13/20) - (7/20) \log_2 (7/20) = 0.93$
3. Entropy for split Gender = Weighted entropy of sub-nodes = $(10/30)*0.72 + (20/30)*0.93 = 0.86$
4. Entropy for Class IX node, $-(6/14) \log_2 (6/14) - (8/14) \log_2 (8/14) = 0.99$ and for Class X node, $-(9/16) \log_2 (9/16) - (7/16) \log_2 (7/16) = 0.99$.
5. Entropy for split Class = $(14/30)*0.99 + (16/30)*0.99 = 0.99$

Above, you can see that entropy for *Split on Gender* is the lowest among all, so the tree will split on *Gender*. We can derive information gain from entropy as **1- Entropy**.

Reduction in Variance

Till now, we have discussed the algorithms for categorical target variable. Reduction in variance is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population:

$$\text{Variance} = \frac{\sum(X - \bar{X})^2}{n}$$

Above X-bar is mean of the values, X is actual and n is number of values.

Steps to calculate Variance:

1. Calculate variance for each node.
2. Calculate variance for each split as weighted average of each node variance.

Example: - Let's assign numerical value 1 for play cricket and 0 for not playing cricket. Now follow the steps to identify the right split:

1. Variance for Root node, here mean value is $(15*1 + 15*0)/30 = 0.5$ and we have 15 one and 15 zero. Now variance would be $((1-0.5)^2 + (1-0.5)^2 + \dots 15 \text{ times} + (0-0.5)^2 + (0-0.5)^2 + \dots 15 \text{ times}) / 30$, this can be written as $(15*(1-0.5)^2 + 15*(0-0.5)^2) / 30 = \mathbf{0.25}$
2. Mean of Female node = $(2*1 + 8*0)/10 = 0.2$ and Variance = $(2*(1-0.2)^2 + 8*(0-0.2)^2) / 10 = 0.16$
3. Mean of Male Node = $(13*1 + 7*0)/20 = 0.65$ and Variance = $(13*(1-0.65)^2 + 7*(0-0.65)^2) / 20 = 0.23$
4. Variance for Split Gender = Weighted Variance of Sub-nodes = $(10/30)*0.16 + (20/30)*0.23 = \mathbf{0.21}$
5. Mean of Class IX node = $(6*1 + 8*0)/14 = 0.43$ and Variance = $(6*(1-0.43)^2 + 8*(0-0.43)^2) / 14 = 0.24$
6. Mean of Class X node = $(9*1 + 7*0)/16 = 0.56$ and Variance = $(9*(1-0.56)^2 + 7*(0-0.56)^2) / 16 = 0.25$
7. Variance for Split Gender = $(14/30)*0.24 + (16/30)*0.25 = \mathbf{0.25}$

Above, you can see that Gender split has lower variance compare to parent node, so the split would take place on *Gender* variable.