**DECISION TREE CLASSIFIER**

1. A Decision Tree algorithm is one of the most popular machine learning algorithms. It uses a tree like structure and their possible combinations to solve a particular problem.
2. It belongs to the class of supervised learning algorithms where it can be used for both classification and regression purposes.
3. A decision tree is a structure that includes a root node, branches, and leaf nodes.
4. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

We make some assumptions while implementing the Decision-Tree algorithm. These are listed below:-

1. At the beginning, the whole training set is considered as the root.
2. Feature values need to be categorical. If the values are continuous then they are discretized prior to building the model.
3. Records are distributed recursively on the basis of attribute values.
4. Order to placing attributes as root or internal node of the tree is done by using some statistical approach.
5. **A root node** that has no incoming edges and zero or more outgoing edges.
6. **Internal nodes**, each of which has exactly one incoming edge and two or more outgoing edges.
7. **Leaf or terminal nodes**, each of which has exactly one incoming edge and no outgoing edges.

**Decision Tree algorithm intuition**[**¶**](https://www.kaggle.com/code/prashant111/decision-tree-classifier-tutorial/log#4.-Decision-Tree-algorithm-intuition-)

The Decision-Tree algorithm is one of the most frequently and widely used supervised machine learning algorithms that can be used for both classification and regression tasks. The intuition behind the Decision-Tree algorithm is very simple to understand.

The Decision Tree algorithm intuition is as follows:-

1. For each attribute in the dataset, the Decision-Tree algorithm forms a node. The most important attribute is placed at the root node.
2. For evaluating the task in hand, we start at the root node and we work our way down the tree by following the corresponding node that meets our condition or decision.
3. This process continues until a leaf node is reached. It contains the prediction or the outcome of the Decision Tree.

**Attribute selection measures**

* The primary challenge in the Decision Tree implementation is to identify the attributes which we consider as the root node and each level. This process is known as the attributes selection.
* There are different attributes selection measure to identify the attribute which can be considered as the root node at each level.
* There are 2 popular attribute selection measures. They are as follows:-

1. Information gain
2. Gini index

While using Information gain as a criterion, we assume attributes to be categorical and for Gini index attributes are assumed to be continuous. These attribute selection measures are described below.

**1. Information gain**

By using information gain as a criterion, we try to estimate the information contained by each attribute. To understand the concept of Information Gain, we need to know another concept called Entropy.

*IG*(*D*,*A*)=*H*(*D*)−*H*(*D*∣*A*)

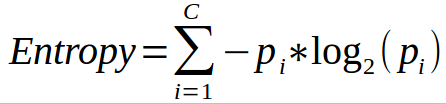
Where,

* IG(D, A) is the Information Gain of feature A concerning dataset D.
* H(D) is the entropy of dataset D.
* H(D∣A) is the conditional entropy of dataset D given feature A.

**1. Entropy H(D)**

* Entropy measures the impurity in the given dataset.
* In Physics and Mathematics, entropy is referred to as the randomness or uncertainty of a random variable X. In information theory, it refers to the impurity in a group of examples.
* Information gain is the decrease in entropy. Information gain computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values.

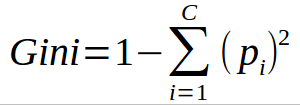
Entropy is represented by the following formula:-



* Here, c is the number of classes and pi is the probability associated with the ith class.
* The ID3 (Iterative Dichotomiser) Decision Tree algorithm uses entropy to calculate information gain.
* So, by calculating decrease in entropy measure of each attribute we can calculate their information gain.
* The attribute with the highest information gain is chosen as the splitting attribute at the node.

**5.2 Gini index**

* Another attribute selection measure that CART (Categorical and Regression Trees) uses is the Gini index. It uses the Gini method to create split points.
* Gini index can be represented with the following diagram:-
* Gini index



* Here, again c is the number of classes and pi is the probability associated with the ith class.
* Gini index says, if we randomly select two items from a population, they must be of the same class and probability for this is 1 if the population is pure.
* It works with the categorical target variable “Success” or “Failure”. It performs only binary splits. The higher the value of Gini, higher the homogeneity.
* CART (Classification and Regression Tree) uses the Gini method to create binary splits.

Steps to Calculate Gini for a split

1. Calculate Gini for sub-nodes, using formula sum of the 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.

In case of a discrete-valued attribute, the subset that gives the minimum gini index for that chosen is selected as a splitting attribute. In the case of continuous-valued attributes, the strategy is to select each pair of adjacent values as a possible split-point and point with smaller gini index chosen as the splitting point. The attribute with minimum Gini index is chosen as the splitting attribute.

A questionnaire with text on it

AI-generated content may be incorrect.

*Let us solve the example using Decision Tree ID3 Algorithm*

*Step 1: Calculate Information Gain for each attribute*

*1.1 Calculate IG for Outlook Attribute*

*A screenshot of a math formula

AI-generated content may be incorrect.*

*1.2. Calculate IG for Temperature Attribute*

*A screenshot of a graph

AI-generated content may be incorrect.*

*1.3. Calculate IG of Wind Attribute*

*A screenshot of a math program

AI-generated content may be incorrect.*

*A table with text and numbers

AI-generated content may be incorrect.*

*By observing IG of all Attribute ,It is high for Outlook attribute ,so we will take this attribute as Root Node. The the Tree at level 0 will looks like the below figure.*

*A diagram of a flowchart

AI-generated content may be incorrect.*

*Level 2: After selection Outlook which attribute can be select as Decision node. For that again calculate the IG for Tempertaure, Humidity and wind.*

*2.1 Calculate IG for Temperature in Sunny*

*A screenshot of a math program

AI-generated content may be incorrect.*

*2.2 Calculate IG for Humidity in Sunny*

*A screenshot of a math program

AI-generated content may be incorrect.*

*2.3 Calculate IG for Wind in Sunny*

*A screenshot of a math program

AI-generated content may be incorrect.*

*A math equations with numbers

AI-generated content may be incorrect.*

*At level 2 , Humidity Attribute is select as node. Then the tree becomes*

*A diagram of a algorithm

AI-generated content may be incorrect.*

*Level 3:After selection Humidity the two leaf nodes have pure information, we have to find which attribute can be select as Decision node. For that again calculate the IG for Temperature ,wind when it is Rainy.*

*3.1 calculate IG for Temperature when it is Rainy*

*A screenshot of a math program

AI-generated content may be incorrect.*

*3.2 calculate IG for Humidity when it is Rainy*

*A screenshot of a math program

AI-generated content may be incorrect.*

*3.2 calculate IG for Windy when it is Rainy*

*A screenshot of a math program

AI-generated content may be incorrect.*

*A screenshot of a computer

AI-generated content may be incorrect.*

*We will take windy as our decision node at this level 3.Then tree will be*

*A diagram of a diagram

AI-generated content may be incorrect.*

*References:*

<https://github.com/milaan9/Python_Decision_Tree_and_Random_Forest/blob/main/001_Decision_Tree_PlayGolf_ID3.ipynb>

<https://medium.datadriveninvestor.com/decision-tree-algorithm-with-hands-on-example-e6c2afb40d38>