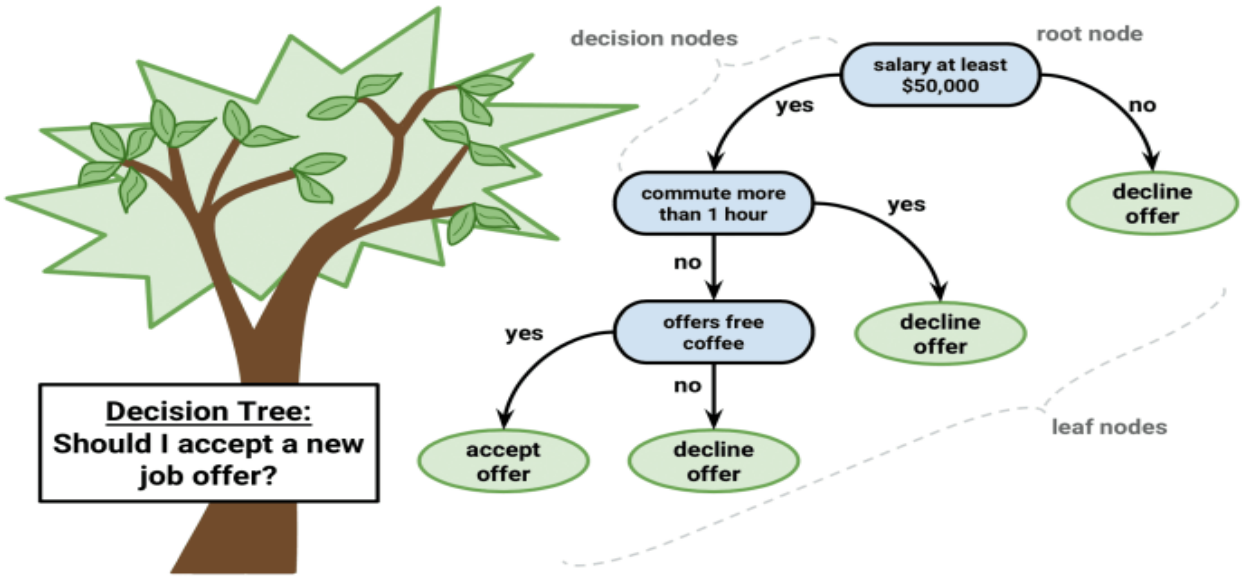
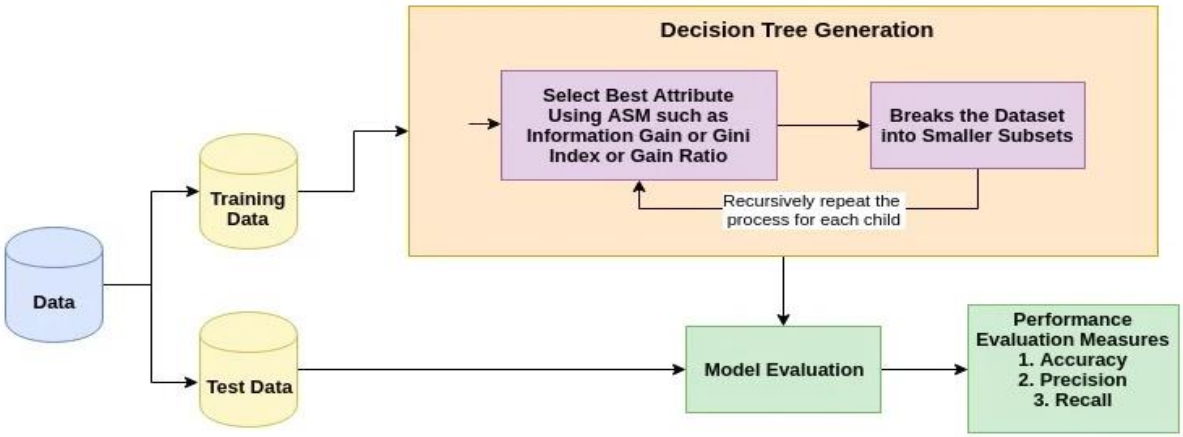
National University of Computer and Emerging Sciences CL 461 Artificial Intelligence

Lab Manual 10

INTRODUCTION TO DECISION TREES



- Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
- 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.

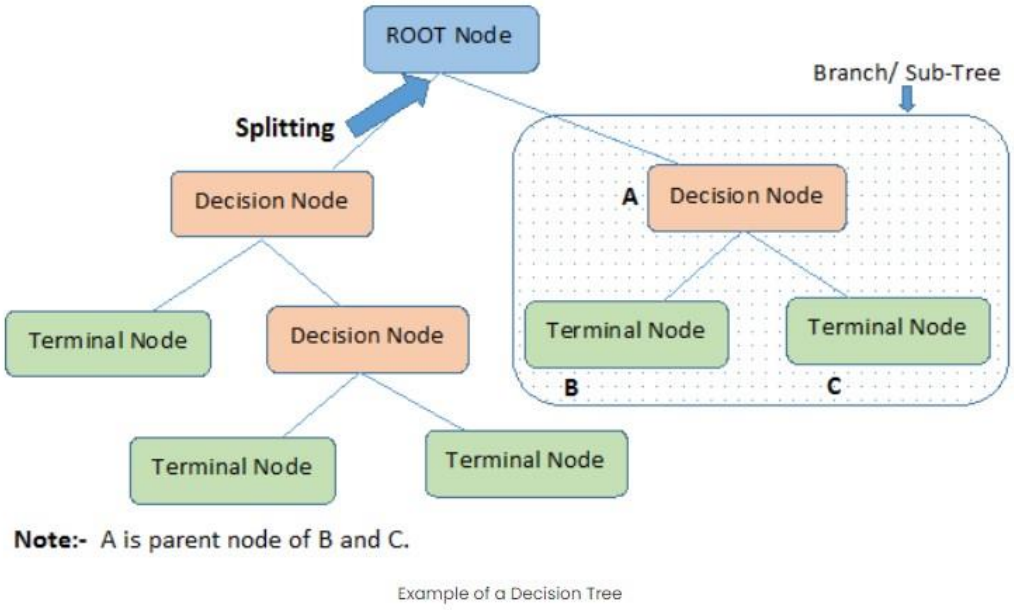


In Decision Tree the major challenge is to identify the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

1. Information Gain (ID3)
2. Gini Index (CART)

Key Terminology

Below is an image explaining the basic structure of the decision tree. Every tree has a **root node**, where the inputs are passed through. This root node is further divided into sets of decision nodes where results and observations are conditionally based. The process of dividing a single node into multiple nodes is called **splitting**. If a node doesn't split into further nodes, then it's called a **leaf node**, or **terminal node**. A subsection of a decision tree is called a **branch** or **subtree** (e.g. in the box in the image below).

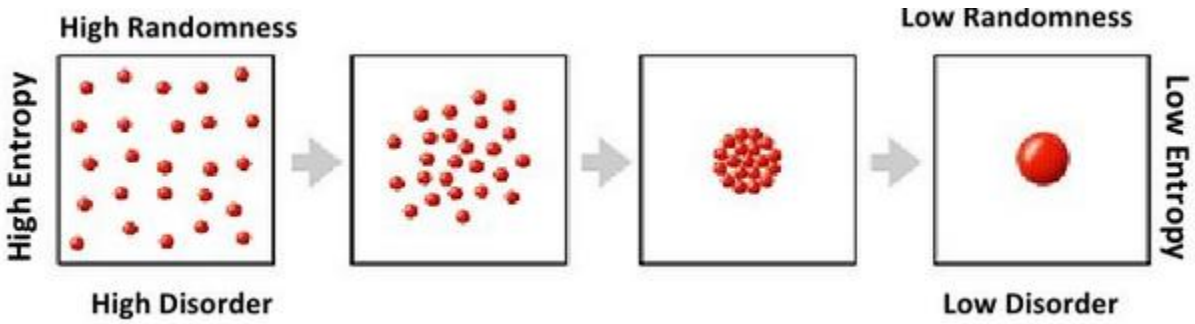


Example of a Decision Tree

Let's get started with the commonly used techniques to split, and thereby, construct the Decision tree.

Information Gain

Information Gain depicts the amount of information that is gained by an attribute. It tells us how important the attribute is. Since Decision Tree construction is all about finding the right split node that assures high accuracy, Information Gain is all about finding the best nodes that return the highest information gain. This is computed using a factor known as **Entropy**. Entropy defines the degree of disorganization in a system. The more the disorganization is, the more is the entropy. When the sample is wholly homogeneous, then the entropy turns out to be zero, and if the sample is partially organized, say 50% of it is organized, then the entropy turns out to be one.
This acts as the base factor in determining the information gain. Entropy and Information Gain together are used to construct the Decision Tree, and the algorithm is known as **ID3**.



Example:

We will use this link to explain in detail.
https://medium.com/analytics-vidhya/entropy-calculation-information-gain-decision-tree-learning-771325d16f

Lab Task:

1. Implement the code in python given in this tutorial and dry run the example.

https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53

2. Implement ID3 algorithm from scratch.

Gini Impurity/Index

If all elements are correctly divided into different classes (an ideal scenario), the division is considered to be pure. The Gini impurity (pronounced like "genie") is used to ***gauge the likelihood that a randomly chosen example would be wrongly classified by a certain node***. It is known as an "impurity" measure since it gives us an idea of how the model differs from a pure division.

The degree of the Gini impurity score is always between 0 and 1, where 0 denotes that all elements belong to a certain class (or the division is pure), and 1 denotes that the elements are randomly distributed across various classes. A Gini impurity of 0.5 denotes that the elements are distributed equally into some classes. The mathematical notation of the Gini impurity measure is given by the following formula:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

Example:

We will use this link to explain in detail.
https://medium.com/analytics-vidhya/classification-in-decision-tree-a-step-by-step-cart-classification-and-regression-tree-8e5f5228b11e

Lab Task:

3. Implement CART algorithm from scratch.
4. Implement Decision Tree in Azure on any dataset of your choice.

Reading Assignment:

https://arifromadhan19.medium.com/regrssion-in-decision-tree-a-step-by-step-cart-classification-and-regression-tree-196c6ac9711e