SUNBEAM

Machine Learning

Dr.Akshita Chanchlani

Topic :Decision Tree

Overview

- A **decision tree** is a decision support tool that uses a tree like model of decisions and their possible consequences
- A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules
- It is one way to display an algorithm that only contains conditional control statements
- Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods
- Tree based methods empower predictive models with high accuracy, stability and ease of interpretation
- Unlike linear models, they map non-linear relationships quite well
- Decision Tree algorithms are referred to as **CART (Classification and Regression Trees)**

Terminologies

- **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
- **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
- **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
- **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

Applications of Decision Tree

- It is one of the more popular classification algorithms being used in Data Mining
- Determination of likely buyers of a product using demographic data to enable targeting of limited advertisement budget
- Prediction of likelihood of default for applicant borrowers using predictive models generated from historical data
- Help with prioritization of emergency room patient treatment using a predictive model based on factors such as age, blood pressure, gender, location etc.
- Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goaln, and other measurements
- Because of their simplicity, tree diagrams have been used in a broad range of industries and disciplines including civil planning, energy, financial, engineering, healthcare, pharmaceutical, education, law, and business

How does Decision Tree work?

- Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems
- It works for both categorical and continuous input and output variables
- In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables

Steps

- Place the best attribute of the dataset at the **root** of the tree.
- Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.(homogenous)
- Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

Assumptions

- At the beginning, the whole training set is considered as the **root**
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are **distributed recursively** on the basis of attribute values
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach

Decision Tree Types

- **Categorical Variable Decision Tree (Classification)**
  - Decision Tree which has categorical target variable then it called as categorical variable decision tree
  - E.g.:- In an scenario of students data, where the target variable was “Student will play cricket or not” i.e. YES or NO.
- **Continuous Variable Decision Tree (Regression)**
  - Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree

Advantages of Decision Tree

- **Easy to Understand**
  - Decision tree output is very easy to understand even for people from non-analytical background
  - It does not require any statistical knowledge to read and interpret them
  - Its graphical representation is very intuitive and users can easily relate their hypothesis
- **Useful in Data exploration**
  - Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables
  - With the help of decision trees, we can create new variables / features that has better power to predict target variable
  - It can also be used in data exploration stage
  - For e.g., we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.