MACHINE LEARNING
Decision Tree

Trainer : Sujata Mohite
Email: sujata.mohite@sunbeaminfo.com

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 if-else
- 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

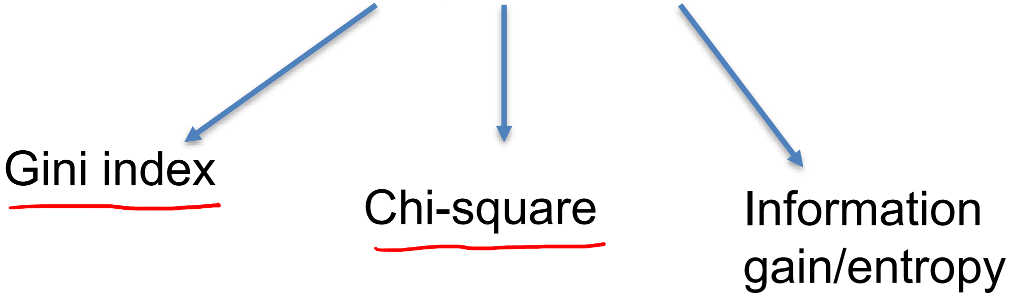
- **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
- **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



Gini index

Chi-square

Information gain/entropy

Steps

- Place the best attribute(most significant) 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(labels/class values)** 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.

Advantages of Decision Tree

- Decision trees implicitly perform variable screening or feature selection
- Decision trees require relatively little effort from users for data preparation
- **Less data cleaning required**
  - It requires less data cleaning compared to some other modelling techniques
  - It is not influenced by outliers and missing values to a fair degree.
- **Data type is not a constraint**
  - It can handle both numerical and categorical variables
- **Non-Parametric Method**
  - Decision tree is considered to be a non-parametric method
  - This means that decision trees have no assumptions about the space distribution and the classifier structure
- Non-linear relationships between parameters do not affect tree performance.
- The number of hyper-parameters to be tuned is almost null.

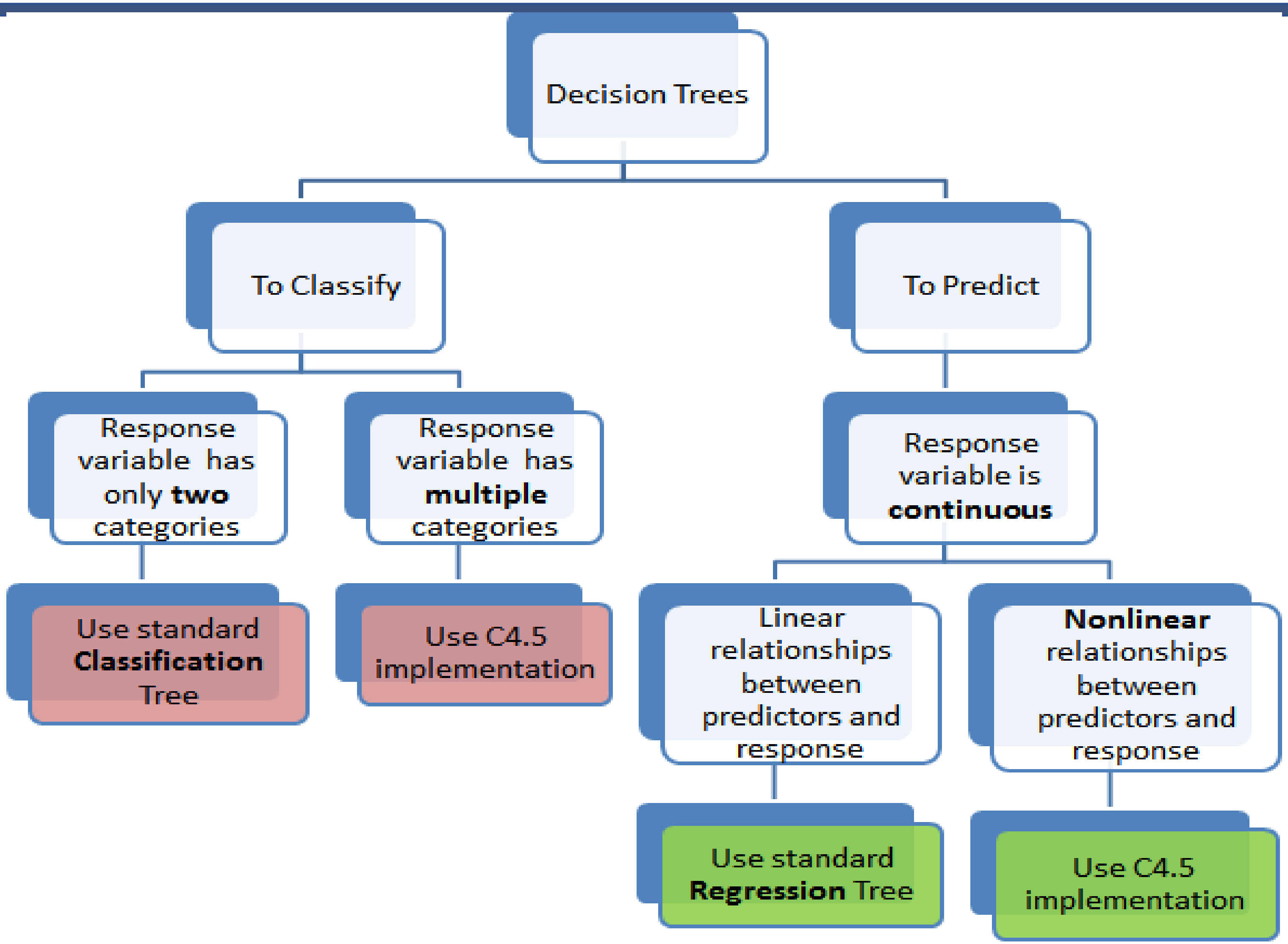
Disadvantages of Decision Tree

- **Over fitting (tree with too many levels)**
  - Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.
  - Over fitting is one of the most practical difficulty for decision tree models
  - This problem gets solved by setting constraints on model parameters and tuning
- **Not fit for continuous variables**
  - While working with continuous numerical variables, decision tree loses information, when it categorizes variables in different categories.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting
- Decision tree learners create *biased* trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree
- Calculations can become complex when there are many class label

Regression and Classification Tree

- Regression trees are used when dependent variable is continuous. Classification Trees are used when dependent variable is categorical
- In case of Regression Tree, the value obtained by terminal nodes in the training data is the mean response of observation falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mean value.
- In case of Classification Tree, the value (class) obtained by terminal node in the training data is the mode of observations falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mode value.
- Both the trees divide the predictor space (independent variables) into distinct and non-overlapping regions.
- Both the trees follow a top-down greedy approach known as recursive binary splitting.
- This splitting process is continued until a user defined stopping criteria is reached

Regression and Classification Tree



Decision Trees

To Classify

To Predict

Response variable has only **two** categories

Response variable has **multiple** categories

Response variable is **continuous**

Use standard **Classification** Tree

Use C4.5 implementation

Linear relationships between predictors and response

Nonlinear relationships between predictors and response

Use standard **Regression** Tree

Use C4.5 implementation

Commonly used algorithms - Gini Index

- Gini index 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
- It works with categorical target variable "Success" or "Failure".
- It performs only Binary splits
- Higher the value of Gini higher the homogeneity.
- CART (Classification and Regression Tree) uses Gini method to create binary splits.

Commonly used algorithms - Chi-Square

- It is an algorithm to find out the statistical significance between the differences between sub-nodes and parent node
- We measure it by sum of squares of standardized differences between observed and expected frequencies of target variable
- It works with categorical target variable “Success” or “Failure”.
- It can perform two or more splits.
- Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node
- It generates tree called CHAID (Chi-square Automatic Interaction Detector)

Commonly used algorithms - Information Gain

- 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
  Entropy = -p log2 p — q log2q
- **Steps to calculate entropy for a split:**
  - Calculate entropy of parent node
  - Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.
  - We can derive information gain from entropy as **1- Entropy.**

Thank You!!