Decision trees regression

Ari Quintal – Christian Adriel – Esau May – Jorge Canul – Pablo Martin

01

Decision tree builds

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes

The core algorithm for building decision trees called ID3 by J. R. Quinlan employs a top-down search through the space of possible branches with no backtracking. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain with Standard Deviation Reduction.

Ensemble Methods for Tree Regression

Bagging

Bagging generates multiple training sets by repeatedly and randomly sampling the original data with replacement . It applies the regression tree algorithm to each of these data sets independently. After training multiple regression tree models, Bagging combines their predictions by averaging them.

Boosting

Boosting is an ensemble learning method that creates a robust model by training sequential models that focus on improving predictions for previously misclassified data points. Once all models are trained, they are combined by giving more weight to the better-performing ones.

Random Forest

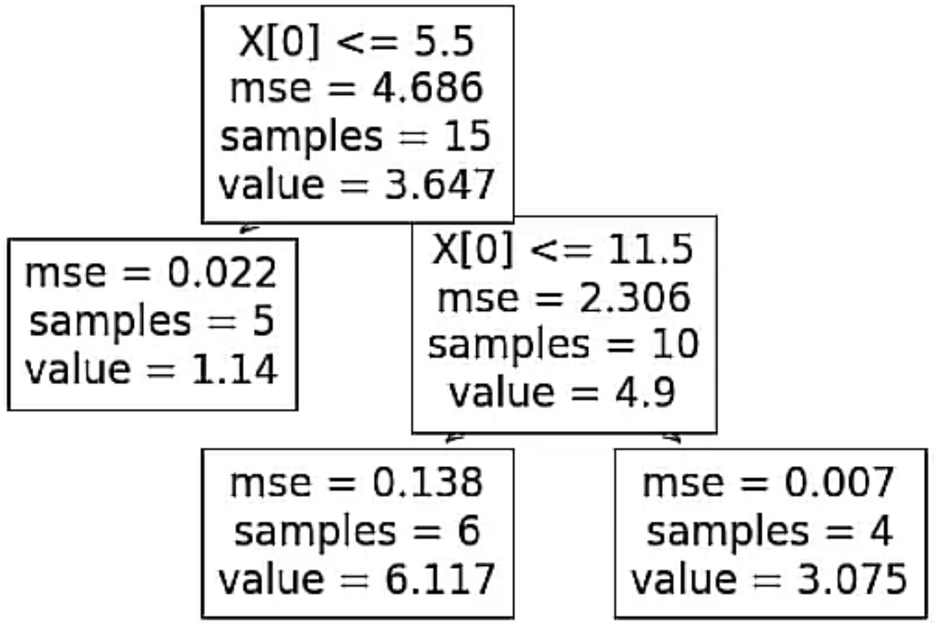
Random Forests have found applications in various domains, including finance, healthcare, natural language processing, and image analysis. It is useful for improving the accuracy of predictions, handling complex and noisy data, and assessing the significance of features.

Construction of DT

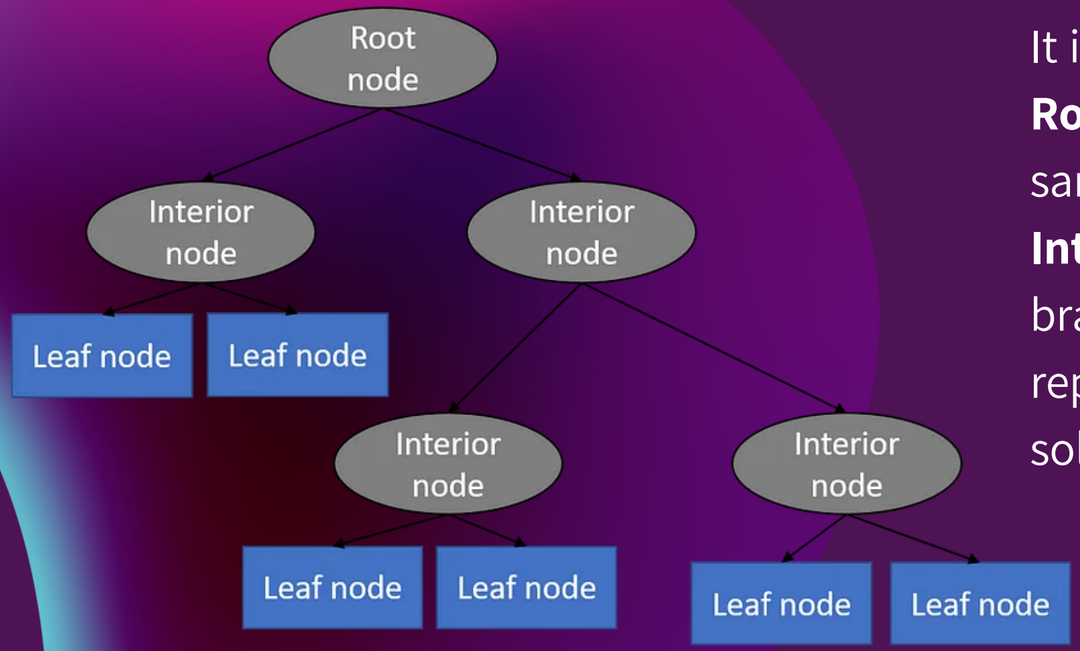
Standard deviation

$$S = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

Example



Root Node - Interior Nodes - Leaf Nodes

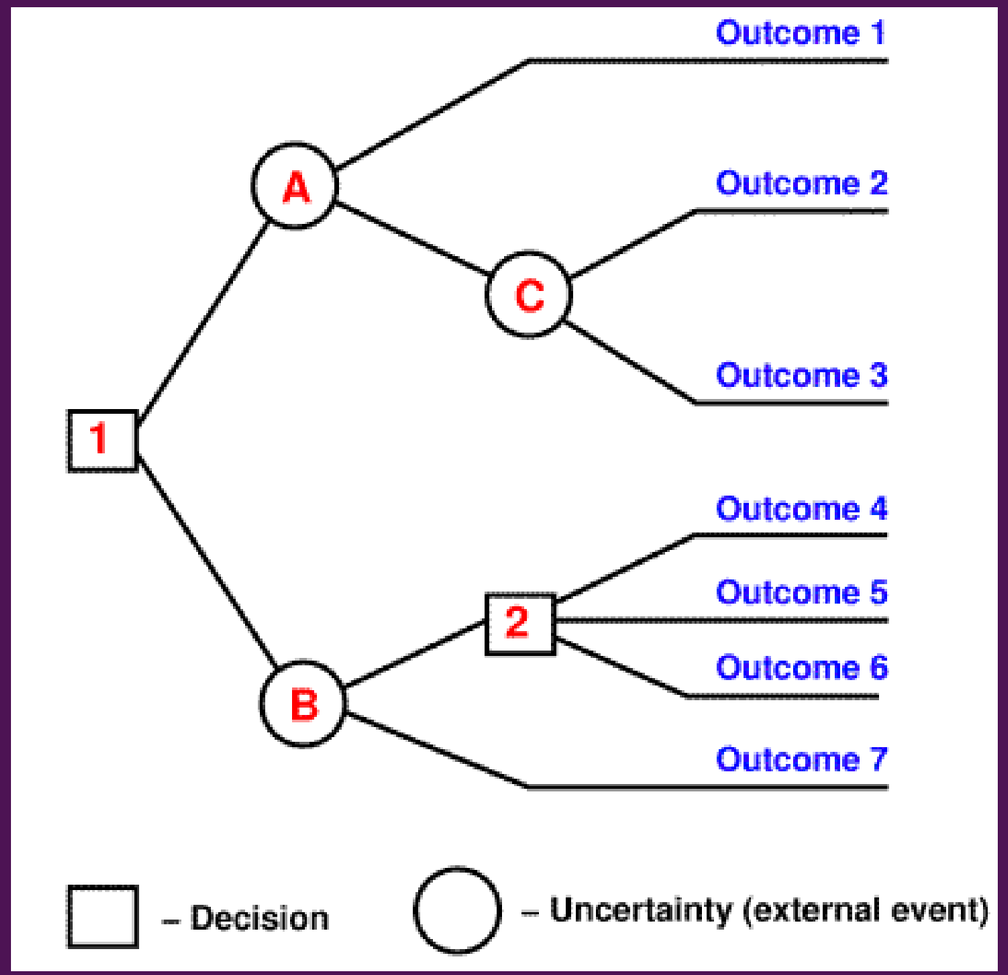


It is a tree-structured classifier with three types of nodes. The **Root Node** is the initial node which represents the entire sample and may get split further into further nodes. The **Interior Nodes** represent the features of a data set and the branches represent the decision rules. Finally, the **Leaf Nodes** represent the outcome. This algorithm is very useful for solving decision-related problems.

With a particular data point, it is run completely through the entirely tree by answering True/False questions till it reaches the leaf node. The final prediction is the average of the value of the dependent variable in that particular leaf node. Through multiple iterations, the Tree is able to predict a proper value for the data point.

Why we have to use decision tree algorithm?

1. It is considered to be the most understandable machine learning algorithm.
2. It can be used for classification and regression problems
3. Decision trees models data as a “tree” of hierarchical branches. The leaves represent predictions. Can easily model non linear relationships.



How does the decision tree work in regression models?

Structure of regression tress:

- Nodes: Represent specific features or attributes of the dataset
- Branches: Show the possible outcomes or values based of the features values
- Leaf nodes: Terminal nodes that contains the predicted values for the target variable

Building a regression tree

- The algorithm chooses the feature that best divides the data bases on a chosen criterio.
- Partition the data into subsets based on the selected features, repeating this process recursively until a stopping criterion is met.
- Refers to the condition that determines when the tree construction process stops. This condition prevents the tree from further dividing into smaller subsets, aiming to prevent overfitting and improve the model's ability to generalize well with new data.

Prediction

Once the tree is constructed, predictions for new data points involve traversing the tree based on the feature values until reaching a leaf node, which contains the predicted numerical value.

The Code.