Decision Trees
Epoch IIT Hyderabad

Chakka Surya Saketh
AI22BTECH11005

Introduction

Formally a decision tree is a graphical representation of all possible solutions to a decision. Most of the time they are used for classification problem. In a classification tree the value obtained by leaf nodes in the training data is the mode response of observation falling in that region but in a regression tree the value obtained by leaf nodes in the training data is the mean response of observation falling in that region.

Algorithms to split

1) Gini Impurity :
   It performs only binary splits
   Higher the value of gini higher the homogeneity

   $$\text{Gini impurtiy} = 1 - (p^2 + q^2) \text{ where p} = \Pr(success) \text{ and q} = \Pr(failure)$$

   Calculate gini score for spit using the weighted gini score of each node of that split and select the feature with the least gini impurity.

2) Chi squared :
   It can perform 2 or more splits.
   Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node.

   $$\text{Chi squared} = \sqrt{\frac{(Acutal - Expected)^2}{expected}})$$

   Select the split where chi squared is maximum.

3) Reduction in Variance :
   Reduction in variance is an algorithm used for continuous target variables.
   This algorithm uses the standard formula of variance to choose the best split.
   The split with lower variance is selected as the criteria to split the population.
   The node with lower variance is selected as the criteria to split.

Algorithm

Decision trees are constructed using a top-down, recursive partitioning approach. At each node, the algorithm selects the best feature to split the data based on certain criteria, such as Gini impurity or information gain. The process continues until a stopping condition is met, resulting in a tree with leaf nodes representing the final predictions.

Overfitting

Usually, real-world datasets have a large number of features, which will result in a large number of splits, which in turn gives a huge tree. Such trees take time to build and can lead to overfitting. That means the tree will give very good accuracy on the training dataset but will give bad accuracy in test data.
We can tackle this problem by setting parameters like max_depth, min_samples_split, max_leaves, max_features where the splitting stops if any of these conditions are met.we use grid search cv to find the best values for these parameters.
We can also prune a tree-

1) Pre-pruning: We can stop growing the tree earlier, which means we can prune/remove/cut a node if it has low importance while growing the tree.
2) Post-pruning: Post-construction, pruning removes unnecessary branches to generalize the model better and enhance predictive performance on unseen data.

Ensemble techniques like random forests and boosting combine various decision trees to improve their accuracy on unseen data (prevent overfitting).

Conclusion

Decision trees are versatile algorithms that have proven effective in various applications due to their simplicity, interpretability, and ability to handle non-linear relationships.While decision trees are prone to overfitting following the methods mentioned above can enhance their accuracy.Decision trees find applications in various domains like healthcare, finance, marketing etc. Understanding decision trees and their applications is crucial for practitioners seeking powerful and interpretable machine learning models.