Decision Trees

David Robinson

Decision Trees

A **decision tree** is a supervised learning algorithm used for classification tasks. It is structured like a tree, where each branch node represents a decision based on feature values and each leaf node represents the final decision of that tree.

Iterative Dichotomiser 3

The Iterative Dichotomiser 3 (ID3) generates a decision tree from a given dataset using a top-down, greedy approach.

1. At each node, it tests each attribute to determine the best split
2. It selects the attribute that maximus information gain or minimizes entropy.
3. The resulting tree is used to classify future data samples.

Algorithm for Building a Decision Tree

1. **Calculate Entropy:** Compute the entropy of each attribute in the dataset to measure the uncertainty in class distribution.

   Entropy measures the homogeneity or uncertainty within a dataset. A completely homogeneous dataset (all samples belong to one class) has an entropy of 0. A dataset that is equally divided among classes has an entropy of 1.

   $$\textbf{Entropy}(S) = -\sum_{i=1}^{n} p(i) \log_2 p(i)$$

   where S is the dataset or sample, $p(i)$ is the proportion of elements in class i within S, and n is the number of classes.

2. **Split the Data:** Split the dataset into subsets using the attribute that results in the lowest entropy.
3. **Create a Node:** Create a decision tree node based on the attribute with the lowest entropy.
4. **Recursive Process:** Repeat the process for each subset, using the remaining attributes until no further splits are possible.

Limitations

1. Overfitting is a common issues with decision trees. A tree that perfectly classifies the training data may not generalize well to unseen data. Also, the tree may fit noise in the training data, leading to poor performance on test data.
2. Decision trees can struggle with continuous numerical variables, as they split them into discrete intervals, potentially leading to a loss of information.

Evaluation Methods

Pre-pruning stops growing the tree during construction when there is not enough data to make reliable decisions. **Post-pruning** first grows the entire tree, then removes nodes that lack sufficient data for making reliable predictions.

Methods

- **Cross-validation:** Use a hold-out set to evaluate the utility of subtrees and determine if pruning will improve performance.
- **Statistical Testing:** Test whether the observed patterns are statistically significant or likely to have occurred by chance.
- **Minimum Description Length (MDL):** Compare the complexity of the tree (hypothesis) with the simplicity of remembering execptions. If the tree is overly complex for the data it explains, then it should be pruned.

Gini Index

The **Gini index** is a measure of impurity used to evaluate splits in decision trees, where a lower Gini index value indicates a purer node which is dominated by samples from a single class.

$$\textbf{Gini Index} = 1 - \sum_j p_j^2$$

where p_j is the probability of a sample belonging to class j.

Random Forest

A **random forest** is an ensemble learning method that combines multiple decision trees to create a stronger model.

Advantages

1. Versatile for classification and regression
2. Can handle high-dimensional data
3. Handles class imbalances