Chapter 1

Decision Trees for Classification

A decision tree is a popular machine learning algorithm used for both regression and classification problems. In this discussion, we will focus on its application in classification tasks.

1.1 Decision Trees for Classification

A decision tree for classification uses a tree structure to predict the class of an object based on its features. The tree is made up of nodes that split the data based on a feature value, and leaves that represent a class label. The idea is to create a tree that has minimum impurity, i.e., at the end of the tree, we would like each leaf to contain data points that belong to a single class.

1.2 Gini Impurity

Gini impurity is a measure of misclassification, which applies in a multiclass classifier context. It gives an idea of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

The Gini impurity for a node of the tree is calculated as:

$$\text{Gini}(p) = 1 - \sum_{i=1}^{J} (p_i)^2 \tag{1.1}$$

where p_i is the fraction of items classified to label i at a node and J is the total number of classes.

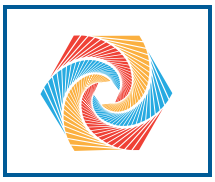
A Gini impurity of 0 is the best score, where all elements in a partition fall into a single category.

1.3 How Gini Impurity is Used

During the construction of a decision tree, the best feature to split on at each node is chosen by minimizing the Gini impurity of the child nodes. The algorithm will consider all features and all possible split points for each feature to find the split that yields the lowest weighted average Gini impurity.

Appendix A

Answers to Exercises



Index