DECISION TREES

Rishitha Pallala

8 August 2023

1 Introduction

The decision tree are predictive models which have flow-chart like structure. the main components of the decision trees are nodes. The nodes from which the tree starts is called 'ROOT' node.The nodes at which the outcome is given and the tree ends is called 'LEAF' node . The nodes between root and leaf nodes are called 'INTERMEDIATE' nodes.
the trees are upside down and each node (apart from leaf node) are true-false statements leading to further nodes.The decision trees can be divided into classification trees for dividing things into categories and into regresssion trees for numerical values.

2 Classification trees

2.1 Impurities and Gini impurity

To select the position of the node in decision trees, we calculate the impurity of the nodes. A node is impure if its data is split into both the true and false categories. To calculate the impurity of the node 'gini impurity ' is mostly is used. Other methods are Information and entropy impurities.

To calculate the gini impurity of the mother node , we calculate the impurity of the daughter nodes given by the equation (1)

$$impurity = y = 1 - Pr(statement : false)^2 - Pr(statement : true)^2 \quad (1)$$

$$gini(mnode) = \frac{3}{7} \times nodeA + \frac{4}{7} \times nodeB \quad (2)$$

Then we calculate the weighted average of the daughter cell to get gini impurity of mother node. For example if mother node has 3 datapoints to daughter nodeA and 4 datapoints to daughter nodeB, Equation(2) gives gini impurity of mother node. The node with least Impurity will be selected.

3 Regression trees

The regression tree is different from the classification tree as it's leaf nodes have numerical values. To reduce the variance of the tree , we set a minimum threshold on the node to further divide.

The nodes are decided by trying different thresholds and calculate the SSR(sum of squared residuals) for all the values in that threshold. the threshold with least value is considered. The equation(3) refers to SSR calculation.

$$ssr = \sum_{i=1}^{n} (y_{average} - y_{observed})^2 \tag{3}$$

Thus,Decision trees are used for classification and regression tasks, providing easy-to-understand models.