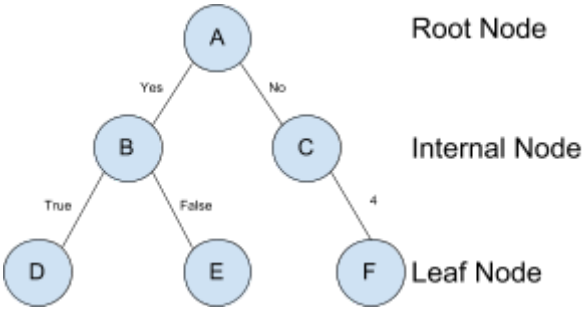
Everything about Decision Trees

Structure

A decision tree is a set of branching options that end in a classification. Decision trees work better with nominal variables but can also be used with numeric ones, by splitting them into bins. The structure of a decision tree consists of nodes and branches. Each node represents a single variable. Each branch represents a possible value of that variable. The first node in a decision tree is called root node or top node. Nodes that don't branch out to other nodes are called leaf nodes. When a leaf node is reached, the object in question is classified. The following graph depicts the structure of a simple decision tree.



Learning

The learning process of most decision tree is based on the ID3 tree building algorithm, which implements the divide and conquer approach. The algorithm starts splitting on the variable that gives the highest information gain. Information can be seen as a measure of uncertainty. It is calculated using the formula:

$$I(e) = -log_2(P(e))$$

Where P(e) is the probability of an event e. The weighted average information across all possible values of a variable is called Entropy. It is calculated as the sum of the probability for each event, multiplied by the variables information value.

$$H(X) = -\sum P(x_i) log_2(P(x_i))$$

Entropy is used to calculate the Information Gain of a variable.

$$IG = H(Outcome) - H(Outcome|Input)$$

The ID3 algorithm picks the top node of a tree by calculating the information gain of the output class for each input variable. It chooses the one that minimises uncertainty; has the highest Info Gain. It then creates branches for each possible value that variable can take. Branches are added by repeating the information gain calculations, based on the data of the location of the current node in the tree. If all objects in a leaf node are in the same class, no more branches are added. Otherwise, the algorithm stops when all data has been accounted for.