What is Information Gain and Gini Index in Decision Trees?

Introduction

Decision Trees are supervised machine learning algorithms that are best suited for classification and regression problems. These algorithms are constructed by implementing the particular splitting conditions at each node, breaking down the training data into subsets of output variables of the same class.

This process of classification begins with the root node of the decision tree and expands by applying some splitting conditions at each non-leaf node, it divides datasets into a homogeneous subset.

The 'knowledge' learned by a decision tree through training is directly formulated into a hierarchical structure. This structure holds and displays the knowledge in such a way that it can easily be understood, even by non-experts." (From)

However, pure homogeneous subsets is not possible to achieve, so while building a decision tree, each node focuses on identifying the an attribute and a split condition on that attribute which miminzing the class labels mixing, consequently giving relatively pure subsets.

In order to check "the goodness of splitting criterion" or for evaluating how well the splitting is, various splitting indices were proposed. Some of them are gini index and information gain.

In the blog discussion, we will discuss the concept of entropy, information gain, gini ratio and gini index.

What is Entropy?

Entropy is the degree of uncertainty, impurity or disorder of a random variable, or a measure of purity. It characterizes the impurity of an arbitrary class of examples.

Entropy is the measurement of impurities or randomness in the data points.

Here, if all elements belong to a single class, then it is termed as "Pure", and if not then the distribution is named as "Impurity".

It is computed between 0 and 1, however, heavily relying on the number of groups or classes present in the data set it can be more than 1 while depicting the same significance i.e. extreme level of disorder.

In more simple terms, If a dataset contains homogeneous subsets of observations, then no impurity or randomness is there in the dataset, and if all the observations belong to one class, the entropy of that dataset becomes zero

What is Information Gain?

The concept of entropy plays an important role in measuring the information gain. However, "Information gain is based on the information theory".

Information gain is used for determining the best features/attributes that render maximum information about a class. It follows the concept of entropy while aiming at decreasing the level of entropy, beginning from the root node to the leaf nodes.

Information gain computes the difference between entropy before and after split and specifies the impurity in class elements.

Information Gain = Entropy before splitting - Entropy after splitting

Given a probability distribution such that

P = (p1 , p2 ,.....pn),
and where (pi) is the probability of a data point in the subset of

Therefore, Entropy is defined as the
Generally, it is not preferred as it involves 'log' function that results in the computational complexity. Moreover;

Information gain is non-negative.
Information Gain is symmetric such that switching of the split variable and target variable, the same amount of information gain is obtained. (Source)
Information gain determines the reduction of the uncertainty after splitting the dataset on a particular feature such that if the value of information gain increases, that feature is most useful for classification.
The feature having the highest value of information gain is accounted for as the best feature to be chosen for split.

What is Gain Ratio?

Proposed by John Ross Quinlan, Gain Ratio or Uncertainty Coefficient is used to normalize the information gain of an attribute against how much entropy that attribute has. Formula of gini ratio is given by

Gain Ratio=Information Gain/Entropy

From the above formula, it can be stated that if entropy is very small, then the gain ratio will be high and vice versa.
Be selected as splitting criterion, Quinlan proposed following procedure,

First, determine the information gain of all the attributes, and then compute the average information gain.
Second, calculate the gain ratio of all the attributes whose calculated information gain is larger or equal to the computed average information gain, and then pick the attribute of higher gain ratio to split.

What is Gini Index?

The gini index, or gini coefficient, or gini impurity computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. It works on categorical variables, provides outcomes either be "successful" or "failure" and hence conducts binary splitting only.

The degree of gini index varies from 0 to 1,
Where 0 depicts that all the elements be allied to a certain class, or only one class exists there.
The gini index of value as 1 signifies that all the elements are randomly zdistributed across various classes, and
A value of 0.5 denotes the elements are uniformly distributed into some classes.
It was proposed by Leo Breiman in 1984 as an impurity measure for decision tree learning and is given by the equation/formula;
where P=(p1 , p2 ,.....pn) , and pi is the probability of an object that is being classified to a particular class.
Also, an attribute/feature with least gini index is preferred as root node while making a decision tree.

Gini Index vs Information Gain

Following are the fundamental differences between gini index and information gain;
Gini index is measured by subtracting the sum of squared probabilities of each class from one, in opposite of it, information gain is obtained by multiplying the probability of the class by log (base= 2) of that class probability.
Gini index favours larger partitions (distributions) and is very easy to implement whereas information gain supports smaller partitions (distributions) with various distinct values, i.e there is a need to perform an experiment with data and splitting criterion.
The gini index approach is used by CART algorithms, in opposite to that, information gain is deployed in ID3, C4.5 algorithms.
While working on categorical data variables, gini index gives results either in "success" or "failure" and performs binary splitting only, in contrast to this, information gain measures the entropy differences before and after splitting and depicts the impurity in class variables.

Conclusion

It is often observed that decision trees are very catchy to understand because of their visual representation/interpretation. They can handle the pool of quality data that can be validated by statistical techniques and are cost-effective computationally.
It can also handle high dimensional data with actual good accuracy. Besides that, various features selection methods are used in building the decision tree from root nodes to leaf nodes and the same are listed in the blog.
Hope, this article is helpful in understanding the basis of the decision tree in the context of entropy, information gain, gini ratio and gini index.

equation/formula;
where P=(p1 , p2 ,.....pn) , and pi is the probability of an object that is being classified to a particular class.

Also, an attribute/feature with least gini index is preferred as root node while making a decision tree.

Gini Index vs Information Gain

Following are the fundamental differences between gini index and information gain;
Gini index is measured by subtracting the sum of squared probabilities of each class from one, in opposite of it, information gain is obtained by multiplying the probability of the class by log (base= 2) of that class probability.
Gini index favours larger partitions (distributions) and is very easy to implement whereas information gain supports smaller partitions (distributions) with various distinct values, i.e there is a need to perform an experiment with data and splitting criterion.
The gini index approach is used by CART algorithms, in opposite to that, information gain is deployed in ID3, C4.5 algorithms.
While working on categorical data variables, gini index gives results either in "success" or "failure" and performs binary splitting only, in contrast to this, information gain measures the entropy differences before and after splitting and depicts the impurity in class variables.