K-Nearest Neighbor (KNN)

The supervised machine learning method k-nearest neighbors (KNN) is a basic, simple technique that may be used to tackle simultaneously classification and regression issues. The KNN method is a basic supervised machine learning technique that may be used to tackle classification and regression issues. It's simple to set up and comprehend. (Harrison, 2018)

The K closest neighbors method, often known as the KNN algorithm, is a simple algorithm that uses all this dataset as its training dataset. When a predictions is produced for an undefined data instance, it searches the whole testing dataset for the k-most similar instances and delivers the data with the most similar examples as the predictions. Both classification and regression prediction issues can be solved with KNN.

There are 3 broken down parts for KNN that is necessary for implementing and applying KNN's algorithm in the process of classification and regression prediction modeling problems :

Step 1: Calculate Euclidean Distance

In this step, first the value of distance between two rows should be identify in the dataset. Rows of data are primarily made up of numbers, and drawing a linear line between two row or vector of quantities is a simple approach to compute the distance between them which works in 2D and 3D, and it grows well to greater dimensions. The straight line between 2 vectors can be calculate with using Eculidean distance measure. The formula is Euclidean Distance = sqrt(sum i to N (x1_i – x2_i)^2). As we add over all columns, x1 is the 1st row of data, x2 is the 2nd row of data, and I is the index to a specific column. The lesser the amount of Euclidean distance, the more accurate records would be. A value of zero indicates that no change between the records.

Step 2: Get Nearest Neighbors

In this step, the k nearest examples, as specified by our distance metric, are neighbors for a new data point in the collection. To find the neighbors for a new data point within a dataset, first it is necessary to find out how far each record in the set of data is from the new data point. This may be accomplished with the help of the distance function we developed before. It is also necessary to arrange all of the records in the training data by

relative distance to the updated data after the distances have been determined. The top k can then be returned as the most comparable neighbors. This may be accomplished by storing the distance between each item in the dataset as a tuple, sorting the lists of tuples by and then retrieving the neighbors.

Step 3: Make Predictions

In this step, to create predictions, the most comparable neighbors from the training dataset might be employed. In the process of predictions, we may return the neighborhood's most well-represented class. This may be accomplished by using the max() function on the list of neighbor output values. The max() method accepts a set of different class labels and performs the counts on the lists of class labels for each classifier in the set, given a list of class labels seen in the neighbors. (Brownlee, 2019)

Hence, The selected separation function is critical in deciding the final classification result because the k-NN classification is dependent on calculating the distance between the sample group with each of the training samples. The main objectives of this algorithms was to understand the K-Nearest Neighbors algorithms step-by-step, to test k-Nearest Neighbors using a real dataset, and to create a forecast for fresh data using k-Nearest Neighbors.

Decision Tree

A selection A tree is a graphical depiction of all potential decision-making options. In today's supervised learning settings, tree-based algorithms are the most widely utilized algorithms. They're easy to understand and envision, and they're quite adaptable. Tree-based methods may be used to solve simultaneously regression and classification issues.

The supervised machine learning techniques include the decision tree Algorithm. It may be used to solve a classification challenge or a regression issue. The objective of this technique is to construct a model which predictions the value of the dependent variable, as well as the decision tree solves the issue by using the tree representation, where the leaf node correlates to a classification model and characteristics are recorded on the central node of the tree.

There is a Parent node which is higher hierarchically, Child node which is lower hierarchically, Root node which the tree starts, leaf node which do not have any children are leaf nodes, internal nodes which have both a parent and at least one child, Splitting which divide into sub-nodes, decision node, Pruning, Branch.

There are two types of decision tree i.e. Regression Tree, Classification Tree.

Regression Tree

When the dependent variable is continuous, a regression tree is employed. The mean response of observations occurring in that area is the value acquired by leaf nodes in the training data. As a result, if an unknown data observation falls within that range, the mean value is used to make a forecast.

Classification Tree

When the dependence variable is categorical, a classifications tree is employed. The mode responses of observations occurring in that area is the value produced by leaf nodes in the training data. It takes a wasteful top-down approach.

The most commonly used algorithm for splitting is given bellow:

1. Gini impurity

   According to Gini, if two things are randomly chosen from a group, they must belong to the same classes, and the chance of this is one if the population is accurate. The category

goal variable "Success" or "Failure" is used. It only does binary splits. The greater the Gini coefficient, the greater the homogeneity. Regression and classification To make binary splits, Tree employs the Gini technique.

Ways to calculate Gini impurity for a split:

1) Minus the sum of the squares of likelihood for successes and failures from one to get Gini impurity for sub-nodes.
   1-(p2+q2), where p = P(Success) and q=P(Failure) (Failure)
2) Calculate Gini again for split based on the weighted Gini scores of every split node.
3) For the split, choose the attribute with lowest Gini impurity.

2. Chi-Square

It's a method for determining the statistically significance of differences among sub-nodes and the parent node. The square root of standardized deviations between observed and predicted frequency of the dependent variable is used to calculate it. The category goal variables "Success" or "Failure" is used.

It has the ability to do two or more splits. The statistical evidence of discrepancies among sub-node and Parent node increases as the Chi-Square value increases. The formula for calculating the Chi-Square of every node is Chi-square = ((Actual — Expected)2 / Expected)1/2. It produces the CHAID tree.

Ways to calculate Chi-Square for a split:

1) Calculate Chi-square to every node of the split by adding the deviations for Success and Failure.
2) Calculate Chi-square for split by adding the total of all Chi-square of Success and Failure to every node of the split.
3) Choose the split with the highest Chi-Square.

3. Information Gain

A less impurity node needs fewer detail to be described, whereas a more impurity node needs greater. Entropy is a measure of disorder in a system that is defined by information theory. The entropy of a sample that is fully homogenous is zero, whereas a sample that is evenly split (50 percent — 50 percent) has an entropy of one.

Ways to calculate entropy for a split:

1) Calculate the parent node's entropy.

2) Calculate the rolling average including all sub-nodes accessible in the divide and the entropy for every particular split node. The less entropy there is, the better.

3) Calculate the gain ratio as follows, then divide the node with largest information gain.

4. Reduction in Variance

   The methods for the category response variable have been described up to this point. For continous dependent variable, a reductions in variance technique is utilized (regression problems). It's used to represent dependent variable. To determine the optimum split, this method use the usual variance formula. The split with the least variance is chosen as the criterion for splitting the data.

   Ways to calculate Variance:

   1) Calculate the variance for every node separately.
   2) As a rolling average of every node's variance, calculated variance for each split.
   3) The node with the lowest variance is chosen as the divide criterion. (Analytics Vidhya, 2020)

In Summary, According to Gini, if two things are chosen at random from a group, they must belong to the same classes, and the chance of this is one if the population is purified. Reducing in variance is an algorithm used for continuously dependent variable (regression issues) to determine the statistically significance of the difference among sub-nodes and parent nodes. To determine the optimum split, this method use the usual formula od variance. The info needed to describe a less impurity node is less, whereas a more impurity node requires more.