# Use of Distribution Algorithms, for the Construction of a Classification and Regression Tree

**MSc. Adem Meta**
*Cuyahoga Community College, USA*

## Abstract

One of the most important processes in the construction of the classification and regression trees is the distribution of a given data. There are numerous algorithms for predicting continuous variables or categorical variables from a set of continuous predictors and/or categorical factor effects. In this paper I address the problem of learning various types of algorithms to be used to get a optimal decision trees from data base. In particular, we study online machine learning algorithms for learning classification and regression trees, linear model trees, option trees for regression, multi-target model trees, and ensembles of model trees from a given data. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. The core algorithm for building decision trees called, greedy search through the space of possible branches uses *Entropy* and *Information Gain* to construct a decision tree. In this paper, through a concrete example, I will explicitly look at the use of four algorithms such as the Gini Index, Chi-Square, Entropy and the Variance Reduction on which node will be the distribution of a database. Once the data base is small, no doubt that the calculation is much more simple than in the case of a large database.

**Keywords:** distribution algorithms, regression tree, classification.

## Introduction

The basic classification and regression algorithms are considered to be one of the best learning methods and used the most. Methods based on the classification tree provide predictive models with very good precision, stability and very ease of interpretation. They represent non-linear links quite well and are suitable for solving any classification or regression problems. Decision trees use multiple algorithms to decide when to split a node into two or more sub-nodes. The creation of subunits increases the homogeneity of the resulting subunits. Thus, the purity of the node increases with respect to the target variable. The crucial tree divides the nodes into all available variables and then selects the resulting partition with the most homogeneous sub nodes.

The choice of algorithms is also based on the type of responsible variables. Let's look at the four most used algorithms in the decision tree using the following example: Let's take a class of 36 students with three variables Gender (male/female), Class (XI / XII) and Height (160 cm up to 180 cm, (160,170) and (170,180)), 18 of which play basketball at leisure. We want to create a model to predict who will play basketball

during his or her free time? In this problem, we should highlight the students who play basketball in their free time based on gender, grade and height. This is the structure where the decision tree helps us to distinguish students based on all the values of the three variables and identify the variable that creates the best homogeneous groups of students (that are heterogeneous to one another). Below we can see that variable classes are able to identify the best homogeneous groups compared to the other two variables.

A. Distribution by gender

| Gender M/F | | | |
|---|---|---|---|
| **Gender** | **Number of students** | **Play basketball** | **Percentage** |
| **Female** | **16** | **4** | **25%** |
| **Male** | **20** | **14** | **70%** |
| **Totally** | **36** | **18** | **50%** |

Table 1: Distribution by gender

B. Distribution by height

| Hight(>170 or<170) | | | |
|---|---|---|---|
| **Height** | **Number of students** | **Play basketball** | **Percentage** |
| **>170cm** | **20** | **12** | **60%** |
| **<170 cm** | **16** | **6** | **37.5%** |
| **Totally** | **36** | **18** | **50%** |

Table 2: Distribution by height

C. Distribution by classes

| Class(XI or XII) | | | |
|---|---|---|---|
| **Classes** | **Number of students** | **Play basketball** | **Percentage** |
| **XI** | **16** | **6** | **37.5%** |
| **XII** | **20** | **12** | **60%** |
| **Totally** | **36** | **18** | **50%** |

Table 3: Distribution by classes

As mentioned above, the decision tree identifies the most important variable and what is the value that gives the best homogeneous population groups. How is variability and division identified? To do this, the decision tree uses different algorithms, which we will discuss below.

**How to decide when a tree should be distributed?**

The decision to make strategic separation greatly affects the accuracy of a tree. The

decision criterion is different for classification and regression trees. Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of subunits increases the homogeneity of the resulting subunits. In other words, we can say that the purity of the node increases with respect to the responsible variable. The crucial tree divides the nodes into all available variables and then selects the resulting division into the most homogeneous sub nodes. The choice of algorithms is also based on the type of responsible variables.

**Gini index**

The Gini index says, if we choose two quantities from a population at random then they should be in the same class and the probability for this is 1 if the population is pure.
1. In the case of categorical variables, our target may be "Success" or "Non-Success";
2. Perform only Binary divisions;
3. Gini's higher value, with higher homogeneity;
4. CART (Tree Classification and Regression) uses the Gini method to create binary separation.
Steps to calculate the Gini index for a split:
1. How to calculate Gin for a subunit, using the sum of the formula for the lowest probability for success and failure $Gini = p^2 + (1-p)^2$where (p-success and 1 -p-failure);
2. Calculate Gini for a division using the weighted Gini score of each node of the division split;
In the table above, we divide the population using two variables given, such as Gender and Class. Now, I want to identify which partition produces more homogeneous subunits using the Gini index.

- **Calculating Gini for distribution on node gender**

1. Calculate, Gini for female sub-node = $0.25^2 + 0.75^2 = 0.62.5$;
$0.25^2 + 0.75 = 0.62.5$2. Calculate, Gini for male sub-node=$0.70^2 + 0.30^2 = 0.58$
3. Calculate weighted Gini for distribution Gender = $\frac{20}{36} * 0.58 + \frac{16}{36} * 0.625 = 0.62$

- **The same way for the distribution according to classes**

1. Gini for sub-nodes Class XI = $0.375^2 + 0.625^2 = 0.53$;
2. Gini for sub-nodes Class XII =$0.6^2 + 0.4^2 = 0.52$;
3. Calculate weighted Gini for distribution Classes=$\frac{16}{36} * 0.53 + \frac{20}{36} * 0.52 = 0.52$;

- **The same way for the distribution according to height**

1. Gini for sub-nodes for height <170cm = $0.6^2 + 0.4^2 = 0.52$;
2. Gini for sub-nodes for height >170 cm =$0.0.375^2 + 0.625^2 = 0.53$;
3. Calculate weighted Gini for distribution for height=$\frac{16}{36} * 0.52 + \frac{20}{36} * 0.53 = 0.53$

From the above calculations, we notice that Gin's gender score is higher than the distribution in the class and the high, so the division of the joints will be made for the gender.

**Chi-Square**

It is an algorithm that detects the statistical significance between the differences of a sub-node and parent node. We measure it with the sum of the square of the difference between the observed values and the expected values by dividing it with the expected values of the target variables.

1. Works with the target categorical variable "Success" or "Do not Succeed".
2. It can perform two or more partitions.
3. The higher the value of Chi-Square is the higher the statistical significance of the differences between the sub-nodes and the parental node.
4. Hi-square of each node is calculated using the formula: Chi-square=$\frac{(O_i - E_i)^2}{E_i}$, where Oi observed values and Ei expected values
5. Generates a tree named CHAID (Chi-square Automatic Interaction Detector). This type of test is a technique used to find a decisive tree based on the fit, or to regulate the importance of the test. CHAID is a tree classification technique that not only evaluates complex interactions among forecasters but also shows the final modeling in a tree diagram easy to interpret. Tree trunk represents the ultimate database modeling. CHAID then creates a first layer of "branches" by displaying the values of the stronger prediction variables. CHAID automatically determines how to group this viewer's values into the number of manageable categories.

**Steps to calculate Chi- square for distribution:**

- Calculate the Chi-square for each individual node by calculating the average quadratic deviation for Success and Non-Success (play and do not play basketball);
- Chi-square Distribution Calculation using the Chi-Squares sum for success or failure for each split node;
- First we look at and calculate the value for the Female node, namely calculate the current value for "Basketball Play" and "Do not Play Basketball", here are 4 and 14 respectively;
- Calculate the expected value for "Play Basketball" and "Do not Play Basketball", here would be 4 and 14 for both, because the parent node has a probability of 50% and we have applied the same probability of counting Women (16);
- Calculate the quadratic average deviations using the formula above;
- Calculate Chi-Squares of the "Basketball Games" and "Do not Play Basketball" using the formula above. This is what we see in the table below;
- Follow the same steps for calculating the Chi-square Value for the male node;
- At the end add all Chi-square for separated gender;

| Nodes | Play Basketball | Don't play basketball | Totally | Expectation to play basketball | Expectation not to play basketball | Deviation play bask | Deviation not play basketball | Chi-square | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Play Basketball | Don't play basketball |
| Female | 6 | 12 | 16 | 8 | 8 | -2 | 4 | 0.5 | 2 |
| Male | 12 | 6 | 20 | 10 | 4 | 2 | -4 | 0.4 | 1.6 |
| sum | 0.9 | 3.6 | | | | | | | |
| | | | | | | | | | |
| Totally | 4.5 | | | | | | | | |

Tabela 4: Chi-square for gender

**Distribution by classes:**

| Node | Play basketball Bask | Don't play basketball | Totally | Deviation to play basketball | Deviation not play basketball | Deviation play basketball | Deviation not play basketball | Chi-square | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Play basketball | Don't play basketball |
| XI | 6 | 10 | 16 | 8 | 8 | -2 | 2 | 0.5 | 0.5 |
| XII | 12 | 8 | 20 | 10 | 10 | 2 | -2 | 0.4 | 0.4 |
| Sum | 0.9 | 0.9 | | | | | | | |
| Totally | 1.8 | | | | | | | | |

Table 5: Chi-kateror for distribution by classes

From the table above, we see that Chi-square also identifies gender division is more important than class division.

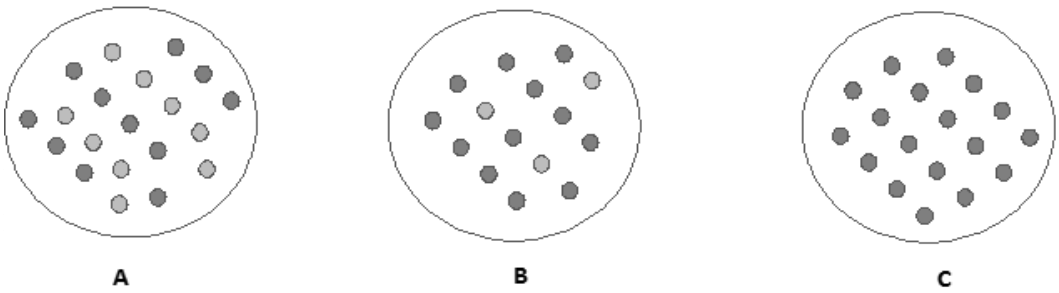


Figure 1 : Image A,B,C

We can conclude that the least impurity node requires less information to describe it. The most impurity node requires more information. Information theory is a measure to determine this degree of disorganization in a system known as Entropy. If the example is completely homogeneous then the entropy is zero and if the example is equally divided (50% - 50%), entropy is one. Entropy can be calculated using the formula: $Entopy = -plog_2p - (1-p)log_2(1-p)$.

Where p and 1-p are the probability of success and non-success respectively in that node. Entropy is also used with categorical target variables. It chooses the division that has the lowest entropy compared to the parent node and the other divisions. The smaller the entropy, the better the distribution.

Let's use this method to identify the best partition for the above example.

- Entropy of the parent node - (18/36) log2 (18/36) - (18/36) log2 (18/36) = 1. This indicates that it is an impure node;
- Entropy for the female node = - (4/16) log2 (4/16) - (12/16) log2 (12/16) = 0.81 and male node - (14/20) log2 (14/20 ) - (6/20) log2 (6/20) = 0.88;
- Entropy for division Gender = Weighted Entropy of Sub-Joints = (16/36) * 0. 81+ (20/36) * 0.88 = 0.85;
- Entropy for node Class XI, - (6/16) log2 (6/16) - (10/16) log2 (10/16) = 0.95 and node Class XII, - (12/20) log2 (12/20 ) - (8/20) log2 (8/20) = 0.970;
- Entropy for division by Class = (16/36) * 0.95+ (20/36) * 0.97 = 0.96;

From above we can see that the entropy for gender division is the lowest among all, so the tree will be divided into gender. We can get information from entropy like 1- Entropy.

**Reduction of Variables**

So far, we have discussed the algorithms for the categorical variables. Reduction of variance is an algorithm used for the constant variable in regression problems. This algorithm uses the standard change formula to choose the best partition. Separation with low variance is chosen as a criterion for distribution of a population.

$$Variance = \frac{\sum_{i=1}^{n}(X - \bar{X})^2}{n}$$

**Steps in calculating variance**

Calculate the variance for each division as the weighted average of each string of nodes. Let's set the numeric value 1 for those who play basketball and 0 for those who do not play basketball. Now we're following the steps to identify the right partition. So far, we have discussed the algorithms for the categorical variable.

1. The variance of the root node, the mean value is: (18 * 1 + 18 * 0) / 36 = 0.5 and in this case according to the above note we have 18 units and 18 zeros. And now the variance is: ((1-0.5) ^ 2 + (1-0.5) ^ 2 + ... .10 here + (0-0.5) ^ 2 + (0-0.5) ^ 2 + ... 8 here) / 36, which can be written: (18 * (1-0.55) ^ 2 + 18 * (0-0.5) ^ 2) / 36 = 0.25.
2. The average of the female node = (4 * 1 + 12 * 0) / 16 = 0.25 and Variance = (4 * (1-

0.25) ^ 2 + 12 * (0-0.25) ^ 2) / 16 = 0.19.
3. The median of the male node = (14 * 1 + 6 * 0) / 20 = 0.7 and Variance = (14 * (1-0.7) ^ 2 + 6 * (0-0.7) ^ 2) / 20 = 0.21.
4. Variance for Gender Distribution = Weighted Variance = (16/36) * 0.19 + (20/36) * 0.21 = 0.21.
5. Average for the class node XI = (6 * 1 + 10 * 0) / 16 = 0.375 and Variance = (6 * (1-0.375) ^ 2 + 10 * (0-0.375) ^ 2) / 16 = 0.23.
6. Masters for this class XII = (12 * 1 + 8 * 0) /20 = 0.6 and Variance = (12 * (1-0.6) ^ 2 + 8 * (0-0.6) ^ 2) / 20 = 0.24.
7. Variance for distribution class = (16/36) * 0.23 + (20/36) * 0.24 = 0.24.
From the above calculations we see that gender division has lower variance compared to the parent node, so the division will occur in the gender variable. Finally, we see that the use of four algorithms gives us the same conclusion on which node should be the distribution. This shows that all these algorithms are valid and result are the same.

## References

"Classification and Regression Trees (CART)." Electronic Textbook StatSoft.StatSoft, Inc., 2002. Web. 20 Apr. 2012. <http://www.obgyn.cam.ac.uk/cam-only/statsbook/stcart.html>.
Stine, R. (2011). "Lecture 8: Classification & Regression Trees." University of Pennsylvania Data Mining.Web. 19 Apr. 2012. <http://wwwstat.wharton.upenn.edu/~stine/mich/ DM08.pdf>.
"Lesson 10: Classification/Decision Trees ." File last modified on 2012. Penn State STAT 557 Data Mining. The Pennsylvania State University.Drupal.Web. 19 Apr. 2012. <https://onlinecourses.science.psu.edu/stat557/book/export/html/83>.
"Classification and Regression Trees (C&RT)." Electronic Textbook.StatSoft, Inc., 2002. Web. 20 Apr. 2012. <http://www.obgyn.cam.ac.uk/cam-only/statsbook/stcart.html>.
"Classification and Regression Trees (C&RT)." *Electronic Textbook*. StatSoft, Inc., 2002. Web. 20 Apr. 2012. <http://www.obgyn.cam.ac.uk/cam-only/statsbook/stcart.html>.
*CRAN R Project*. Vers. 3.1-52. N.p., Mar.-Apr. 2012. Web. 20 Apr. 2012. <http://cran.r-project.org/web/packages/rpart/rpart.pdf>.
https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/