Decision Trees in Machine Learning

An Introduction

Nikita Masand

Mentor: Prof. Pranav Nerurkar

Dept of Computer Engineering and IT, VJTI

Abstract

Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated data sets without imposing a complicated parametric structure. This poster is an introduction to the widely used Decision Trees.

Introduction

A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome(categorical or continues value). The whole idea is to create a tree for the entire data and process a single outcome at every leaf(or minimize the error in every leaf). 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.Each feature of the data set becomes a root[parent] node, and the leaf[child] nodes represent the outcomes. The decision on which feature to split on is made based on resultant entropy reduction or information gain from the split.

Splitting

Only input variables related to the target variable are used to split parent nodes into purer child nodes of the target variable. Both discrete input variables and continuous input variables (which are collapsed into two or more categories) can be used. When building the model one must first identify the most important input variables, and then split records at the root node and at subsequent internal nodes into two or more categories or bins based on the status of these variables. Characteristics that are related to the degree of purity of the resultant child nodes (i. e. , the proportion with the target condition) are used to choose between different potential input variables; these characteristics include entropy, Gini index, classification error, information gain, gain ratio, and twoing criteria.

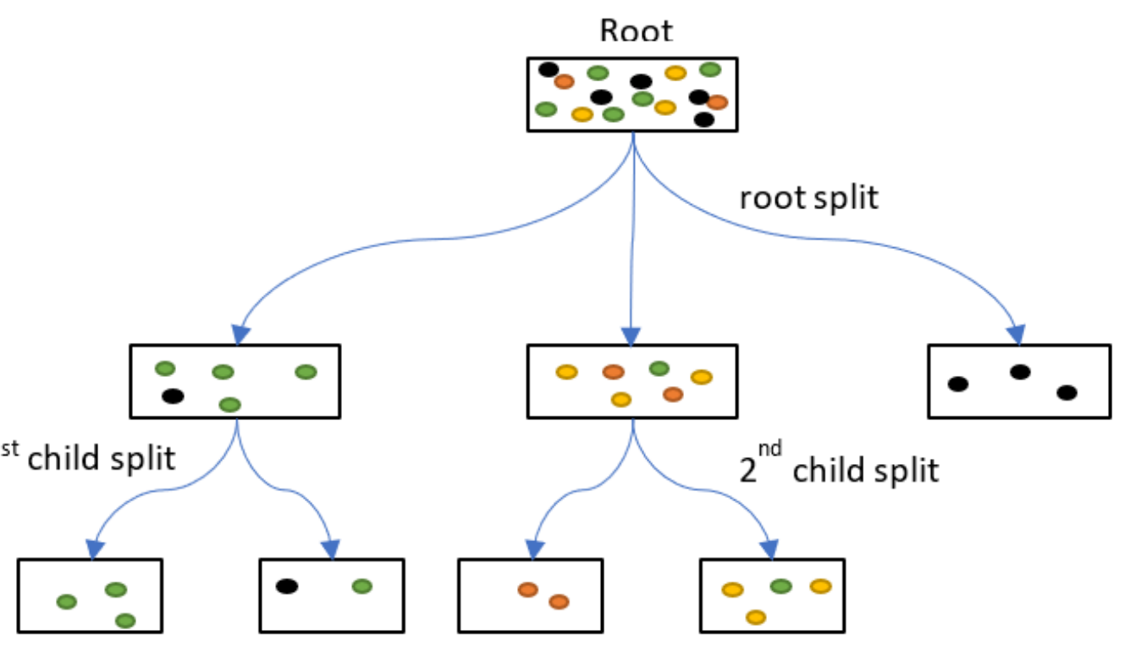


Figure 1: Splitting in a tree

Source:https://www.displayr.com/how-is-splitting-decided-for-decision-trees/

Characteristics

Entropy

- Entropy is the measure of impurity, disorder or uncertainty in a bunch of examples.
- Entropy controls how a Decision Tree decides to split the data. It actually effects how a Decision Tree draws its boundaries.

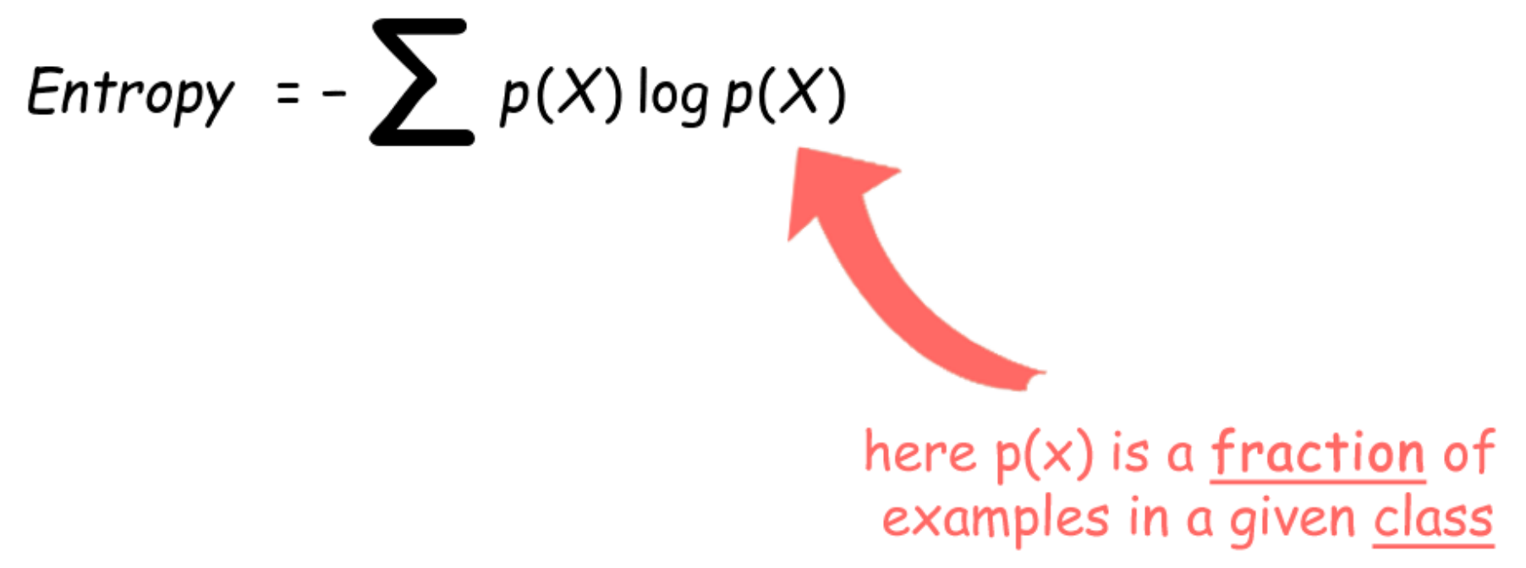
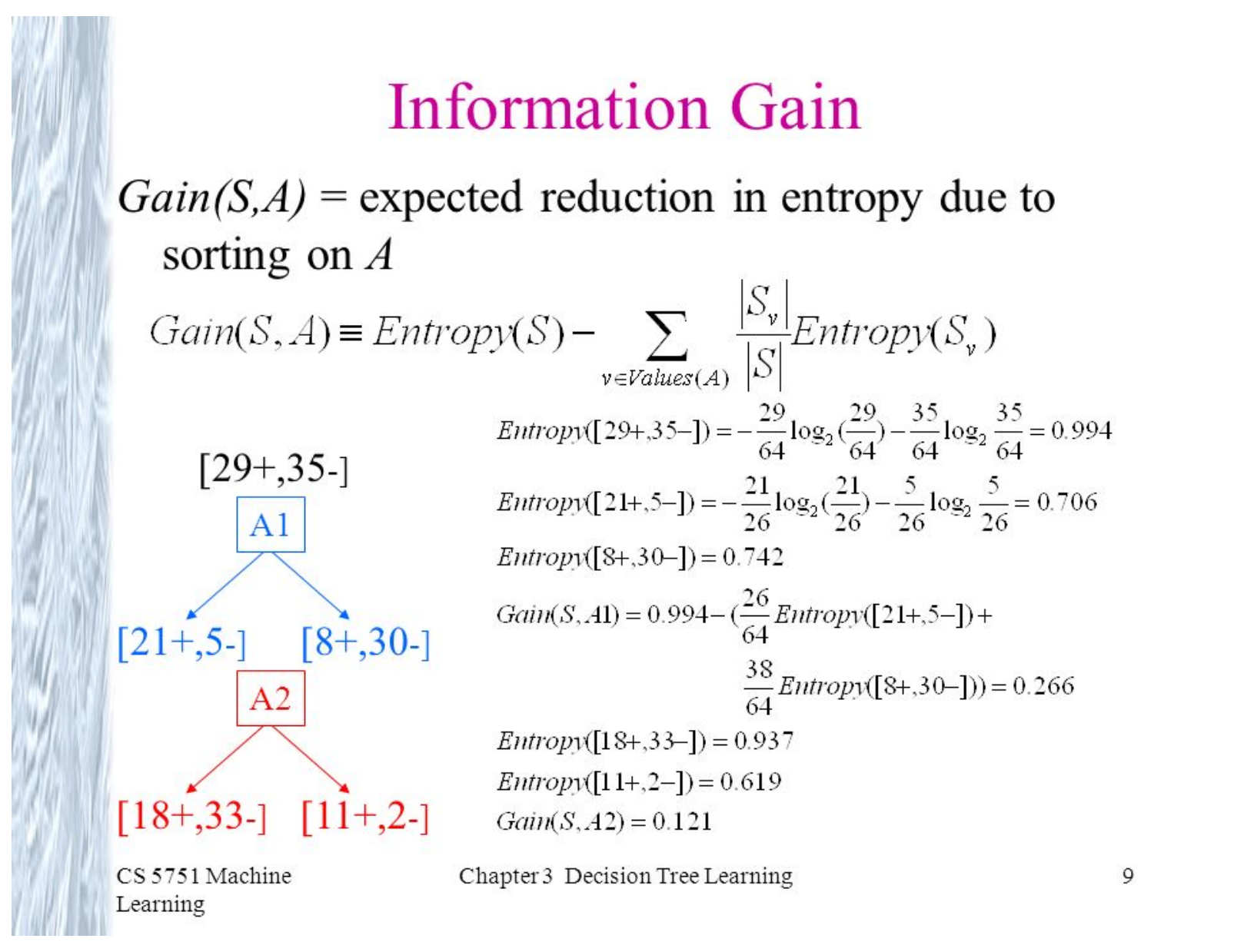

$Entropy = -\sum p(X)\log p(X)$
here p(x) is a fraction of examples in a given class

Figure 2: Entropy

Source:https://www.saedsayad.com/decision$_{tree.htm}$

Information gain

1. Information gain (IG) measures how much information a feature gives us about the class.
2. Information gain is the main key that is used by Decision Tree Algorithms to construct a Decision Tree.
3. Decision Trees algorithm will always tries to maximize Information gain.
4. An attribute with highest Information gain will tested/split first.



Information Gain

Gain(S,A) = expected reduction in entropy due to sorting on *A*

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy([29+,35-]) = -\frac{29}{64}\log_2(\frac{29}{64}) - \frac{35}{64}\log_2\frac{35}{64} = 0.994$$

$$Entropy([21+,5-]) = -\frac{21}{26}\log_2(\frac{21}{26}) - \frac{5}{26}\log_2\frac{5}{26} = 0.706$$

$$Entropy([8+,30-]) = 0.742$$

$$Gain(S,A1) = 0.994 - (\frac{26}{64}Entropy([21+,5-]) + \frac{38}{64}Entropy([8+,30-])) = 0.266$$

$$Entropy([18+,33-]) = 0.937$$

$$Entropy([11+,2-]) = 0.619$$

$$Gain(S,A2) = 0.121$$

CS 5751 Machine Learning Chapter 3 Decision Tree Learning 9

Figure 3: calculating information gain

Source:https://www.saedsayad.com/decisiontree.htm

Gini Index

1. Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower gini index should be preferred.
2. Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.
3. It performs only Binary splits Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses Gini method to create binary splits.

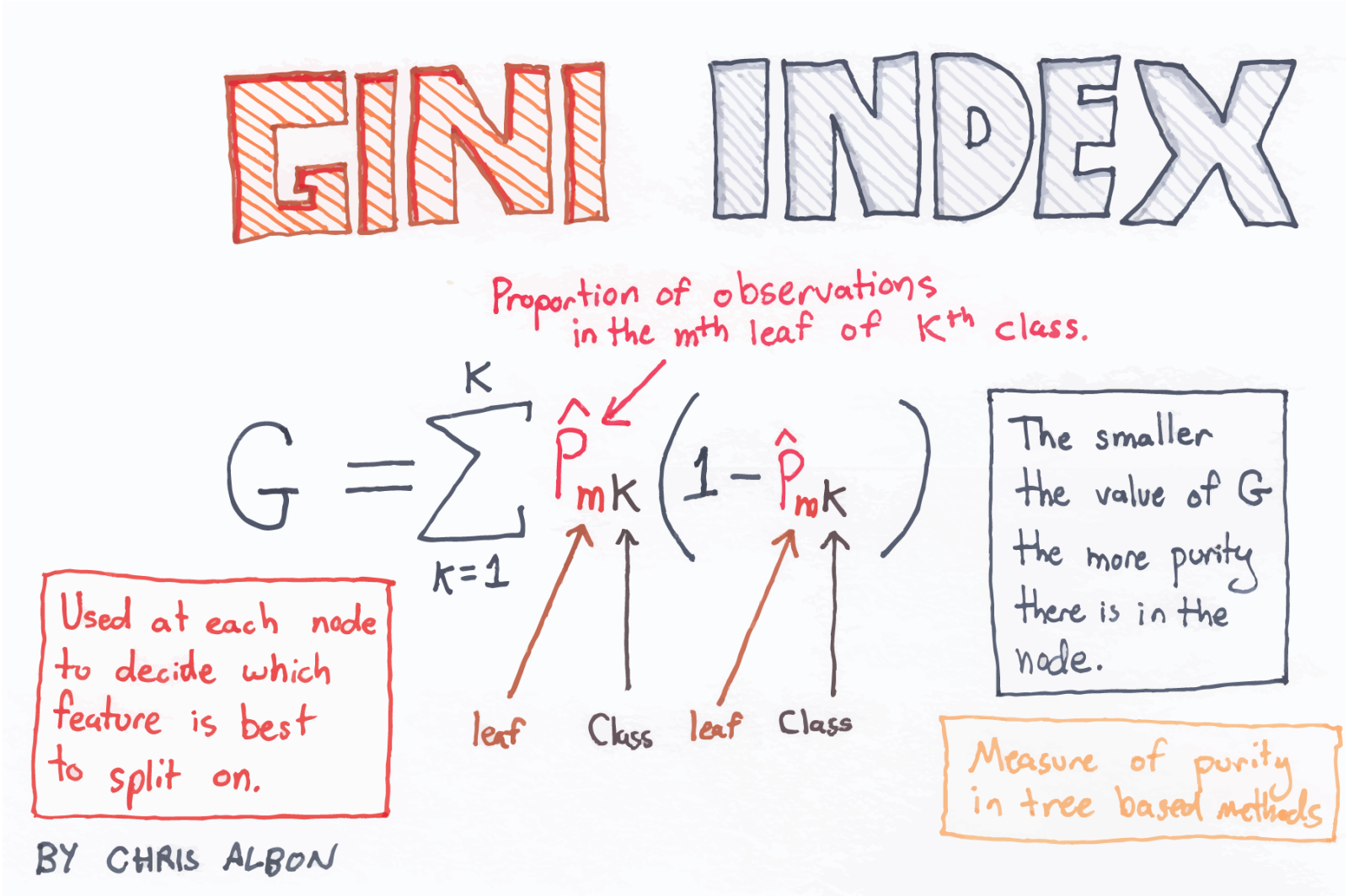

GINI INDEX
Proportion of observations in the mth leaf of Kth class.
$G = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk})$
Used at each node to decide which feature is best to split on.
leaf
Class
leaf
Class
The smaller the value of G the more purity there is in the node.
Measure of purity in tree based methods
BY CHRIS ALBON

Figure 4: Gini Index

Source:https://t4tutorials.com/gini-index-data-mining/

CART-Classification And Regression Tree

- The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be.
- The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.
- The main elements of CART are:
  1. Rules for splitting data at a node based on the value of one variable
  2. Stopping rules for deciding when a branch is terminal and can be split no more
  3. Finally, a prediction for the target variable in each terminal node.

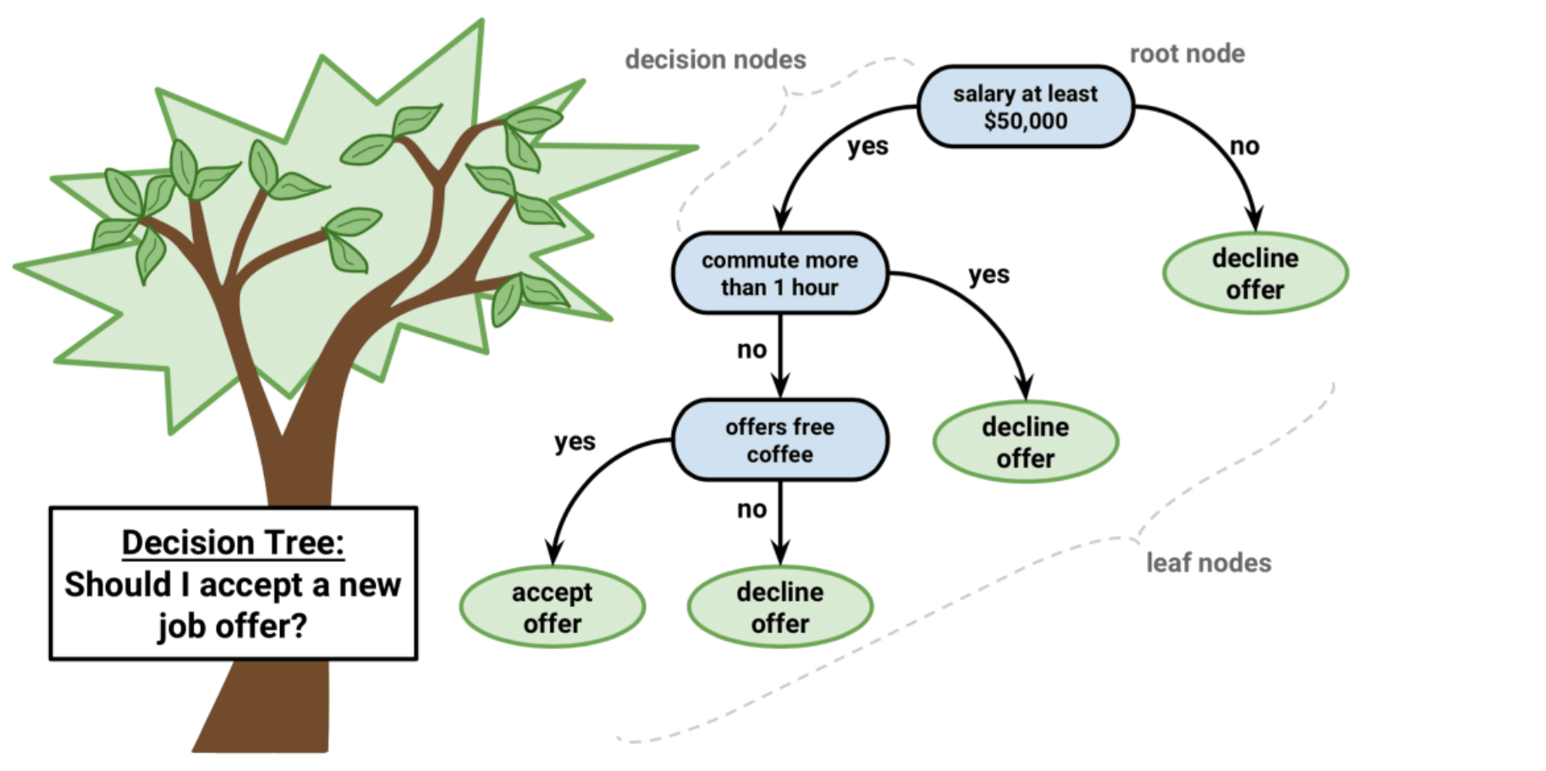


Figure 5: Decision tree CART

Source:https://machinelearningmastery.com/cart-for-machine-learning/

Pruning

1. As the name implies, pruning involves cutting back the tree.
2. After a tree has been built it may be overfitted. The CART algorithm will repeatedly partition data into smaller and smaller subsets until those final subsets are homogeneous in terms of the outcome variable.

Figure 6: pruning

Source:https://www.finegardening.com/pruning-tips-and-techniques

Implementation

Many data mining software packages provide implementations of one or more decision tree algorithms. Examples include Salford Systems CART (which licensed the proprietary code of the original CART authors),[3] IBM SPSS Modeler, RapidMiner, SAS Enterprise Miner, Matlab, R (an open-source software environment for statistical computing, which includes several CART implementations such as rpart, party and randomForest packages), Weka (a free and open-source data-mining suite, contains many decision tree algorithms), Orange, KNIME, Microsoft SQL Server [1], and scikit-learn (a free and open-source machine learning library for the Python programming language).

Forthcoming Research

In a decision tree, all paths from the root node to the leaf node proceed by way of conjunction, or AND. In a decision graph, it is possible to use disjunctions (ORs) to join two more paths together using minimum message length (MML). Decision graphs have been further extended to allow for previously unstated new attributes to be learnt dynamically and used at different places within the graph.[

References

[1] ml-decision-tree.

[2] Rishabh Jain. Decision trees. *decision-trees-it-begins-here*.

[3] Renu Khandelwal. Decision tree and random forest. *datadriveninvestor*.

[4] Madhu Sanjeevi. Decision trees algorithms. *deep-math-machine-learning-ai*.

[5] SeattleDataGuy. What is a decision tree algorithm?

[6] Chirag Sehra. Decision trees explained easily.

[7] Nasir Islam Sujan. What is entropy and why information gain matter in decision trees? *coinmonks*.