Supervised Machine learning technique using decision trees

Group-5

Name - Kirath Singh | Shashwat Arya

Student Number – 9082129405 | 9082094328

Email id - ksingh52@wisc.edu | sarya3@wisc.edu

Abstract

The aim of this project is to conceptualise the decision tree algorithm for supervised machine learning (clustering and regression) as well as understanding different mechanisms to prevent overfitting of the decision tree models.

Learning objectives

1. The concept of Gini impurity used in decision tree- Gini impurity is used to measure the degree or probability variable being incorrectly classified when it is randomly chosen .For ex. Gini impurity of value 0 means sample are perfectly homogeneous and all element are similar, whereas, Gini impurity of value 1 means maximal inequality among elements.

2. The mathematics behind the classification tree algorithm for supervised machine learning (classification and regression)-the use Gini index as a cost function in order to evaluate split in feature selection for building classification tree.

Background

For a bank to consider whether to offer someone a loan they often go through a sequential list of questions to figure out if it is safe to give said loan to an individual. Those questions can start as simple as what kind of income does the person have? If it is between $30–70k they move on to the next question. How long have they held their current job? If 1–5 years it leads to their next question of do, they make their credit card payments? If yes, then they offer the Loan and if no they do not. This process at its most basic form is a Decision Tree. A decision tree is a largely used non-parametric effective machine learning modelling technique for regression and classification problems. To find solutions a decision tree makes sequential, hierarchical decision about the outcome's variable based on the predictor data.

Decision tree models where the target variable uses a discrete set of values are classified as Classification Trees. In these trees, each node, or leaf, represent class labels while the branches represent conjunctions of features leading to class labels. A decision tree where the target variable takes a continuous value, usually numbers, are called Regression Trees. The two types are commonly referred to together at CART (Classification and Regression Tree).

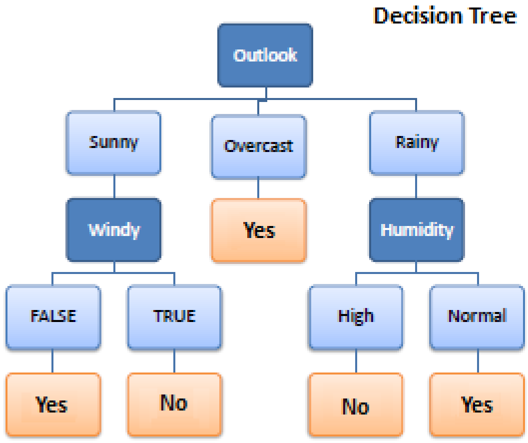
Each CART model is a case of a Directed Acyclic Graph. These graphs have nodes representing decision points about the main variable given the predictor and edges are the connections between the nodes. In the Loan scenario above the $30-$7ok would be an edge and the "Years Present in Job" are nodes.

As the goal of a decision tree is that it makes the optimal choice at the end of each node it needs an algorithm that can do just that. That algorithm is known as Hunt's algorithm, which is both greedy, and recursive. Greedy meaning that at step it makes the most optimal decision and recursive meaning it splits the larger question into smaller questions and resolves them the same way. The decision to split at each node is made according to the metric called **purity**. A node is 100% impure when a node is split evenly 50/50 and 100% pure when all its data belongs to a single class.

In order to optimize our model, we need to reach maximum purity and avoid impurity. To measure this, we use the Gini impurity, which measures how often a randomly chosen element is labelled incorrectly if it was randomly labelled according to distribution. It is calculated by adding the probability, pi, of an item with the label, i, being chosen multiplied by the times the probability (1–pi) of a mistake categorizing the time. Our goal is to have it reach 0 where it will be minimally impure and maximally pure falling into one category. [1]

Now let us apply our knowledge about decision trees on a dataset: - [2]

Outlook	Temp.	Humidity	Windy	Play Golf
Predictors				Target
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Consider the above dataset, the decision tree for the data has been constructed on the left of the dataset.

Camlin Page
Date / /

Entropy – If the sample is completely homogeneous the entropy is 0, if the sample is equally divided, then its Entropy is 1

In order to build a decision tree We need to calculate 2 types of entropy

a) Entropy using frequency table of 1 attribute

No. of samples = 1

Samples for which play golf is 'Yes' = 9

Samples for which play golf is 'No' = 5

Hence, $E(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$, ~~Entropy = -0.5 log p~~ $\text{Entropy} = -p\log_2 p - q\log_2 q$

$\text{Entropy}(5,9) = \text{Entropy}(0.36, 0.64)$

$= -(0.36\log_2(0.36)) - (0.64\log_2(0.64))$

$= 0.94$

b) Entropy Using frequency table of 2 attributes

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

E(Play Golf, Outlook) = P(Sunny) * E(Sunny Yes, Sunny No) (PlayGolf, Play Golf)
+
P(Overcast) * E(Play Golf Yes, No)
+
P(Rainy) * E(Play Golf Yes, No)

Camlin Page
Date / /

$= (5/14) * 0.971 + (4/14) * 0.0 + (5/14) * 0.971$

$= 0.693$

Information Gain - It is based on the decrease in entropy after a dataset is split on an attribute

Entropy (Play Golf) = 0.94

Entropy (Play Golf, Outlook) = 0.693

Information Gain

$= 0.94 - 0.693$

$= 0.247$

We usually choose attribute with the largest information gain as the decision node, divide the dataset, repeat the same process

Calculation of Gini index/Gini impurity

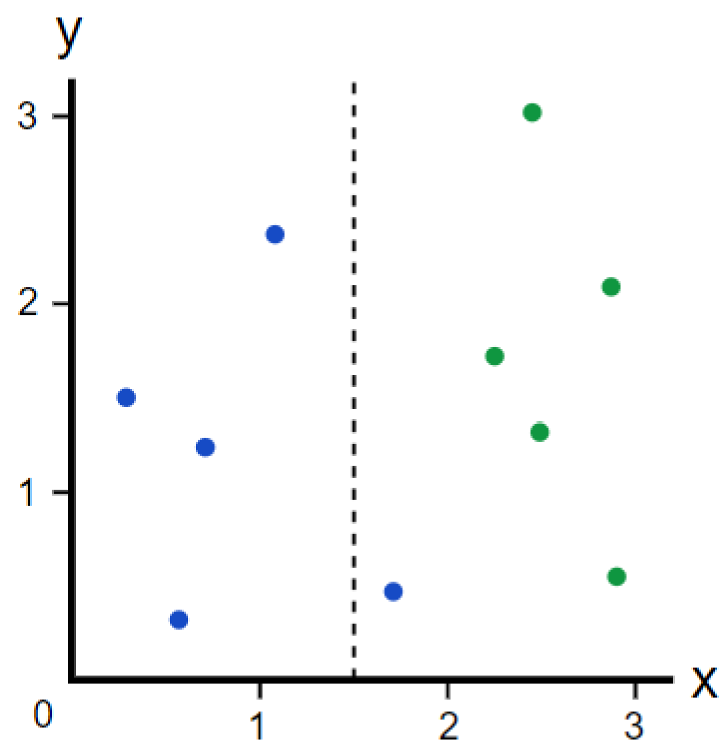
Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

1. It works with categorical target variable "Success" or "Failure".
2. It performs only Binary splits
3. Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses Gini method to create binary splits.

Steps to calculate Gini index/impurity

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split

Example: -



An Imperfect Split

Consider the above graph containing equal number of green and blue points. Let us make a split at x=1.5.

This imperfect split breaks our dataset into these branches: -

- Left branch, with 4 blues.
- Right branch, with 1 blue and 5 greens. It's obvious that this split is worse, but **how can we quantify that?**

This is where Gini impurity comes into picture.

Let's calculate the Gini Impurity of our entire dataset. If we randomly pick a datapoint, it's either blue (50%) or green (50%). Now, we randomly classify our datapoint according to the class distribution. Since we have 5 of each colour, we classify it as blue 50% of the time and as green 50% of the time. [3]

What's the probability we classify our datapoint **incorrectly**?

Event	Probability
Pick Blue, Classify Blue ✓	25%
Pick Blue, Classify Green ✗	25%
Pick Green, Classify Blue ✗	25%
Pick Green, Classify Green ✓	25%

We only classify it incorrectly in 2 of the events above. Thus, our total probability is 25% + 25% = 50%, so the Gini Impurity is $\boxed{0.5}$.

Gini index

B

Total classes = C

Probability of picking a datapoint of class i = p(i)

Gini impurity $(G) = \sum_{i=1}^{C} p(i) * (1-p(i))$

$G = p(1) * (1-p(1)) + p(2) * (1-p(2))$

$= 0.5 * (1-0.5) + 0.5 * (1-0.5)$

Now that we have studied about the decision tree model, let us test our knowledge with the help of a short quiz, comprising 5 questions.

Quiz

Question 1. Decision tree model comes under which type of machine learning.

A) Unsupervised machine learning

B) Supervised machine learning

C) Both A and B

D) It is not a machine learning model

Answer - Option B

Question 2 Select the correct options.

A) The information gain is based on the decrease in entropy after a dataset is split on an attribute.

B) Decision tree is the same as binary tree

C) Entropy is calculated with the help of a frequency table

D) Decision tree is used in clustering

Answer – Options A, C

Question 3 What is true about Entropy in decision tree model?

A) If the sample is homogenous, the entropy is zero

B) If the sample is homogenous, the entropy is one

C) If the sample is equally divided, the entropy is zero

D) If the sample is equally divided, the entropy is one

Answer – Options A, D

Question 4 What is a decision node?

A) It is just an ordinary node

B) When a sub-node splits into further sub-nodes, then it is called decision node.

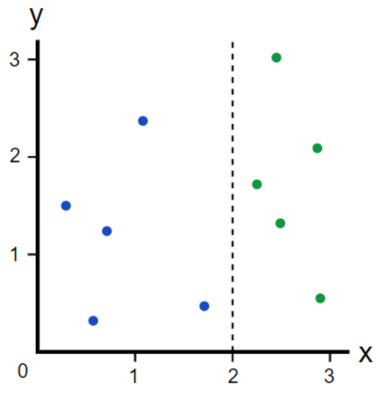
C) The node that makes a decision

D) The last node of the decision tree

Answer – Option B

Activity Based on the Lecture

Question 1 – Refer to the plot given below and answer the following questions related to the plot.



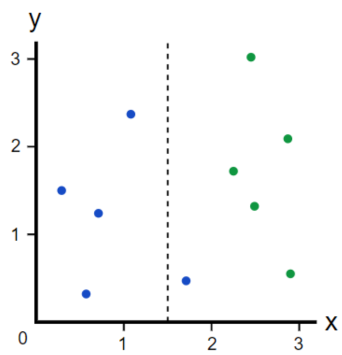
A Perfect Split

a) Calculate the Gini impurity for the left branch (i.e. only blue points in the plot)

b) Calculate the Gini impurity for the right branch (i.e. only green points in the plot)

c) Comment on the Gini impurity values obtained in parts a) and b)

Question 2- Refer to the plot given below and answer the following questions related to the plot.



An Imperfect Split

a) Calculate the Gini impurity for the left branch (i.e. only blue points in the plot)

b) Calculate the Gini impurity for the right branch (i.e. only green points in the plot)

c) Comment on the Gini impurity values obtained in parts a) and b)

d) Refer to the Gini impurity value for the dataset (before the split) given in the lecture notes and calculate the quality of the split by weighting the impurity of each branch by the total number of elements it has.

e) Calculate the total amount of impurity removed in the split also known as Gini gain.

f) Comment on the new Gini value obtained above and compare it with the Gini value calculated before the split.

Activity Solutions

Page No.
Date : / /201

Question 1

a) $\because$ The left branch has only blue points

$\therefore$ The Gini impurity is

$$G_{left} = 1 * (1-1) + 0 * (1-0)$$
$$= 0 + 0$$
$$= 0$$

b) The Right branch has only green points

$\therefore$ The Gini impurity is

$$G_{right} = 0 * (1-0) + 1 * (1-1)$$
$$= 0 + 0$$
$$= 0$$

c) Both the branches (i.e. Left & Right) have 0 impurity.
That indicates that our splitting of the dataset was perfect as it divided the dataset into 2 branches with 0 impurity

Page No.
Date : / /201

Question 2

a) The Left branch has only blues

Hence reviewing the results calculated earlier

$G_{left} = 0$

b) The Right branch has 1 blue & 5 greens

$$\therefore G_{right} = \frac{1}{6} * \left(1 - \frac{1}{6}\right) + \frac{5}{6} * \left(1 - \frac{5}{6}\right)$$

$$= \frac{5}{18}$$

$$= 0.278$$

c) Considering the impurities of both left & right branches; it is clearly observable that:-

i) The Left branch does a perfect job as it has 0 impurity in classifying datapoints

ii) However due to the presence of 1 blue datapoint in the ~~left~~ right branch, the right split will have some inaccuracy while classifying datapoints.

Page No.
Date : / /201

Question 2
d)

Gini impurity for the dataset before the split (given in Lecture notes) = 0.5

$Gini_{Left} = 0$

$Gini_{Right} = 0.278$

Points in the Left branch = 4

Points in the Right branch = 6

$\therefore$ Quality of Split = $(0.4 * 0) + (0.6 * 0.278)$

$= 0.167$

e) Gini impurity for the dataset before the split (given in Lecture notes) = 0.5

Gini impurity for the dataset after the split (calculated in part d) = ~~0.167~~ 0.167

$\therefore$ Total amount of impurity removed with the split = $0.5 - 0.167$

$= 0.333$

Page No.
Date : / /201

Question 2

b) After reviewing the Gini impurity obtained before and after splitting the dataset into left and right branches, it is clearly visible that the Gini impurity was reduced significantly from 0.5 to (0.5 − 0.167) = 0.333.

$\therefore$ 0.5 > 0.333

We were able to lower the overall gini impurity of the given dataset by splitting it into left & right branches respectively

Citations

1. Plapinger Thomas, what is decision tree? 30th Jul. 2017, Accessed on 26th Apr. 2020 . [Online]. Available: https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1

2. Jain Rishabh, Decision Tree, it begins here, 21st Mar 2017, Accessed on 26th Apr. 2020. [Online]. Available: https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134

3. Zhou Victor, A simple explanation on Gini index, 29th Mar 2019, Accessed on 26th Apr. 2020. [Online]. Available: https://victorzhou.com/blog/gini-impurity/