ASSIGNMENT 2

Decision Tree Using Gini Index

Machine Intelligence and Expert System (IT-5213)

Submitted By:

Name	Roll	M.Sc. Session
1. *Tajim Md. Niamat Ullah Akhund*	1120	2017-2018
2. *M. Mahfuzul Haq*	1124	
3. *Priyanka Dhar*	1074	
4. *Rafia Akther*	1086	
5. *Kazi Zannath Nowshin*	1077	

Submitted To:

Dr. Shamim Al Mamun
Associate Professor,
Institute of Information Technology,
Jahangirnagar University.

Institute of Information Technology,
Jahangirnagar University,
Savar, Dhaka-1342, Bangladesh.

Decision Tree Using Gini Index

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each *internal node of the tree corresponds to an attribute*, and each *leaf node corresponds to a class label*.

The primary challenge in the decision tree implementation is to identify which attributes do we need to consider as the root node and each level. Handling this is know the attributes selection. We have different attributes selection measure to identify the attribute which can be considered as the root note at each level.

The popular attribute selection measures:

- Information gain
- Gini index

Gini Index

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*.

Example: Construct a Decision Tree by using "Gini index" as a criterion

Our dataset is as follows:

Serial No.	A	B	C	D	E
1.	4.8	3.4	1.9	0.2	POSITIVE
2.	5	3	1.6	0.2	POSITIVE
3.	5	3.4	1.6	0.4	POSITIVE
4.	5.2	3.5	1.5	0.2	POSITIVE
5.	5.2	3.4	1.4	0.2	POSITIVE
6.	4.7	3.2	1.6	0.2	POSITIVE
7.	4.8	3.1	1.6	0.2	POSITIVE
8.	5.4	3.4	1.5	0.4	POSITIVE
9.	7	3.2	4.7	1.4	NEGATIVE
10.	6.4	3.2	4.5	1.5	NEGATIVE
11.	6.9	3.1	4.9	1.5	NEGATIVE
12.	5.5	2.3	4	1.3	NEGATIVE

13.	6.5	2.8	4.6	1.5	NEGATIVE
14.	5.7	2.8	4.5	1.3	NEGATIVE
15.	6.3	3.3	4.7	1.6	NEGATIVE
16.	4.9	2.4	3.3	1	NEGATIVE

Here, we have 5 columns out of which 4 columns have continuous data and 5th column consists of class labels.

A, B, C, D attributes can be considered as predictors and E column class labels can be considered as a target variable. For constructing a decision tree from this data, we have to convert continuous data into categorical data.

We have chosen some random values to categorize each attribute:

A	B	C	D
>= 5	>= 3.0	>=4.2	>= 1.4
< 5	< 3.0	< 4.2	< 1.4

Gini Index for Var A

Var A has value >=5 for 12 records out of 16 and 4 records with value <5 value.

- For Var A >= 5 & class == positive: 5/12
- For Var A >= 5 & class == negative: 7/12
    - Gini (5,7) = 1- $\{(5/12)^2 + (7/12)^2\}$ = 0.4860
- For Var A <5 & class == positive: 3/4
- For Var A <5 & class == negative: 1/4
    - Gini (3,1) = 1- $\{(3/4)^2 + (1/4)^2\}$ = 0.375

By adding weight and sum each of the Gini indices:

Gini (Target, A) = (12/16) * (.486) + (4/16) * (.375) = **0.45825**

Gini Index for Var B

Var B has value >=3 for 12 records out of 16 and 4 records with value <5 value.

- For Var B >= 3 & class == positive: 8/12
- For Var B >= 3 & class == negative: 4/12
    - Gini (8,4) = 1- $\{(8/12)^2 + (4/12)^2\}$ = 0.446
- For Var B <3 & class == positive: 0/4

- For Var B <3 & class == negative: 4/4
    - Gini (0,4) = 1- $\{(0/4)^2 + (4/4)^2\} = 0$

Gini (Target, B) = (12/16) * 0.446 + (4/16) * 0 = **0.3345**

Gini Index for Var C

Var C has value >=4.2 for 6 records out of 16 and 10 records with value <4.2 value.

- For Var C >= 4.2 & class == positive: 0/6
- For Var C >= 4.2 & class == negative: 6/6
    - Gini (0,6) = 1- $\{(0/8)^2 + (6/6)^2\} = 0$
- For Var C < 4.2& class == positive: 8/10
- For Var C < 4.2 & class == negative: 2/10
    - Gini (8,2) = 1- $\{(8/10)^2 + (2/10)^2\} = 0.32$

Gini (Target, C) = (6/16) * 0+ (10/16) * 0.32 = **0.2**

Gini Index for Var D

Var D has value >=1.4 for 5 records out of 16 and 11 records with value <1.4 value.

- For Var D >= 1.4 & class == positive: 0/5
- For Var D >= 1.4 & class == negative: 5/5
    - Gini (0,5) = 1- $\{(0/5)^2 + (5/5)^2\} = 0$
- For Var D < 1.4 & class == positive: 8/11
- For Var D < 1.4 & class == negative: 3/11
    - Gini (8,3) = 1- $\{(8/11)^2 + (3/11)^2\} = 0.397$

Gini (Target, D) = (5/16) * 0+ (11/16) * 0.397 = **0.273**

Drawing the tree is as follows:

		wTarget	
		Positive	Negative
A	>. 5.0	5	
	<5	3	1
Gini Index of A = 0.45825			

		Target	
		Positive	Negative
B	>. 3.0	8	4
	< 3.0	0	4
Gini Index of B = 0.3345			

		Target	
		Positive	Negative
C	>= 4.2	0	6
	< 4.2	8	
Gini Index of C= 0.2			

		Target	
		Positive	Negative
D	>= 1.4	0	5
	< 1.4	8	3
Gini Index of D= 0.273