POORNIMA FOUNDATION

Estd. 2000

LECTURE NOTES

Campus: PCE Course: BTECH in CSE Class/Section: III Yr. Section- A Date: 24-02-24

Name of Faculty: Praveen Kumar Yadav Name of Subject: Machine Learning Code: 6CS4-02

Date (Prep.): 24-02-24 Date (Del.): 5-03-24 Unit No.: 4 Lect. No: 07

OBJECTIVE: To be written before taking the lecture (Pl. write in bullet points the main topics/concepts etc., which will be taught in this lecture)

Gini Impurity for a Decision Tree

IMPORTANT & RELEVANT QUESTIONS:

what is Gini Impurity for a decision Tree?

FEED BACK QUESTIONS (AFTER 20 MINUTES):

what are the factors to determine the decision Tree?

OUTCOME OF THE DELIVERED LECTURE: To be written after taking the lecture (Pl. write in bullet points about students' feedback on this lecture, level of understanding of this lecture by students etc.)

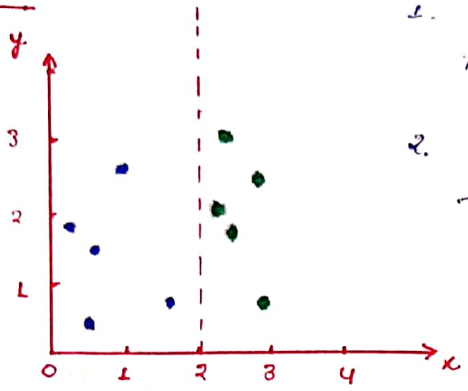
goal.

REFERENCES: Text/Ref. Book with Page No. and relevant Internet Websites:

scikit with ML.

Campus: PCE. Course: BTECH
Name of Faculty: Praveen Kumar Yadav
Class/Section: IV CSE-A
Name of Subject: ML
Date: 29-02-21
Code: 6CS4-02

Gini Impurity:-



1. Randomly pick a datapoint in our dataset, then
2. Randomly classify it according to class distribution in the dataset.

(what is the probability we classify the data point Incorrectly?)

The answer to that question is the Gini Impurity.

For eg-

Event	Probability	
Pick blue classify Blue	25%	
Pick blue classify Green	25%	Here we classify incorrectly in 2 events.
Pick Green " Blue	25%	
Pick Green " Green	25%	

So our total probability is - 25% + 25% = 50%.

Gini Impurity = 0.5

So Gini Impurity is -

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

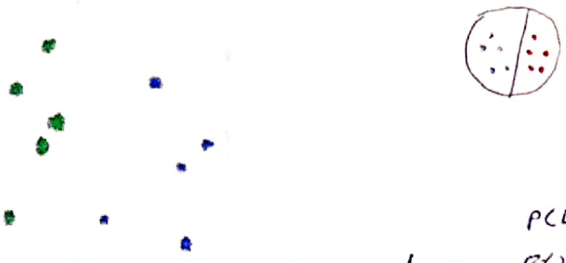
so Here we have $C = 2$, $p(1) = 0.5$

$p(2) = 0.5$

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

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

$$G = 0.5$$



Now calculate Gini Impurity After split in two branches.

$p(1) = 1$
$p(2) = 0$

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

$p(1) = 0$
$p(2) = 1$

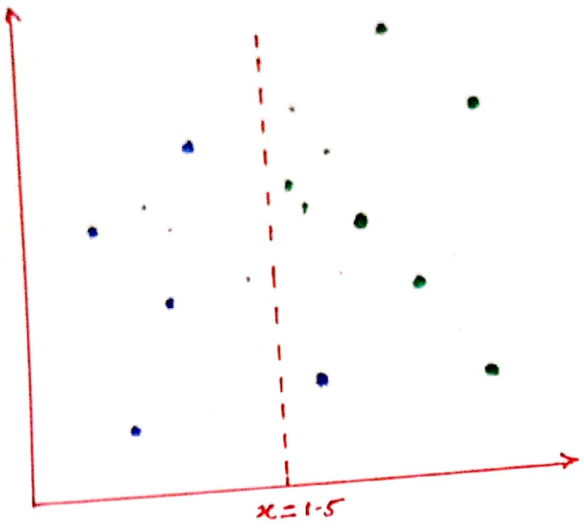
$$G_{right} = 0 * (1 - 0) + 1 * (1 - 1) = 0$$

so perfect split turned a dataset with 0.5 impurity into 2 branches with 0 impurity.

So A Gini Impurity of 0 is the lowest and best possible Impurity. It can only be achieved when everything is from the same class.

Campus: PCE. Course: BTECH | Class/Section: CSE-A | Date: 22-04-21
Name of Faculty: Praveen Kumar Yadav | Name of Subject: ML | Code: BCS-056-01

Imperfect split :-



Here Left branch has only blue so

$G_{left} = 0$

Right branch has 1 blue and 5 greens, so

$G_{right} = 1/6 \times (1 - 1/6) + 5/6 (1 - 5/6)$

$= 5/18 = 0.278$

we have already calculated the Gini Impurity for -

Before split (the entire data set) = 0.5

Left branch : 0

Right " : 0.278

so determine the quality of split by weighting the Impurity of each branch by how many element it has -

i.e (0.4 * 0) + (0.6 * 0.278) = 0.167

So Amount of Impurity we have "removed" with the split is - 0.5 - 0.167 = 0.333

↑ Gini Gain.

This is what's used to pick the best split in decision tree.

Higher Gini Gain = Better split.

so **Gini Impurity** - is the probability of incorrectly classifying a randomly chosen element in the dataset. if it were randomly labeled according to the class distribution in the data set.

Campus: PCE. Course: BTECH
Name of Faculty: Praveen Kumar Yadav
Class/Section: VII CSE - A
Name of Subject: ML
Date: 24-02-20
Code: CS4-02

when training a decision tree, the best split is chosen by "Maximizing the Gini gain" which is calculated by subtracting the weighted impurities of the branches from the original impurity.