Introduction

21 July 2023 10:16

ml algo

- Non-Parametric -> Non linear data
- White-box model
- Mother of all Tree based Algos
- Works with both Classification and Regression

Tree

linear model

DT

RF

Adaboost

GradientB

xgboost

Decision Tree

$\beta_0 \; \beta_1 \; \beta_2$

White-box

explainable

inference

Black Box

NN

Chatgpt

Intuition behind DT

20 July 2023 16:21

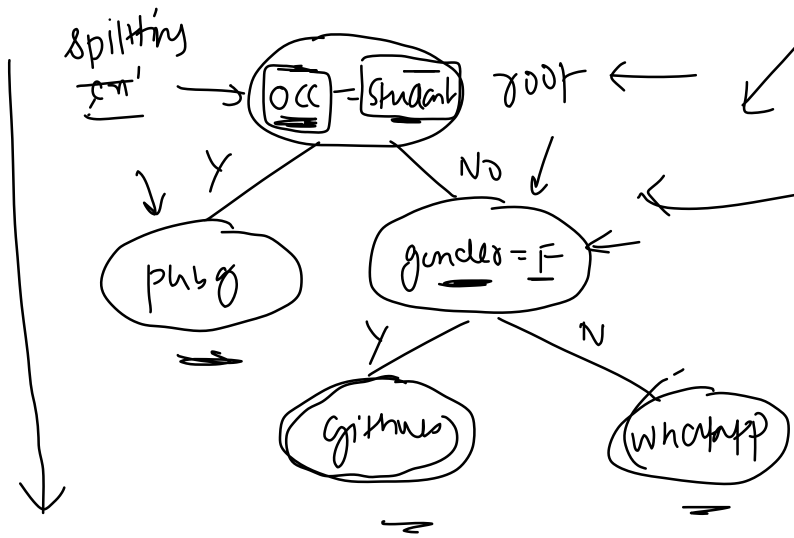
{m, prog}{F, student}

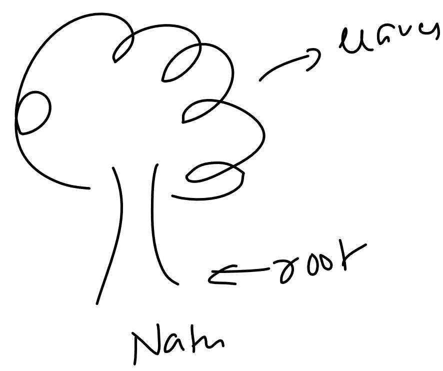
Gender	Occupation	Suggestion
F	Student	PUBG
F	Programmer	Github
M	Programmer	Whatsapp
F	Programmer	Github
M	Student	PUBG
M	Student	PUBG

giant Nested if-else structure

```
if occupation = student
    print pubg
else
    if gender = F
        print github
    else
        print whatsapp
```

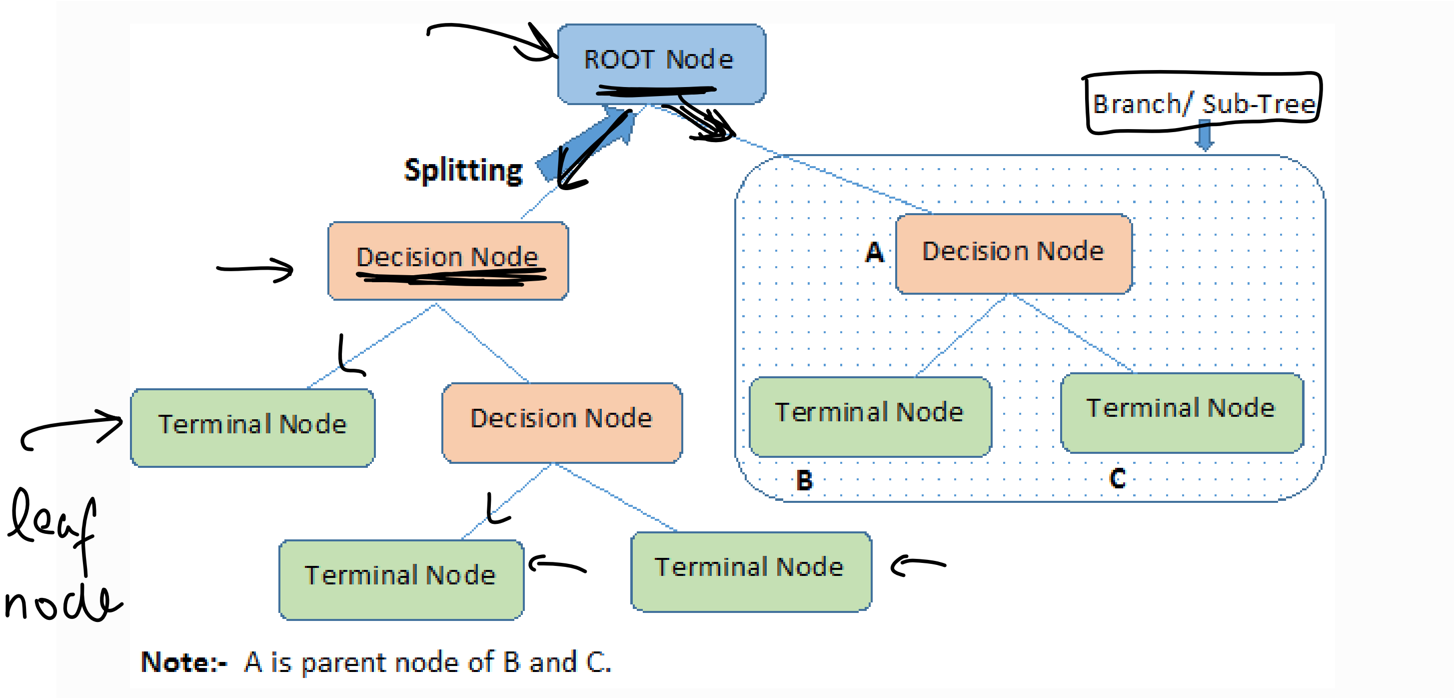
basic dt





Vocab

20 July 2023 16:22



Example 2

20 July 2023 16:25

Input output

	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed
7	Postgraduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed
10	Undergraduate	Arts	91	Unemployed
11	Postgraduate	Arts	96	Employed
12	PhD	Science	87	Employed
13	Undergraduate	Science	90	Unemployed
14	Postgraduate	Science	95	Employed

{UG, Science, 90}

15 rows

→ 1.5 lakh

ID3 C4.5 $\boxed{\text{CART}}$

sklearn

The CART Algorithm - Classification

21 July 2023 10:26

sklearn ?

Given training vectors $x_i \in R^n$, i=1,..., l and a label vector $y \in R^l$, a decision tree recursively partitions the feature space such that the samples with the same labels or similar target values are grouped together.

Let the data at node m be represented by Q_m with n_m samples. For each candidate split $\boxed{\theta} = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ subsets

$$Q_m^{left}(\theta) = \{(x, y) | x_j \le t_m\}$$
$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta)$$

The quality of a candidate split of node m is then computed using an impurity function or loss function $H()$, the choice of which depends on the task being solved (classification or regression)

$$G(Q_m, \theta) = \boxed{\frac{n_m^{left}}{n_m}} H(Q_m^{left}(\theta)) + \boxed{\frac{n_m^{right}}{n_m}} H(Q_m^{right}(\theta))$$

Select the parameters that minimises the impurity

$$\theta^* = \operatorname{argmin}_\theta G(Q_m, \theta)$$

Recurse for subsets $Q_m^{left}(\theta^*)$ and $Q_m^{right}(\theta^*)$ until the maximum allowable depth is reached, $n_m < \min_{samples}$ or $n_m = 1$.

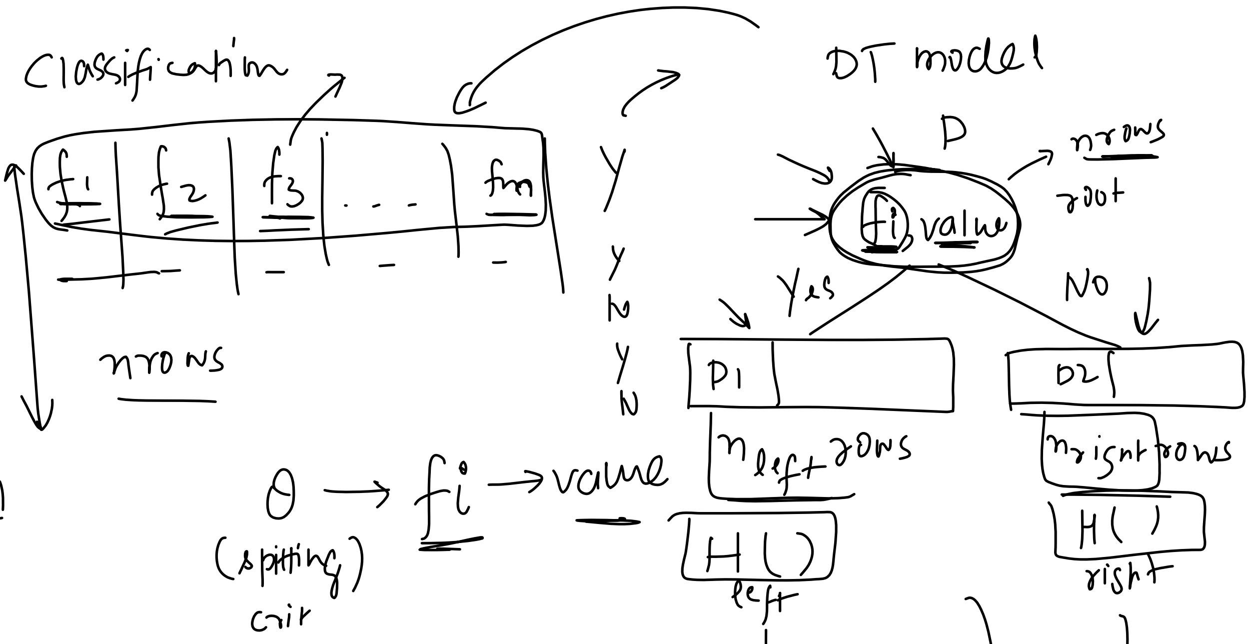
CART → Classification and Regression Trees

→ Algorithm to grow decision tree

→ sklearn

Classi → Regression

Classification



n×m

$\theta \rightarrow fi \rightarrow$ value

(spitting) crit

G( ) overall impurity $= \frac{n_{left}}{n} H_{left}() + \frac{n_{right}}{n} H()_{right}$

(G) impurity value

impurity function H()

classi — gini impurity

Reg — mse

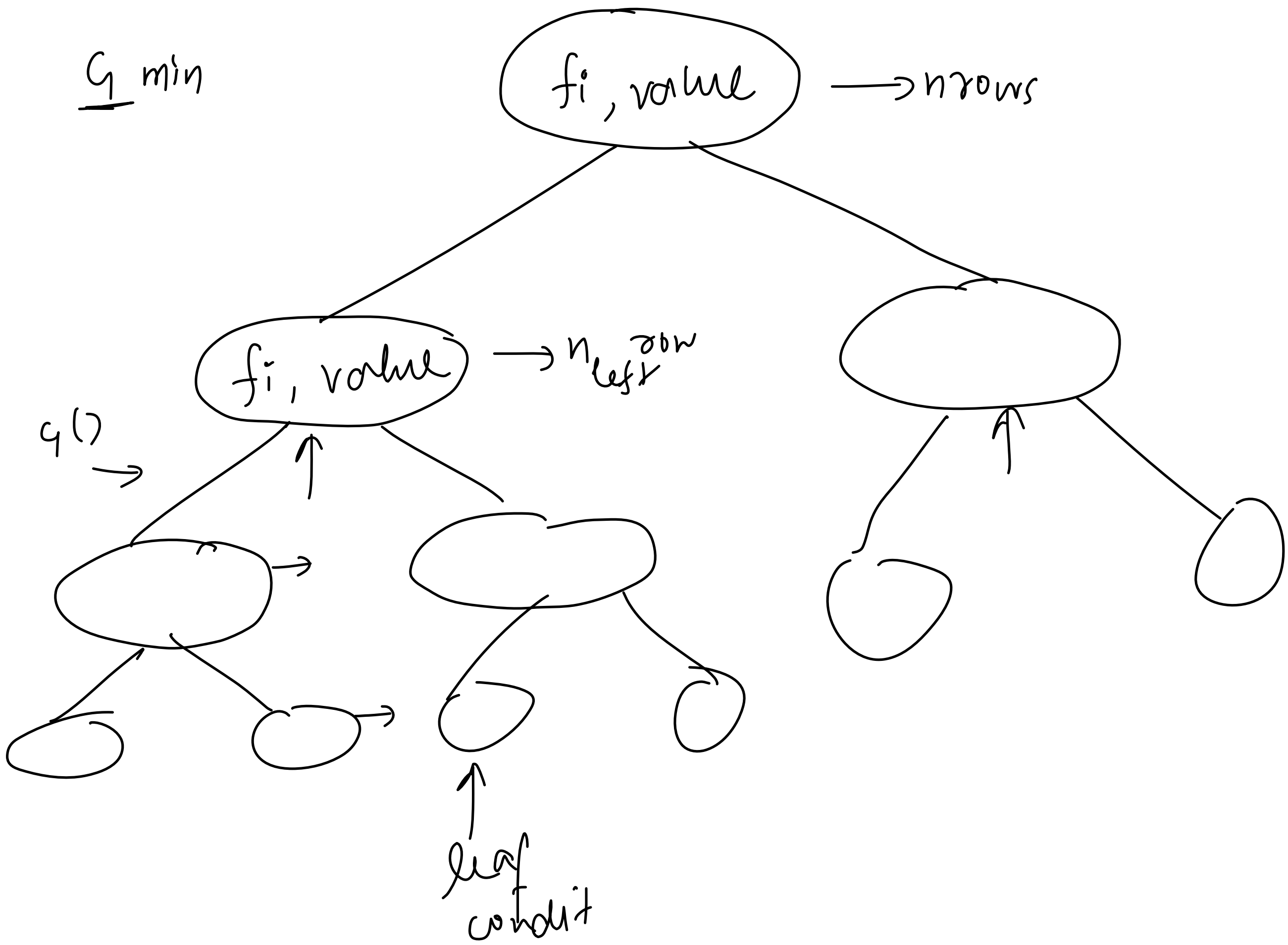
classi
- gini impurity
- entropy

- mse
- mae

(G) impurity valu

~~fi, value,~~ → G() ← (min)
f2, value, → G()
. . . → G()

G min



Splitting Categorical Features

21 July 2023 17:45

(G)

f1 ↓ f2 ↓ f3 ↓

	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed
7	Postgraduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed

multi class | binary class | Numerical

cols

value ←

field = science

yes NO

gini impurity

$$\text{gini impurity} = \boxed{1 - \sum_{i=1}^{k} P_k^2}$$

P_k = prob of each class

$P_{science}\left(\frac{5}{10}\right)$ $P_{art} = \left(\frac{5}{10}\right)$

$$1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2$$

	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed 1
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed 2
3	PhD	Science	85	Employed 3
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed 1
7	Postgraduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed 4

field = science

yes (5) NO (5)

emp 4 unemp 1 ⟷ emp 3 unemp 2

$P_{emp} = \frac{4}{5}$ $P_{unem} = \frac{1}{5}$ $P_{emp} = \frac{3}{5}$ $P_{un} = \frac{2}{5}$

gini

$$1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2$$

$$1 - \frac{16}{25} - \frac{1}{25} = \frac{8}{25}$$

gini

$$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

$$= 1 - \frac{9}{25} - \frac{4}{25} = \frac{12}{25}$$

$$\boxed{\frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{12}{25}}$$

$0.16 + 0.24 \longrightarrow$ $\boxed{0.40}$

G() value

G → (field, science)

0.4 0.4 $\boxed{0.1}$

multi class values → UG / PG / PhD

	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed

$\boxed{\text{degree, ug}}$ $\boxed{\text{degree, PG}}$ $\boxed{\text{degree, PhD}}$

G() = 0.5 G() G()

min

degree = UG

yes NO

(6)

4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed
7	Postgraduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed

degree = PG X

emp	unemp
2	1

emp	unemp
4	3

$$1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2 \qquad 1-\left(\frac{4}{7}\right)^2-\left(\frac{3}{7}\right)^2$$

0.25 0.35

$$\frac{3}{10}\times 0.25+\frac{7}{10}\times 0.35 = 0.65$$

Yes No

4 (6)

emp	unemp
1	3

emp	unemp
5	1

$$1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2-(\ ) \qquad 1-\left(\frac{5}{6}\right)^2-\left(\frac{1}{6}\right)^2$$

0.3 0.4

$$\frac{4}{10}\times 0.3+\frac{6}{10}\times 0.4=\boxed{0.5}$$ X

0.4

Splitting Numerical Features

21 July 2023 17:45

Average_Grade	Job_Outcome
89	Employed
92	Unemployed
95	Employed
85	Employed
98	Unemployed
90	Employed
88	Unemployed
93	Employed
94	Unemployed
86	Employed

e

p

p

$\boxed{\text{grade} > 89.5} \rightarrow \boxed{0.1} \rightarrow$

grade value

0.4 → 85.5
0.3 → 87
0.2 → 88.5
$\boxed{0.1}$ 89.5
91
0.8 92.5
0.7 93.5
94.5

85 – emp
86 – emp
88 – unemp
89 – emp
90 – emp
92 – unem
93 – emp
94 – unem
95 – emp
98 – unem

Grade > 85.5

Yes

emp	unemp
5	4

$$1-\left(\frac{5}{9}\right)^2-\left(\frac{4}{9}\right)^2$$

$$1-\frac{25}{81}-\frac{16}{81}$$

$$1-\frac{41}{81}=\frac{40}{81}$$

No

pure node

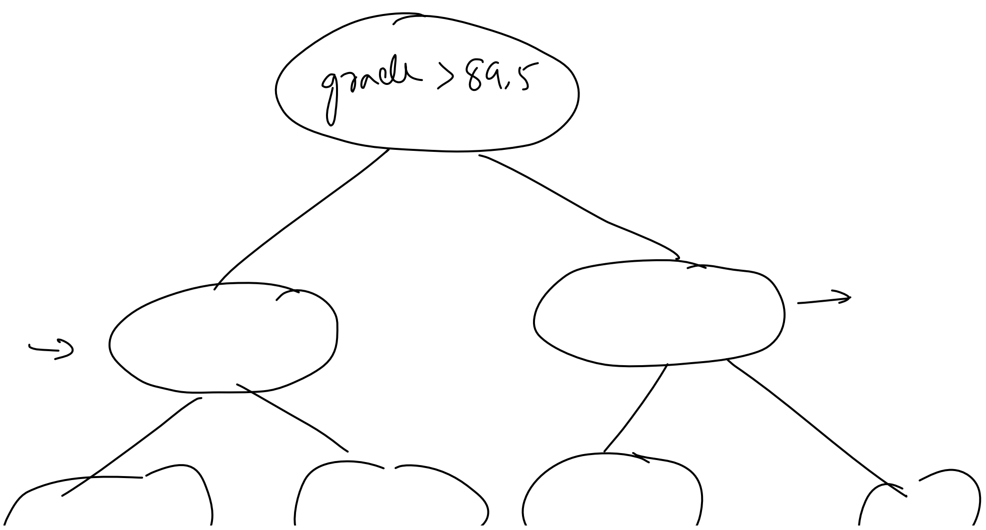
emp	unemp
1	0

→ gini = 0

$$1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2$$

$$1-1=0$$

$$G() = \frac{9}{10}\times\frac{40}{81}+\frac{1}{10}\times 0 = \boxed{0.25}$$




leafnode

Understanding Gini Impurity?

21 July 2023 16:40

The Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

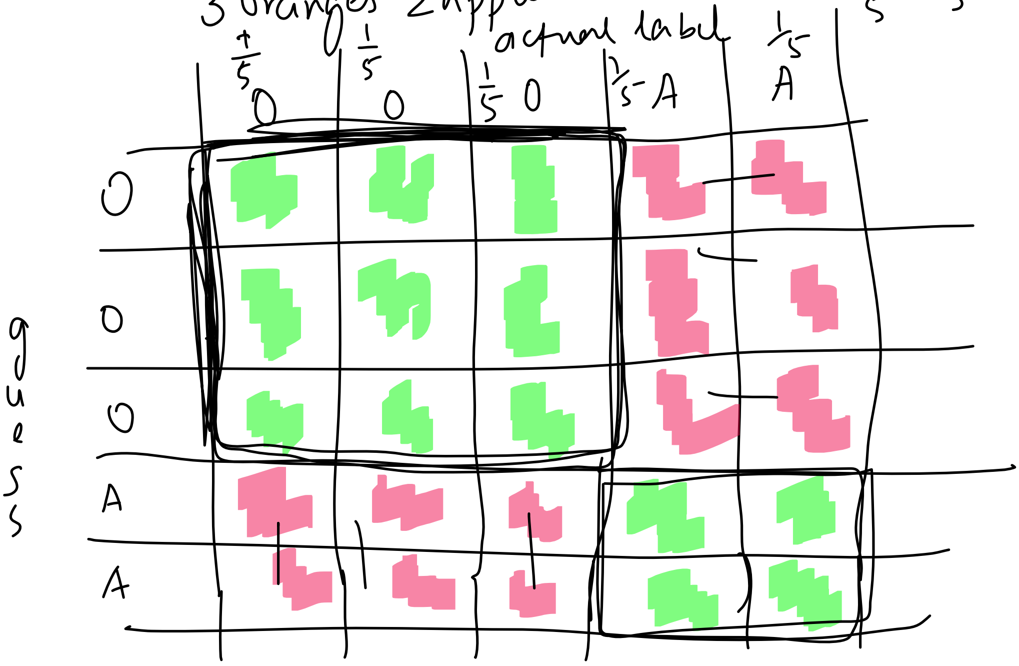
$$\boxed{H() = 1 - \sum_{i=1}^{r} p_i^2}$$

$O_1 \; O_2 \; O_3 \; A_1 \; A_2$

$O_1 \; O_2 \; O_3 \; O_4 \; O_5$

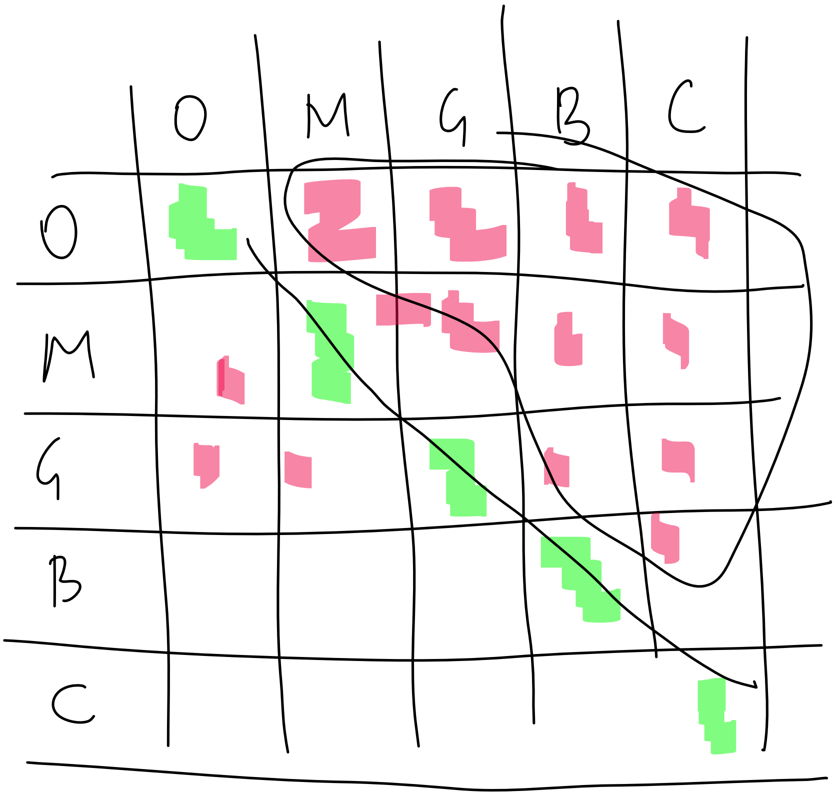
3 oranges 2apples

$\frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5}$

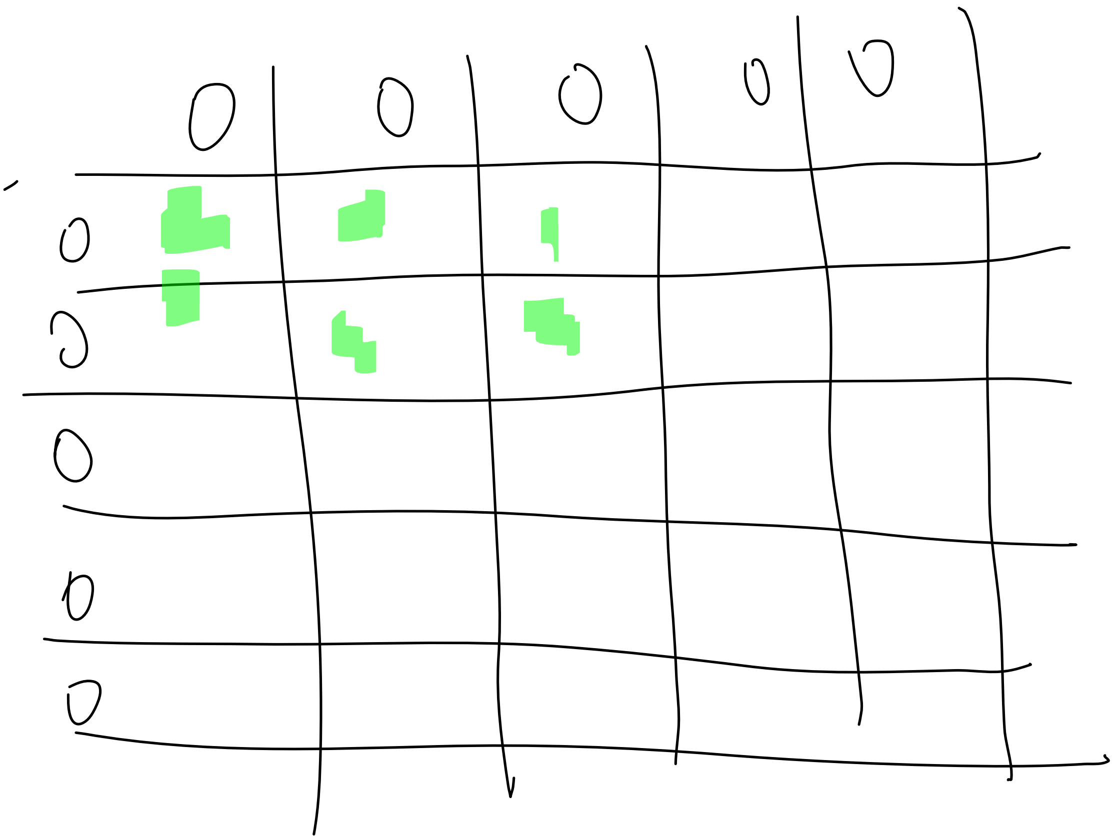


$$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

= gini impurity



$1-\left(\frac{1}{5}\right)^2-\left(\frac{1}{5}\right)^2-\left(\frac{1}{5}\right)^2-(\frac{1}{}$



$1-\left(\frac{5}{5}\right)^2$

$1-1=0$

$\boxed{1 > gini >= 0}$

$0 \rightarrow$ preferred

	o	m	g	b
o				
m				
g				
b				

$$1-\left(\frac{1}{3}\right)^2-\left(\frac{1}{3}\right)^2-\left(\frac{1}{3}\right)^2$$

$$1-\frac{1}{9}-\frac{1}{9}-\frac{1}{9}$$

$$\frac{9-3}{9}=\frac{2}{3}$$

$$\boxed{0.66}$$

$$1-\left(\frac{1}{4}\right)^2-\left(\frac{1}{4}\right)^2-\left(\frac{1}{4}\right)^2-\left(\frac{1}{4}\right)^2$$

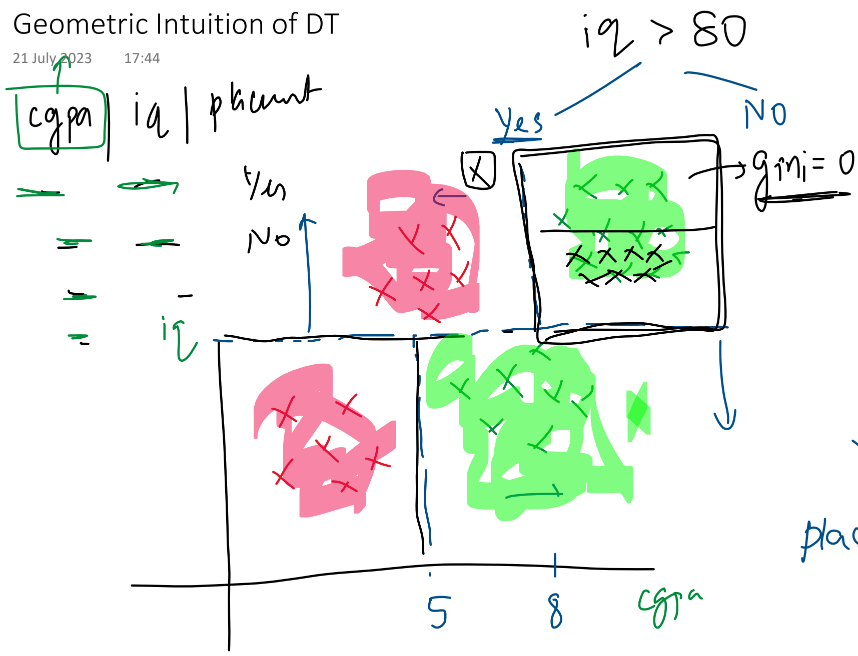
1

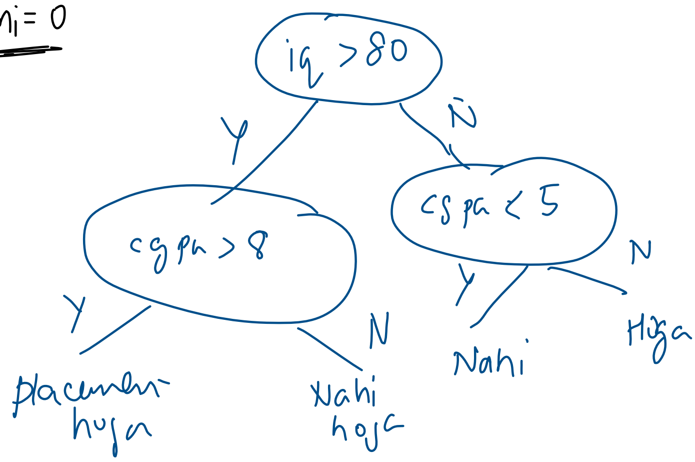
$$1-\frac{4}{16}=\frac{12}{16}$$

0.75

Geometric Intuition of DT

21 July 2023 17:44





Code

21 July 2023 17:49

The CART Algorithm - Regression

	Subject	Grade_Level	Hours_Studied	Test_Score
0	Math	Freshman	4	59
1	Physics	Freshman	1	82
2	Physics	Freshman	4	81
3	Math	Junior	6	60
4	Physics	Sophomore	1	73
5	Physics	Junior	3	85
6	Physics	Junior	4	61
7	Physics	Freshman	9	78

Code

21 July 2023 17:49

Advantages and Disadvantages

21 July 2023 17:36

Advantages

- Simple to understand and to interpret. Trees can be visualized.
- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data.
- Can work on non-linear datasets

Disadvantages

- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- Predictions of decision trees are neither smooth nor continuous, but piecewise constant approximations as seen in the above figure. Therefore, they are not good at extrapolation.
- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.
- Predictions of decision trees are neither smooth nor continuous, but piecewise constant approximations as seen in the above figure. Therefore, they are not good at extrapolation.