Decision Trees

16 February 2022 20:10

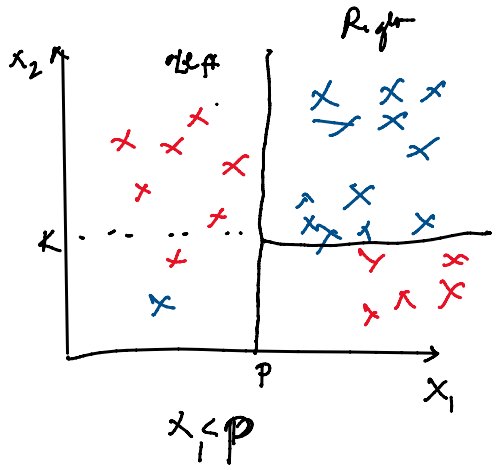
CART - Classification and Regression Trees.

Classification

Linear

1. Multicollinearity
2. Outliers
3. Non linear relationship

Tree

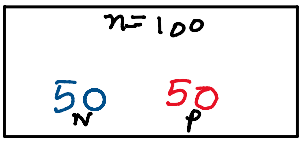


advantages

1. Multicollinearity
2. Outliers
3. Non linear boundaries ✓

Decision Tree - Root Node

$\{x_1, x_2 \ldots x_m\}$



← ? y

100 < 50 No
50 Yes

Measures of Impurity

Purity - When all obs belong to Same class

Impurity - When mix of both the classes

Metrics :- Entropy, Gini Index.

k - no. of classes

$$\text{Entropy} = -\sum_{i=1}^{k} p_i \log_2 p_i$$

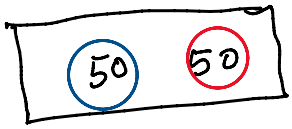
= $-\frac{50}{100}\log_2\frac{50}{100} - \frac{50}{100}\log_2\frac{50}{100}$

Impurity

Class 0

Class 1

Perfect Impure $E = 1$



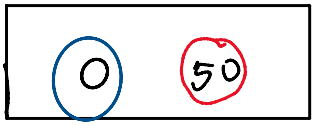
$$= -\frac{50}{100}\log_2\frac{50}{100} - \frac{50}{100}\log_2\frac{50}{100}$$

$$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2$$

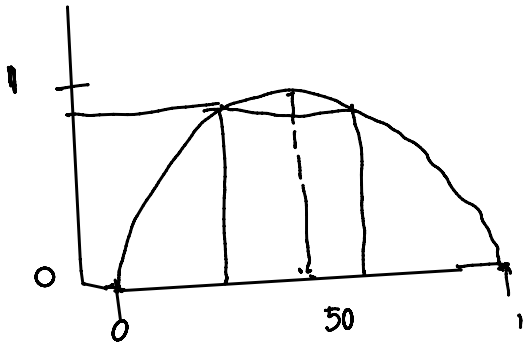
$$= \frac{1}{2} + \frac{1}{2} = 1$$

Pure $E = 0$



$$= \frac{0}{50}\log_2\frac{0}{50} - \frac{50}{50}\log_2\frac{50}{50}$$

$$= 0$$



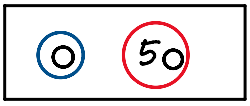
Gini Index

Perfect Impure $G = 0.5$

$$G = 1 - \sum_{i=1}^{k} (P_i)^2$$

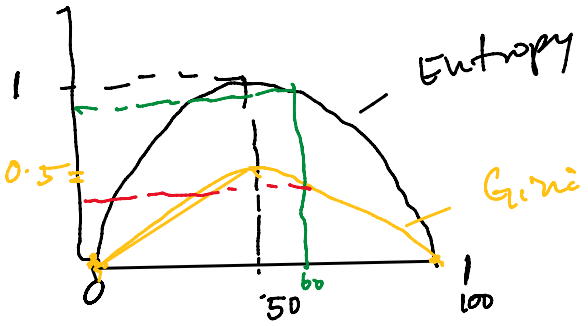
$$= 1 - \left(\frac{50}{100}\right)^2 - \left(\frac{50}{100}\right)^2$$

$$= 1 - \frac{1}{4} - \frac{1}{4} = 0.5$$



$$= 1 - \left(\frac{50}{50}\right)^2 - \left(\frac{0}{50}\right)^2 \qquad \text{Gini} = (0 - 0.5)$$

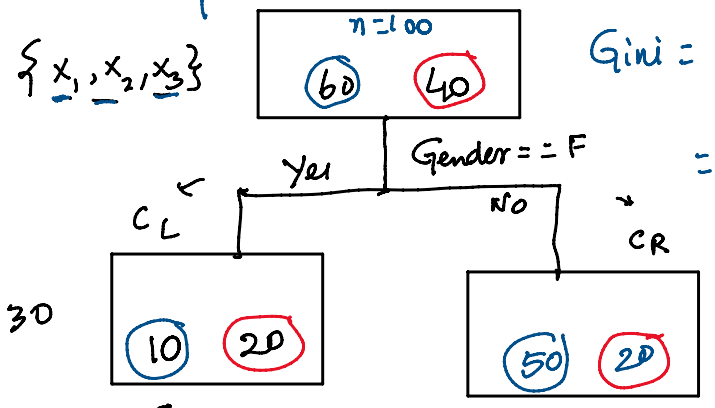
$$= 1 - 1 = 0$$



Root node.

$\{x_1, x_2, x_3\}$

Recursive Binary Partitions



$y - \begin{matrix} 0 - 60 \\ 1 - 40 \end{matrix}$

$$\text{Gini} = 1 - \left(\frac{40}{100}\right)^2 - \left(\frac{60}{100}\right)^2$$

$$= 0.48$$

TO

$$= \frac{30}{100} \times 0.44 + \frac{70}{100} \times 0.41$$

Childnodes?

$G_1 = 0.42$

	Gender		Total
Target Variable	Male	Female	
0 (Stays)	50	10	
1 (Churns)	20	20	
Total	70	30	100

	Age		Total
Target Variable	<35	>35	
0 (Stays)	50	10	
1 (Churns)	10	30	
Total	60	40	100

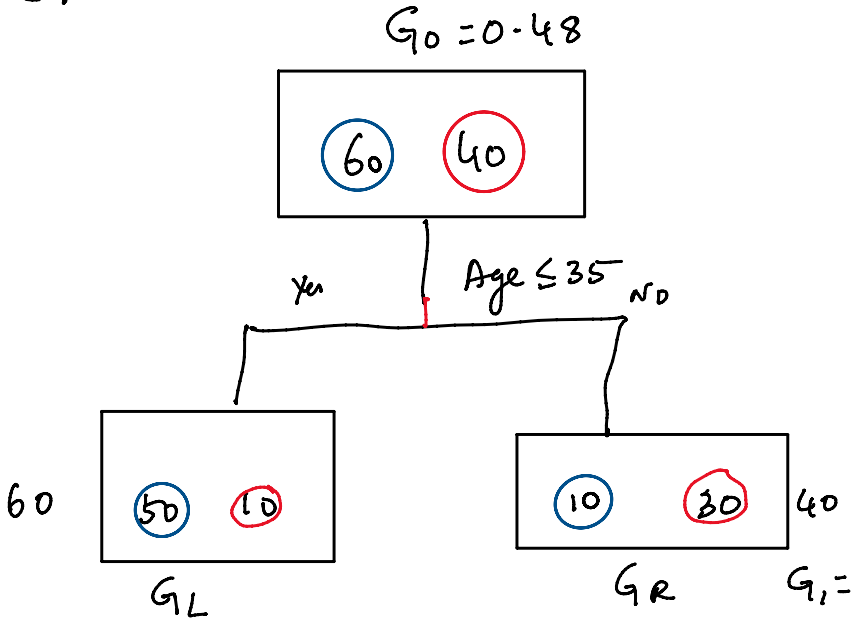
Information Gain, Gini decrease

$$G_C = 1 - \left(\frac{10}{30}\right)^2 - \left(\frac{20}{30}\right)^2 \qquad G_R = 1 - \left(\frac{50}{70}\right)^2 - \left(\frac{20}{70}\right)^2$$

$G_C = 1 - \left(\frac{\cdot}{30}\right) - \left(\frac{\cdot}{30}\right)$ -IR- $\left(\frac{70}{}\right)\left(\frac{70}{}\right)$

$= 0.44$ $= 0.41$

Gini decrease $= G_0 - G_1 = 0.48 - 0.42 = 0.06$
(Gender) = F



$= 1 - \left(\frac{50}{60}\right)^2 - \left(\frac{10}{60}\right)^2$ $1 - \left(\frac{10}{40}\right)^2 - \left(\frac{30}{40}\right)^2$

$=$ 0.27 0.37

$$G_1 = G_L \times \frac{\#L}{T} + G_R \frac{\#R}{T}$$

$$= 0.27 \times \frac{60}{100} + \underline{0.37} \times \frac{40}{100} = 0.18$$

Gini decrease $= G_0 - G_1 = 0.48 - 0.18 = 0.3$
(Age < 35)

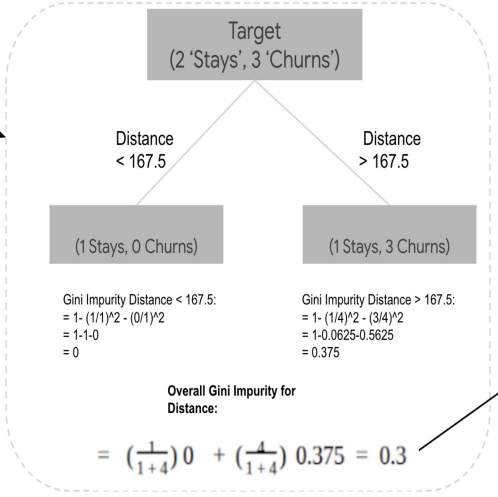
Variable	Gini decrease
Gender	0.06
Age < 35	0.3 ✓
distance < 205	0.21

(Max decrease in Gini)

1. Which variable to split
2. Split Criteria

Distance	Target
155	0
180	1
190	0
220	1
225	1

167.5, 185, 205, 222.5



Distance	Target	
155	0	Gini = 0.3
~~167.5~~		
180	1	Gini = 0.47
~~185~~		
190	0	Gini = 0.27
~~205~~		
220	1	Gini = 0.4
~~222.5~~		
225	1	

distance <= 167.5

Gini decrease

0.48 - 0.3

0.48 - 0.47

0.48 - 0.27 ✓

0.48 - 0.4

distance < 167.5 Gini = 0.3

distance < 185 Gini = 0.47

dist < 205 Gini = 0.27 ✓

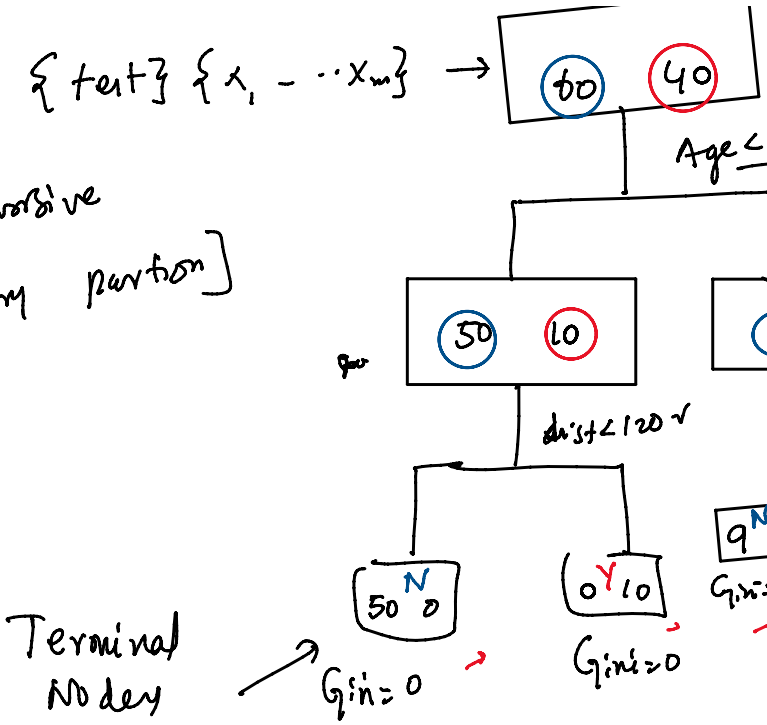
dist < 222.5 Gini = 0.4

Decision Tree

{test} {$x_1 \dots x_m$} → (60) (40)

① find the best split

{test} $\{x_1 - \cdots x_m\}$ →



① find the best Split among a variable

② Compare across variables

③ take the Split that gives max decrease in Gini

④ Recursively apply

⑤ Untill Gini is '0' or it increases

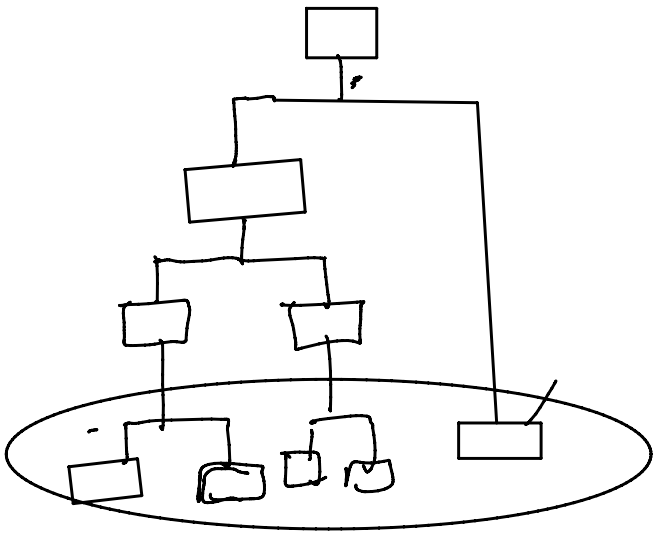
Recursive [Binary partion]

Terminal Nodes / leaf nodes / decision nodes

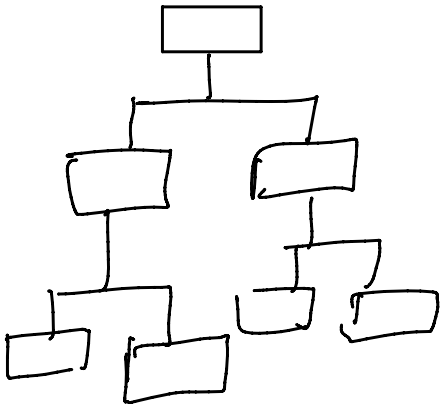
Gini = 0

Entropy

Tree1



Tree2



Entropy - (Computationally complex) ✗

Gini - very fast ✓

{0 1 2 3}

Split Criteria

Cont — X Sort — y

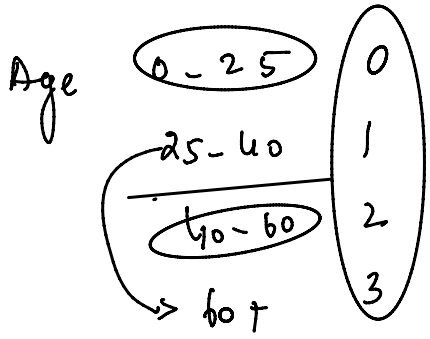
Binary
$x_1 == 0$

Categorical (A, B, C)
$x_1 == A$
$x_1 \{A, B\}$
$x_1 \{A, C\}$
A | B C
AB | C
AC | B

{0 1 2 3}

Ordinal
$x_1 <= 1$
$x_1 <= 2$
$x_1 <= 0$
~~(0, 2 | 1, 3)~~

label encoder :- A B C D
(0 1 2 3)



0, 1 | 2, 3 ~~X~~

0, 2 | 1, 3

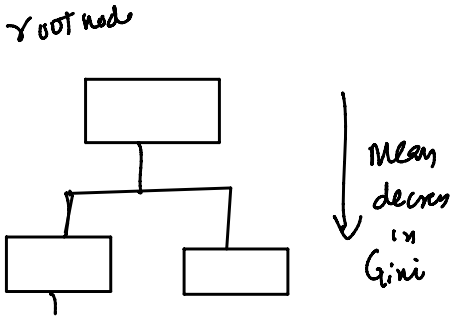
{0 2, 1 3}

0 1, 2 3

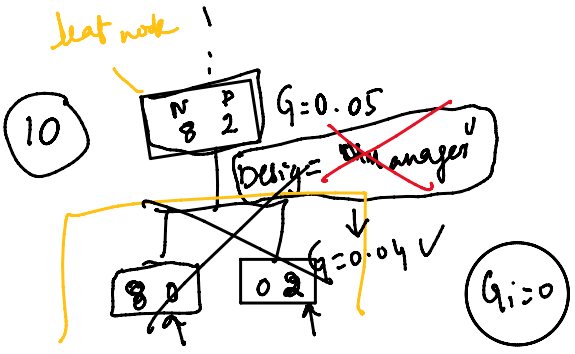
① Greedy ✓

② Overfiting

33 → (X) inputs



① Pruning { Post *, Pre -



Stop tree Growth if- any of

Stop tree Growth if- any of these conditions are satisfied

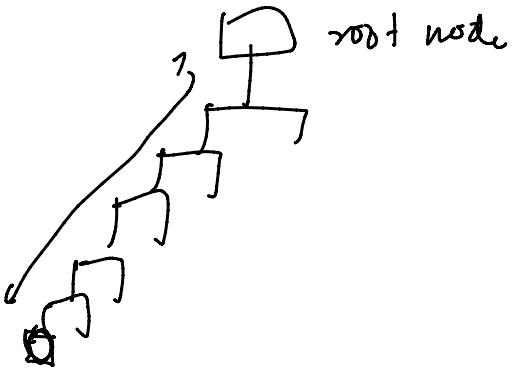
(1) if Max depth is reached — Max depth = 10

(2) if no. of samples in a node is less than 20

(3) Min no. of samples in a leaf node = 5

Will Stop tree growth

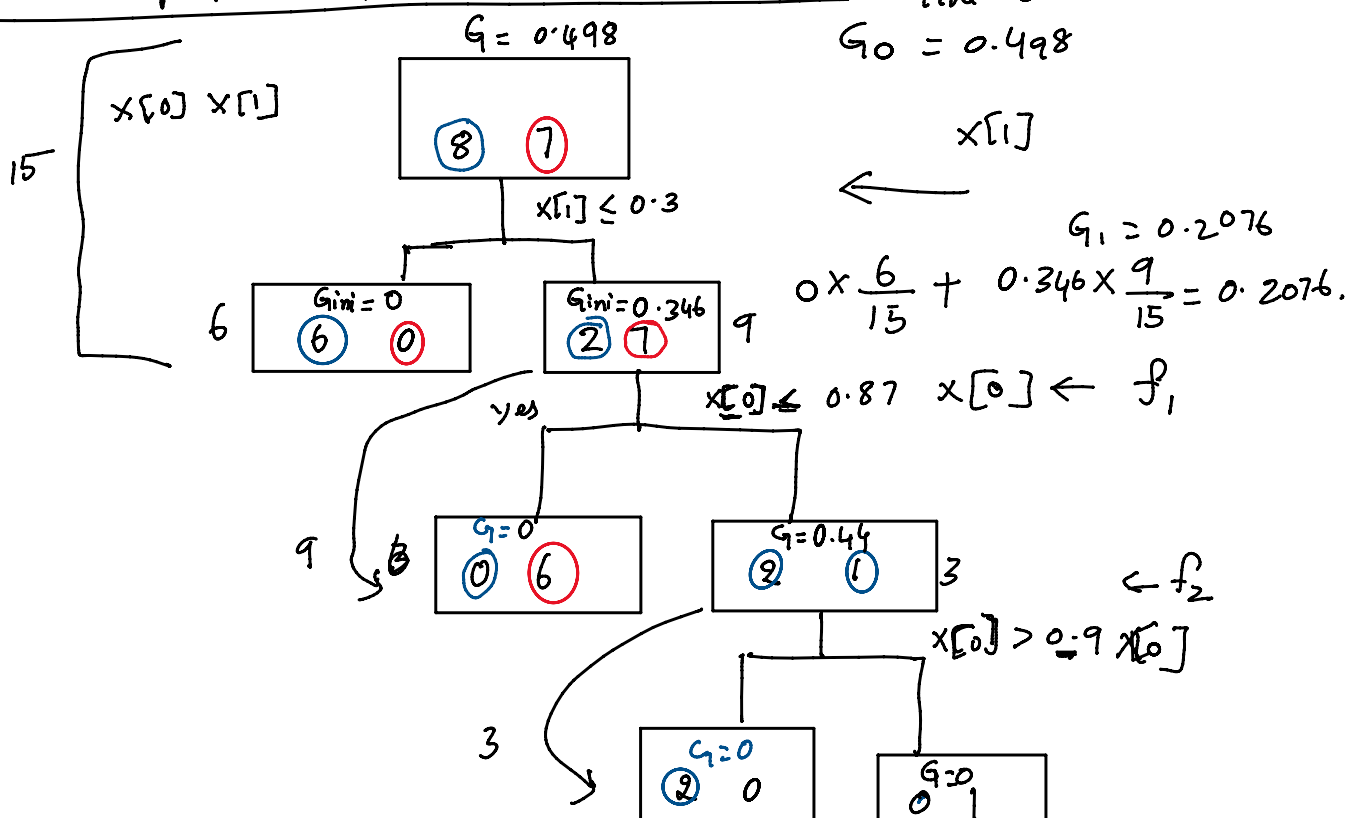
Reduce overfitting.

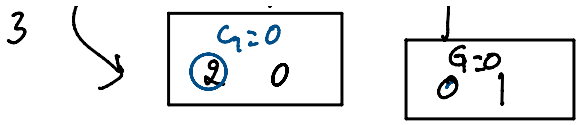


Feature Importance from a decision Tree

pure Impure

Gini 0 - 0.5





Variable Importance of $X[1]$

$$= 0.498 - 0.2076 = 0.2904$$

$$= 0.2904 \times \frac{15}{15} = 0.2904$$

Variable Importance $X[0]$ f_1 $G_0 = 0.346$

G_0 $(G_L + G_R)w.$

$$0.346 - 0 \times \frac{6}{9} + 0.44 \times \frac{3}{9} = 0.1981$$

f_2: $G_0 = 0.44$

$$0.44 - 0 = 0.44$$

$$0.1981 \times \frac{9}{15} + \frac{3}{15} \times 0.44 = 0.2074$$

$X[1] = 0.2904$ $= \frac{0.2904}{0.2904 + 0.2074} = 0.5833$

$X[0] = 0.2074$ $= \frac{0.2074}{0.2074 + 0.2904} = 0.4166$

3, 50, 25 $\left(\mu = 25\right)$ $\frac{1}{3}\left[(25-25)^2 + (4-25)^2 + (51-25)^2 + (50-25)^2\right]$

3, 50L $\mu = 25L$ $(x-25)^2 + (4-25)^2 + (5-1-25)^2 + (50-25)^2$

Regression

x

{age, desg, gen, dept, }