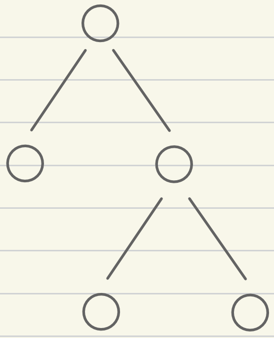
Trees



How do we decide the split ?

For classification

- Gini index
- Entropy

Gini index :

$$G = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk})$$

Entropy :

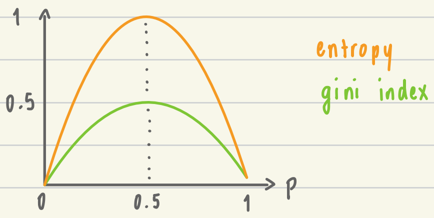
proportion of observations in the m^{th} region that are from the k^{th} class

$$D = - \sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

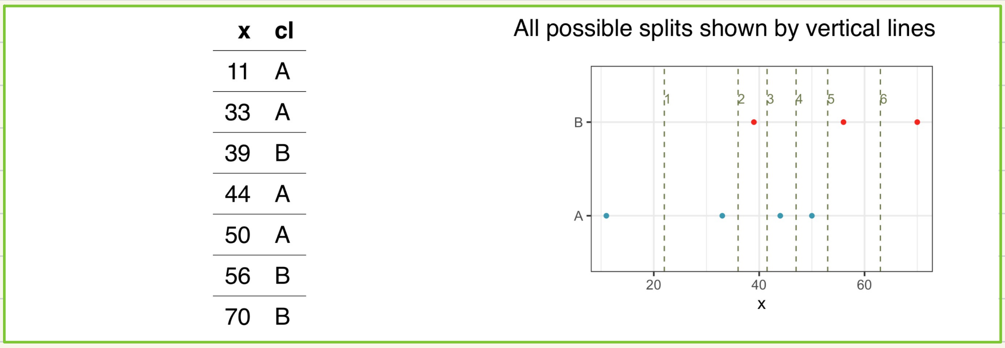
→ class

↳ subset of observations

- a small value in gini index mean the nodes contain mostly the observation from the same class .
- just like gini index , a small value in entropy corresponding to when nodes are pure



Example from the lecture



x	cl
11	A
33	A
39	B
44	A
50	A
56	B
70	B

All possible splits shown by vertical lines

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

- 2 subset, left or right
- 2 classes, A or B

For split 2:

Left : 2As, 0B

Right : 2As, 3Bs

Step 1 - calculates entropy for each subset

Left :

$\hat{p}_{LA} = 2/2$, $\hat{p}_{LB} = 0/0$

$$D_L = -\left\{ \underbrace{[1 \log(1)]}_{\hat{p}_{LA} = 1} + \underbrace{[0 \log(0)]}_{\hat{p}_{LB} = 0} \right\} = 0$$

Right :

$\hat{p}_{RA} = 2/5$, $\hat{p}_{RB} = 3/5$

$$D_R = -\left\{ \underbrace{[0.4 \log(0.4)]}_{\hat{p}_{RA} = 0.4} + \underbrace{[0.6 \log(0.6)]}_{\hat{p}_{RB} = 0.6} \right\}$$

$$= 0.673$$

Step 2 : Combine the weighted sum

$$D = 2/7\, D_L + 5/7\, D_R$$

$$= 0.673$$

For Split 5 :

Left : 4As , 1 B

Right : 0A , 2 Bs

Left :

$\hat{p}_{LA} = 4/5$, $\hat{p}_{LB} = 1/5$

$$D_L = -\left\{ [0.8 \log(0.8)] + [0.2 \log(0.2)] \right\} = 0.5$$

Right :

$\hat{p}_{RA} = 0/2$, $\hat{p}_{RB} = 2/2$

$$D_R = -\left\{ [0 \log(0)] + [1 \log(1)] \right\}$$

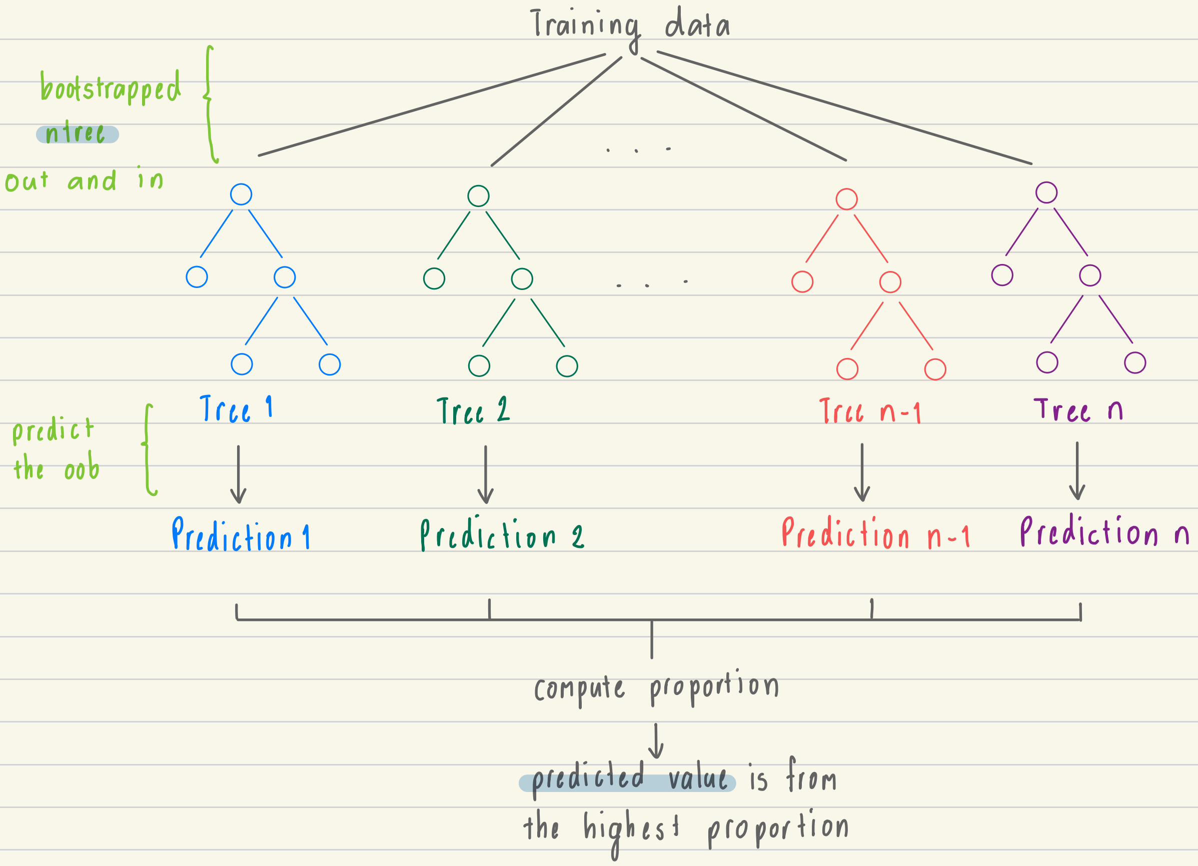
$$= 0$$

$$D = 5/7\, D_L + 2/7\, D_R$$

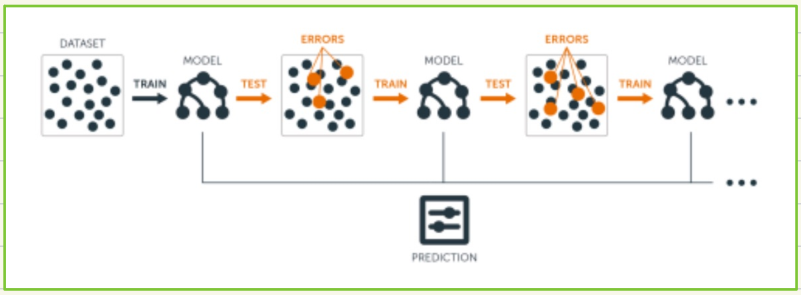
$$= 0.3571$$

Sruthi (2024) Understand Random Forest Algorithms With Examples (Updated 2024)

Random Forest



Boosted trees



- boosted trees can lead to overfitting, while random forest does not lead to overfitting