(1)

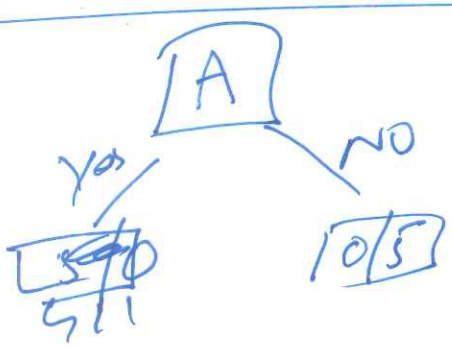
Decision Tree

$$Gini = 1 - \sum p_i^2$$

Target: Y — $6 = 6/10 = 0.6$

N — $4 = 4/10 = 0.4$

$$Gini_{root} = 1 - ((0.6)^2 + (0.4)^2)$$
$$= 1 - (0.16 + 0.36) = 1 - 0.52 = 0.48$$



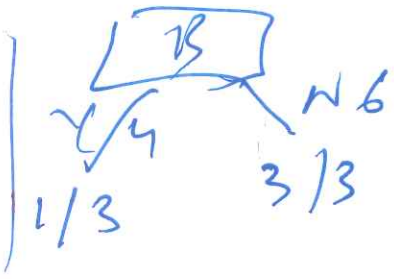
$$Gini_A(Yes) = 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right)$$
$$= 1 - (0.64 + 0.32)$$
$$= 0.32$$

$$Gini_A(NO) = 1 - \left(\left(\frac{0}{5}\right)^2 + \left(\frac{5}{5}\right)^2\right)$$
$$= 1 - (1) = 0$$

$$\text{Weighted } Gini_A = \frac{5}{10} \times 0.32 + \frac{5}{10} \times 0$$
$$= 0.16$$

(2)

$Gini_B(Yes) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right)$



$= 1 - (0.0625 + 0.5625)$

$= 1 - 0.625$

$= 0.375$

$Gini_B(No) = 1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right)$

$= 1 - (0.25 + 0.25)$

$= 0.5$

Weighted $Gini_B = \frac{4}{10} \times 0.375 + \frac{6}{10} \times 0.5$

$= 0.45$

$Gini_C(Y) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)$

$= 0.48$

$Gini(N) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)$

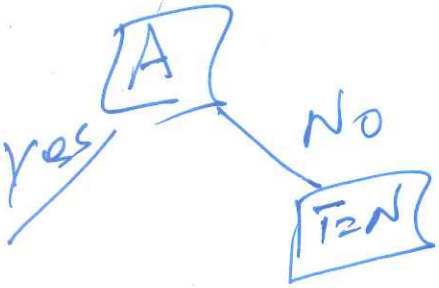
$= 0.48$

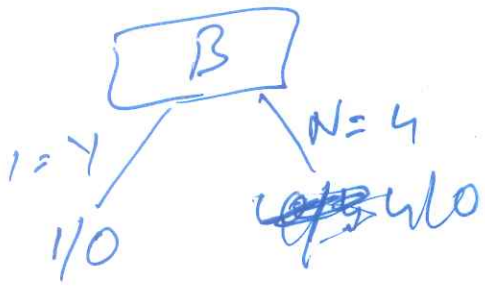
Weighted $Gini_C = \frac{5}{10} \times 0.48 + \frac{5}{10} \times 0.48$

$= 0.48$

GiniA has best split — lowest score.

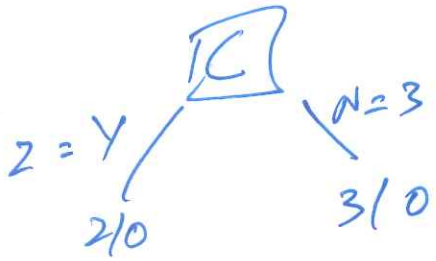
(4)





$$Gini_B(Y) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right)$$
$$= 0$$
$$Gini_B(N) = 1 - \left(\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right)$$
$$= 0$$

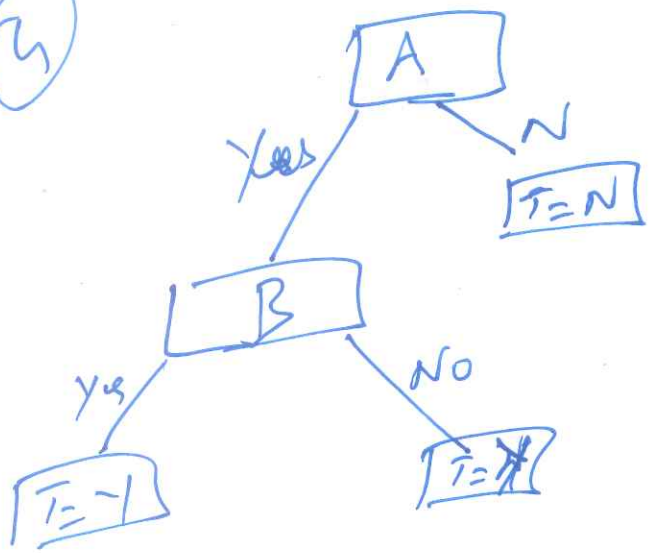
$$\text{Weighted } Gini\ B = \frac{1}{5} \times 0 + \frac{4}{5} \times 0$$
$$= 0$$



$$Gini_C(Y) = 0$$
$$Gini_C(N) = 0$$
$$\text{Weighted } Gini = 0$$

We can select B or C.

(4)



→ easy to build, easy to use, easy to interpret.

→ good to create with data (not good to predict

Overfit — memorize the train data

Sensitive — small change = diff. tree
instability.

Imbalanced data = bias towards dominant class

Random Forest

Bootstraping = Bagging.

- More than one trees
  - Forest.
    - same tree as the same data?

Bootstrapped data set
- ~~Allow~~ Select Sample from the original data
  + Allow to repeat. → Bag
- some data may not be part of the Bag — out of bag data
- Select a few features (randomly)
  - default for classification: $\sqrt{n}$ or 2
- different trees with different features.

How to classify

→ for any new data

- run through all DTs.
- Voting: the class ~~with most~~ in the result of most DTs is the winner

→ how to measure the model

- Out of Bag samples = data not selected
- run these samples on DTs.
- we have label and predicted labels.