Week 4

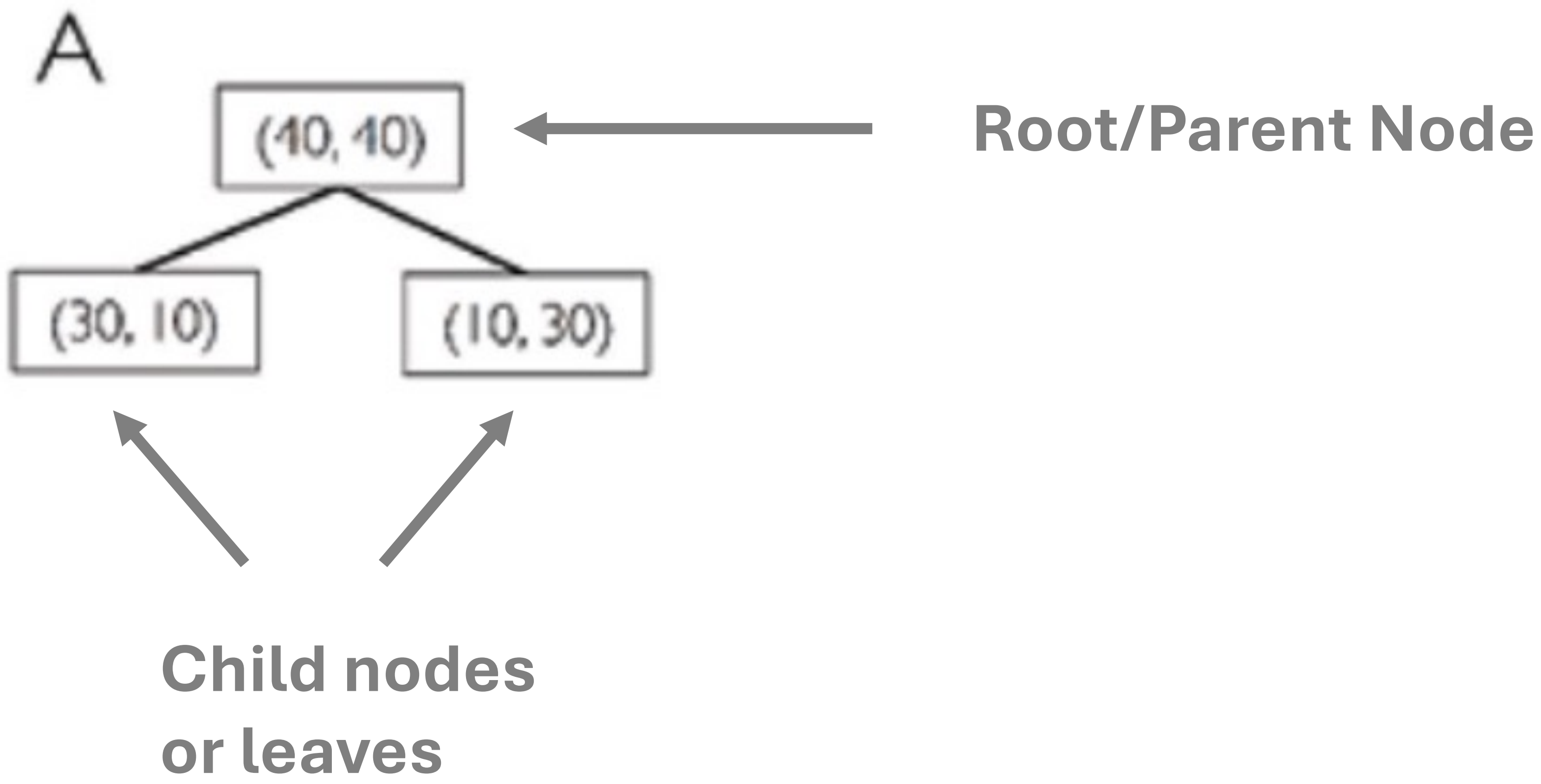
Decision Tree Question

Background

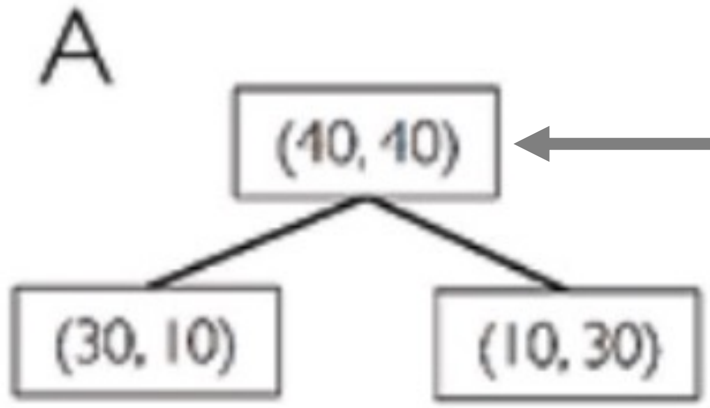
- **Machine learning** = learning from data, often to predict something we want to know
  - Loan default prediction
  - Cancer detection
  - Fraud detection

- **Decision tree** = popular machine learning algorithm which infers decision rules based on features

Decision tree example



Decision tree example



80 customers in total

- 40 are fraudulent
- 40 are genuine

Left child

40 total customers in this leaf

- 30 fraudulent
- 10 genuine

Right child

40 total customers in this leaf

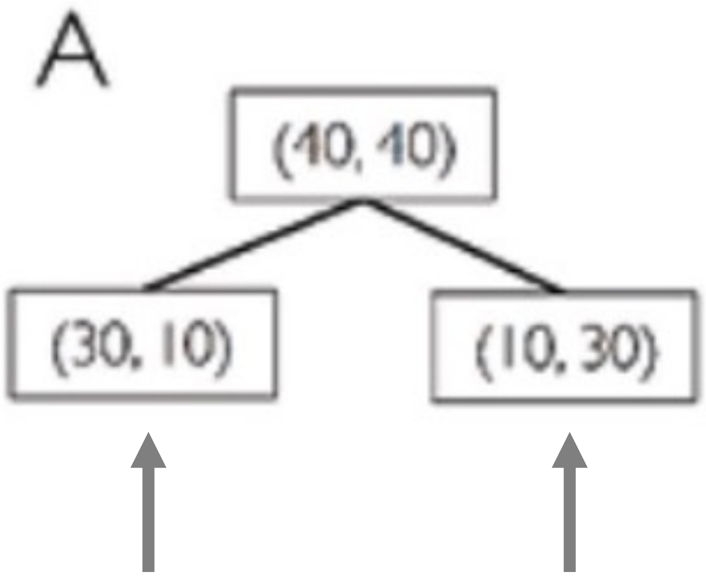
- 10 fraudulent
- 30 genuine

How do Decision Trees learn best way to split?

Which tree (A or B) has better splits?

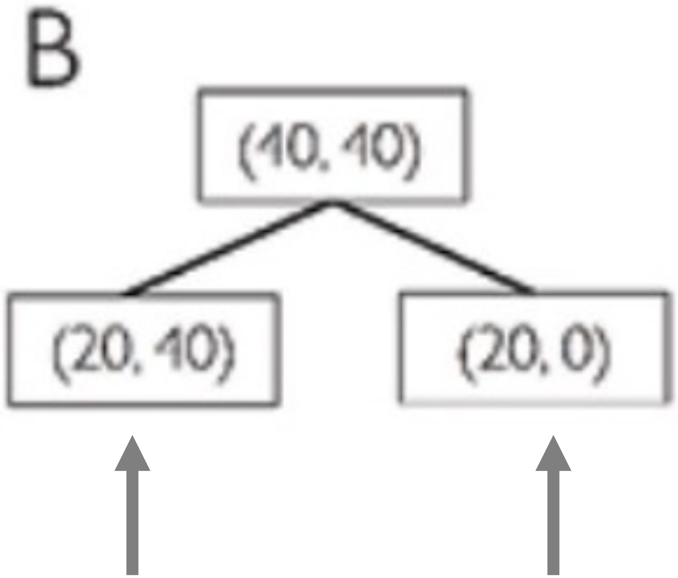
How do Decision Trees learn best way to split?

Which tree (A or B) has better splits?



75% fraudulent
25% genuine

25% fraudulent
75% genuine



33% fraudulent
66% genuine

100% fraudulent
0% genuine

How do Decision Trees learn best way to split?

Can figure this out mathematically using 2 possible formulas

Entropy

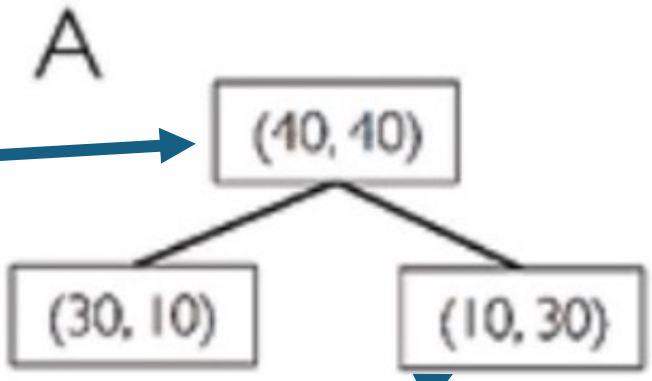
Gini Impurity

Both serve same fundamental purpose –
to quantify how pure some data is

Calculating entropy

Tree A

Entropy $I_H = -\sum_{i=1}^{c} p(i \mid t) \log_2 p(i \mid t)$



$$I_H(D_p) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) \\ = 1$$

$$I_H(D_{left}) = -(0.75 \log_2(0.75) + 0.25 \log_2(0.25)) \\ = 0.81$$

$$I_H(D_{right}) = -(0.25 \log_2(0.25) + 0.75 \log_2(0.75)) \\ = 0.81$$

Calculating entropy

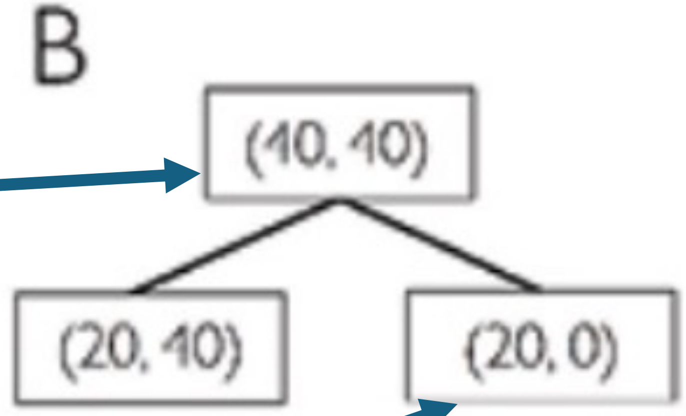
Tree B

Entropy $I_H = -\sum_{i=1}^{c} \; p(i \mid t) \log_2 p(i \mid t)$

$$I_H(D_p) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) \\ = \; 1$$

$$I_H(D_{left}) = -(0.33 \log_2(0.33) + 0.66 \log_2(0.66)) \\ = 0.92$$

$$I_H(D_{right}) = -(1 \log_2(1) + 0 \log_2(0)) \\ = 0$$



How do Decision Trees learn best way to split?

Information Gain calculation

Information gain calculates improvement in Entropy from parent to children
i.e. how much purer does our data become after splitting?

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

Information Gain

Entropy of parent

examples in left leaf over total # examples

Entropy of left leaf

examples in right leaf over total # examples

Entropy of right leaf

Calculating Information Gain

Tree A

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

Entropy of parent

$IG(D_p, f)$ = 1

 # examples left leaf/total

$\frac{N_{left}}{N_p} = \frac{40}{80}$ = 0.5

 # examples right leaf/total

$\frac{N_{right}}{N_p} = \frac{40}{80}$ = 0.5

Entropy of left leaf

$I(D_{left})$ = 0.81

Entropy of right leaf

$I(D_{right})$ = 0.81

Information Gain for tree A

$$IG = 1 - 0.5(0.81) - 0.5(0.81)$$
$$= 0.19$$

Calculating Information Gain

Tree B

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

Entropy of parent

$IG(D_p, f)$ = 1

 # examples left leaf/total

$\frac{N_{left}}{N_p} = \frac{60}{80} = 0.75$

 # examples right leaf/total

$\frac{N_{right}}{N_p} = \frac{20}{80} = 0.25$

Entropy of left leaf

$I(D_{left})$ = 0.92

Entropy of right leaf

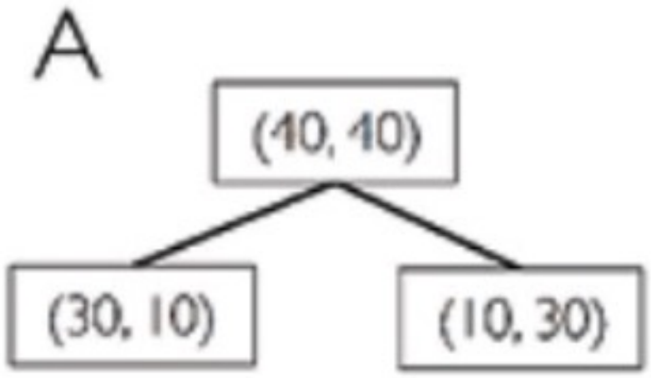
$I(D_{right})$ = 0

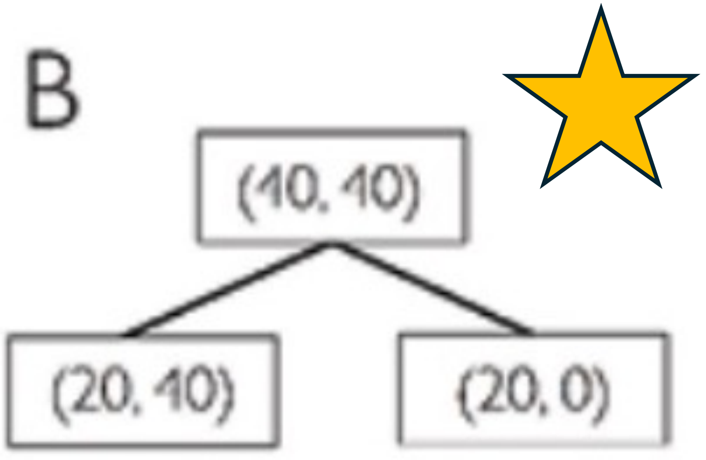
Information Gain for tree B

$$IG = 1 - 0.75(0.92) - 0.25(0) \\ = 0.31$$

Which tree is better?

Comparing both IG calculations, Tree B is better





Information Gain for tree A

$$IG = 1 - 0.5(0.81) - 0.5(0.81)$$
$$= 0.19$$

Information Gain for tree B

$$IG = 1 - 0.75(0.92) - 0.25(0)$$
$$= 0.31$$

Gini Impurity

What is it?

Gini Impurity = another way to quantify purity instead of entropy

Gini Impurity $I_H = -\sum_{i=1}^{c} p(i \mid t)(1 - p(i \mid t)$

Try it out yourself on the examples previously!

Entropy vs Gini Impurity

How are they different?

- Both serve as ways to evaluate quality of splitting your data
- Both should give roughly similar results
- Some packages use Entropy, others use Gini Impurity