CART Algorithm (Briemanetal, 1984)

- Uses an alternative to *entropy* as an impurity metric: *Gini Index*

Gini Index – rate of misclassification

$$Gini(S) = 1 - \sum_{c \in Classes} p_c^2$$

S – subset of training examples

p_c - proportion of examples in S belonging to class c

Higher Gini index → more likely to misclassify

Gini Index example

e.g. (3 yes / 3 no)

Gini Index example

e.g. (3 yes / 3 no)

Gini(S) = $1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$

Gini Index example

e.g. (3 yes / 3 no)

Gini(S) = $1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$

e.g. (4 yes / 0 no)

Gini Index example

e.g. (3 yes / 3 no)

Gini(S) = $1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$

e.g. (4 yes / 0 no)

Gini(S) = $1 - (\frac{4}{4})^2 - (\frac{0}{4})^2 = 0$

Gini Index example

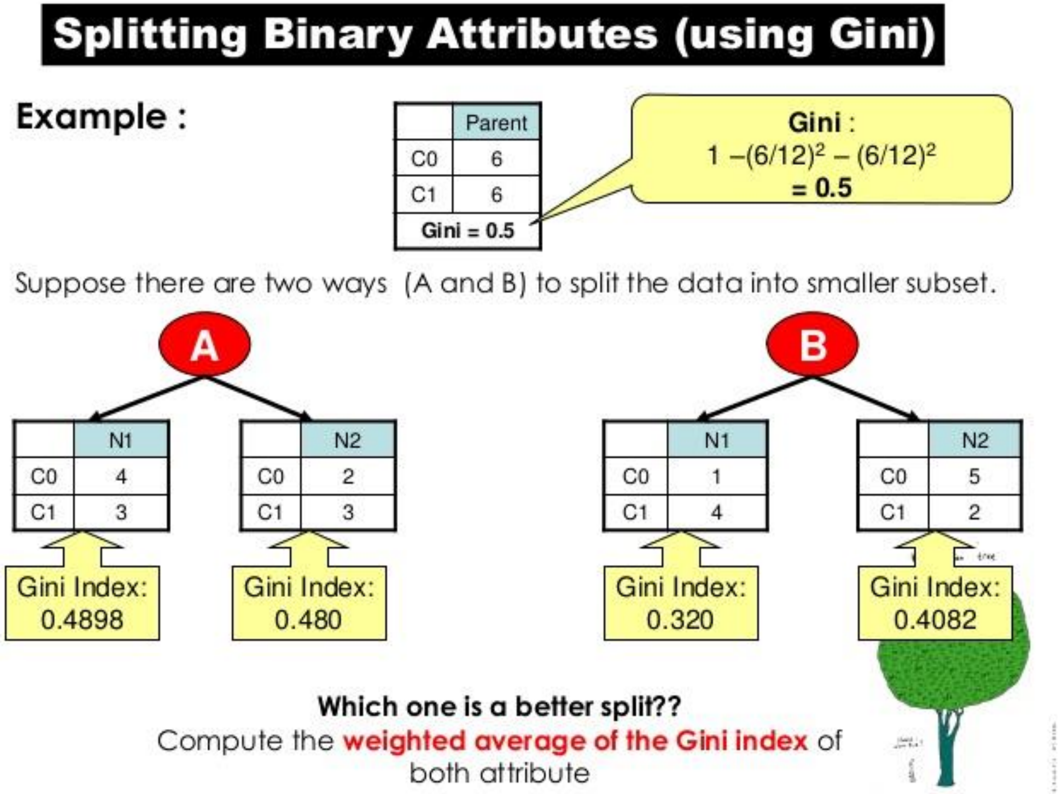
e.g. (3 yes / 3 no)

$$\text{Gini(S)} = 1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$$

e.g. (4 yes / 0 no)

$$\text{Gini(S)} = 1 - (\frac{4}{4})^2 - (\frac{0}{4})^2 = 0$$

Gini Index example



Information Gain

- **Information Gain** – expected reduction in the *misclassification rate* after a split on an attribute

$$Gain(S,A) = Gini(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Gini(S_v)$$

A – Attribute

$Values(A)$ – possible values of A

S – subset of training examples

S_v – subset of S for which attribute A have value v

Entropy vs Gini Index

- So which impurity measure should be used:

  Entropy or Gini index?

- Resulting trees are very similar in practice.
- Best advice is that it is good practice when building decision tree models to
  - Try out different impurity metrics
  - Compare results to see which suits the dataset