Using **GINI Index** for Decision Tree Construction

Measure of Impurity: GINI

- GINI Index for a given node t:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

Where $p(j|t)$ is the relative frequency of class j at node t

- ***Maximum*** $\left(1 - \frac{1}{n_c}\right)$ when records are equally distributed among all classes, implying ***least information***
- ***Minimum (0)*** when all records belong to one class, implying ***most information***

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

Computing GINI

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

Splitting Based on GINI

- Used in CART, SLIQ, SPRINT
- When a node p is split into k partitions (children), the quality of split is computed as

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Where:

n_i is the number of records at child i

n is the number of records at node p

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	Yes	Rich	Cinema
Sunny	No	Rich	Tennis
Windy	Yes	Rich	Cinema
Rainy	Yes	Poor	Cinema
Rainy	No	Rich	Stay In
Rainy	Yes	Poor	Cinema
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Windy	Yes	Rich	Cinema
Sunny	No	Rich	Tennis

Which attribute should be at the top of the tree?

Step 1: Compute Gini of whole dataset.

$$GINI(S) = 1 - \sum_j [p(j|t)]^2$$

$$GINI\ (S) = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2\right] = 0.58$$

Step 2: Computer GINI Index for all attributes.

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	Yes	Rich	Cinema
Sunny	No	Rich	Tennis
Windy	Yes	Rich	Cinema
Rainy	Yes	Poor	Cinema
Rainy	No	Rich	Stay In
Rainy	Yes	Poor	Cinema
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Windy	Yes	Rich	Cinema
Sunny	No	Rich	Tennis

Attribute: Money
Values (Rich) = [Cinema:3 , Tennis: 2, Stay in: 1, Shopping 1]
Values (Poor) = [Cinema:3 , Tennis: 0, Stay in: 0, Shopping 0]

$$GINI\left(Money_{poor}\right) = 1 - \left[\left(\frac{3}{3}\right)^2\right] = 0$$

$$GINI\left(Money_{Rich}\right) = 1 - \left[\left(\frac{2}{7}\right)^2 + \left(\frac{3}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right] = 0.694$$

$$GINI\left(Money\right) = 0 \times \frac{3}{10} + 0.694 \times \frac{7}{10} = 0.486$$

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	Yes	Rich	Cinema
Sunny	No	Rich	Tennis
Windy	Yes	Rich	Cinema
Rainy	Yes	Poor	Cinema
Rainy	No	Rich	Stay In
Rainy	Yes	Poor	Cinema
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Windy	Yes	Rich	Cinema
Sunny	No	Rich	Tennis

Attribute: Parents
Values (Yes) = [Cinema:5 , Tennis: 0, Stay in: 0, Shopping 0]
Values (No) = [Cinema:1 , Tennis: 2, Stay in: 1, Shopping 1]

$$GINI\ (Parents_{Yes}) = 1 - \left[\left(\frac{5}{5}\right)^2\right] = 0$$

$$GINI\ (Parents_{No}) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0.72$$

$$GINI\ (Parents) = 0 \times \frac{5}{10} + 0.72 \times \frac{5}{10} = 0.36$$

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	Yes	Rich	Cinema
Sunny	No	Rich	Tennis
Windy	Yes	Rich	Cinema
Rainy	Yes	Poor	Cinema
Rainy	No	Rich	Stay In
Rainy	Yes	Poor	Cinema
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Windy	Yes	Rich	Cinema
Sunny	No	Rich	Tennis

Attribute: Weather
Values (Sunny) = [Cinema:2 , Tennis: 1, Stay in: 0, Shopping 0]
Values (Rainy) = [Cinema:2 , Tennis: 0, Stay in: 1, Shopping 0]
Values (Windy) = [Cinema:3 , Tennis: 0, Stay in: 0, Shopping 1]

$$GINI\left(Weather_{Sunny}\right) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.444$$

$$GINI\left(Weather_{Rainy}\right) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.444$$

$$GINI\left(Weather_{Windy}\right) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$$

$$GINI\left(Weather\right) = 0.444 \times \frac{3}{10} + 0.444 \times \frac{3}{10} + 0.375 \times \left(\frac{4}{10}\right) = 0.416$$

GINI Index: Numerical Example

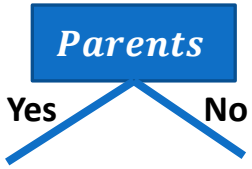
Weather	Parents	Money	Decision
Sunny	Yes	Rich	Cinema
Sunny	No	Rich	Tennis
Windy	Yes	Rich	Cinema
Rainy	Yes	Poor	Cinema
Rainy	No	Rich	Stay In
Rainy	Yes	Poor	Cinema
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Windy	Yes	Rich	Cinema
Sunny	No	Rich	Tennis

$$GINI\ (Money) = 0.486$$

$$GINI\ (Parents) = 0.36$$

$$GINI\ (Weather) = 0.416$$

Parents is selected as root node because it has smallest GINI Index



GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	Yes	Rich	Cinema
Sunny	No	Rich	Tennis
Windy	Yes	Rich	Cinema
Rainy	Yes	Poor	Cinema
Rainy	No	Rich	Stay In
Rainy	Yes	Poor	Cinema
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Windy	Yes	Rich	Cinema
Sunny	No	Rich	Tennis

Lets grow tree further in both directions

Parents

Yes

Weather	Parents	Money	Decision
Sunny	Yes	Rich	Cinema
Windy	Yes	Rich	Cinema
Rainy	Yes	Poor	Cinema
Rainy	Yes	Poor	Cinema
Windy	Yes	Rich	Cinema

$$GINI\ (S) = 1 - \left[\left(\frac{5}{5}\right)^2\right] = 0$$

No need to grow tree further. Decide Class label

Cinema

No

Weather	Parents	Money	Decision
Sunny	No	Rich	Tennis
Rainy	No	Rich	Stay In
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Sunny	No	Rich	Tennis

$$GINI\ (S) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0.28$$

Lets grow tree in this direction

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	No	Rich	Tennis
Rainy	No	Rich	Stay In
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Sunny	No	Rich	Tennis

Attribute: Weather
Values (Sunny) = [Cinema:0 , Tennis: 2, Stay in: 0, Shopping 0]
Values (Rainy) = [Cinema:0 , Tennis: 0, Stay in: 1, Shopping 0]
Values (Windy) = [Cinema:1 , Tennis: 0, Stay in: 0, Shopping 1]

$$GINI\left(Weather_{Sunny}\right) = 1 - \left[\left(\frac{2}{2}\right)^2\right] = 0$$

$$GINI\left(Weather_{Rainy}\right) = 1 - \left[\left(\frac{1}{1}\right)^2\right] = 0$$

$$GINI\left(Weather_{Windy}\right) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$$

$$GINI\left(Weather\right) = 0 \times \frac{2}{5} + 0 \times \frac{1}{5} + 0.5 \times \left(\frac{2}{5}\right) = 0.2$$

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	No	Rich	Tennis
Rainy	No	Rich	Stay In
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Sunny	No	Rich	Tennis

Attribute: Money
Values (Rich) = [Cinema:0 , Tennis: 2, Stay in: 1, Shopping 1]
Values (Poor) = [Cinema:1 , Tennis: 0, Stay in: 0, Shopping 0]

$$GINI\ (Money_{Rich}) = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.625$$

$$GINI\ (Money_{Poor}) = 1 - \left[\left(\frac{1}{1}\right)^2\right] = 0$$

$$GINI\ (Money) = 0.625 \times \frac{4}{5} + 0 \times \frac{1}{5} = 0.5$$

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	No	Rich	Tennis
Rainy	No	Rich	Stay In
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Sunny	No	Rich	Tennis

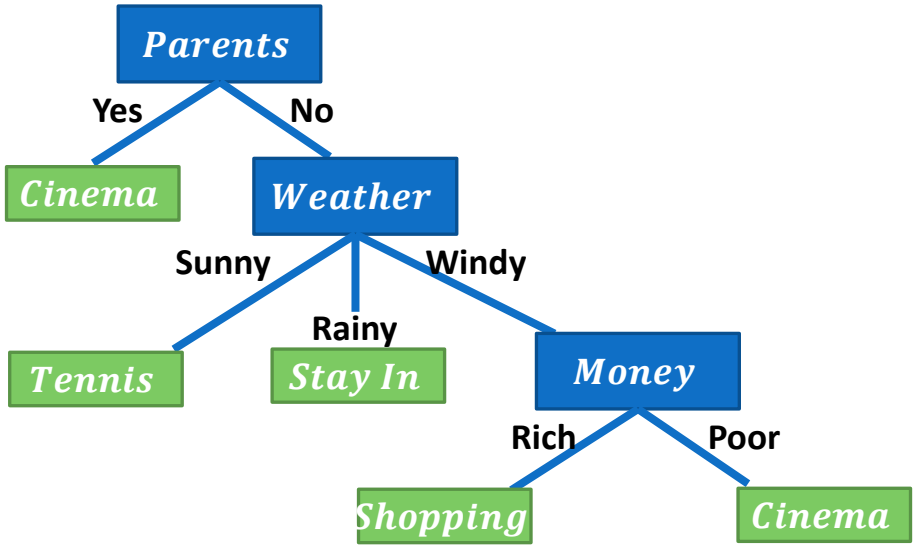
$$GINI\ (Money) = 0.5$$

$$GINI\ (Weather) = 0.2$$

Weather is selected as a node because it has smallest GINI Index

GINI Index: Numerical Example

Weather	Parents	Money	Decision
Sunny	No	Rich	Tennis
Rainy	No	Rich	Stay In
Windy	No	Poor	Cinema
Windy	No	Rich	Shopping
Sunny	No	Rich	Tennis



GINI or Entropy?

- Both can be used.
- **In my experience:**
  - The Gini criterion is much faster because it is less computationally expensive.
  - The results using the entropy criterion are slightly better.
- So it’s a trade-off between computational cost and performance.

- Gini impurity and Information Gain (Entropy) are pretty much the same and people do use the values interchangeably.
- If you are given the choice, use the Gini impurity, as it doesn't require to compute computationally expensive logarithmic functions.

Dimensions and Hyperplanes

Visualizing n −dimensions

- 1D, 2D and 3D being!

- **How transforming data from low-dimensional space to high-dimensional space is helpful?**
  - Polynomial features
  - Polynomial degree

Interesting Videos On Dimensions

- https://youtu.be/0ca4miMMaCE
- https://youtu.be/0T--WC4D1C0?t=1525

Book Reading

- Murphy – Chapter 1, Chapter 14
- Tom Mitchel (TM) – Chapter 3