Complete Decision Tree Construction: Entropy vs GINI Methods

Machine Learning Course

1 Problem Setup

We have the following dataset for predicting outdoor activity enjoyment:

Person	Weather (Outlook)	Temperature	Enjoyed?
1	Sunny	Hot	No
2	Sunny	Hot	No
3	Overcast	Hot	Yes
4	Rainy	Mild	Yes
5	Rainy	Cool	Yes
6	Rainy	Cool	No
7	Overcast	Cool	Yes
8	Sunny	Mild	Yes
9	Sunny	Cool	Yes
10	Rainy	Mild	Yes

Initial distribution: 7 Yes, 3 No

2 Method 1: Entropy and Information Gain

2.1 Step 1: Calculate Initial Entropy

The entropy formula for a node is:

$$E(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

For our root node:

- Total samples: 10
- Yes: 7, No: 3
- $p_{yes} = 7/10 = 0.7$
- $p_{no} = 3/10 = 0.3$

$$E_{initial} = -0.7 \times \log_2(0.7) - 0.3 \times \log_2(0.3)$$

$$E_{initial} = -0.7 \times (-0.5146) - 0.3 \times (-1.737) = 0.3602 + 0.5211 = 0.8813$$

2.2 Step 2: Evaluate Feature Splits

2.2.1 Option A: Split on Weather Feature

Weather has three categories: Sunny, Overcast, Rainy

Sunny branch (Persons 1,2,8,9):

- 4 instances: 2 Yes, 2 No
- $p_{yes} = 0.5$, $p_{no} = 0.5$
- $E_{sunny} = -0.5 \times \log_2(0.5) - 0.5 \times \log_2(0.5) = 1.0$

Overcast branch (Persons 3,7):

- 2 instances: 2 Yes, 0 No
- $p_{yes} = 1.0$, $p_{no} = 0$
- $E_{overcast} = -1 \times \log_2(1) - 0 \times \log_2(0) = 0$

Rainy branch (Persons 4,5,6,10):

- 4 instances: 3 Yes, 1 No
- $p_{yes} = 0.75$, $p_{no} = 0.25$
- $E_{rainy} = -0.75 \times \log_2(0.75) - 0.25 \times \log_2(0.25) = 0.8113$

Weighted average entropy after Weather split:

$$E_{after-weather} = \frac{4}{10} \times 1.0 + \frac{2}{10} \times 0 + \frac{4}{10} \times 0.8113 = 0.4 + 0 + 0.3245 = 0.7245$$

Information Gain for Weather:

$$IG_{weather} = E_{initial} - E_{after-weather} = 0.8813 - 0.7245 = 0.1568$$

2.2.2 Option B: Split on Temperature Feature

Temperature has three categories: Hot, Mild, Cool

Hot branch (Persons 1,2,3):

- 3 instances: 1 Yes, 2 No
- $p_{yes} = 1/3$, $p_{no} = 2/3$
- $E_{hot} = -\frac{1}{3} \times \log_2(\frac{1}{3}) - \frac{2}{3} \times \log_2(\frac{2}{3}) = 0.9183$

Mild branch (Persons 4,8,10):

- 3 instances: 3 Yes, 0 No
- $p_{yes} = 1.0$, $p_{no} = 0$
- $E_{mild} = 0$

Cool branch (Persons 5,6,7,9):

- 4 instances: 3 Yes, 1 No
- $p_{yes} = 0.75$, $p_{no} = 0.25$
- $E_{cool} = 0.8113$

Weighted average entropy after Temperature split:

$$E_{after-temp} = \frac{3}{10} \times 0.9183 + \frac{3}{10} \times 0 + \frac{4}{10} \times 0.8113 = 0.2755 + 0 + 0.3245 = 0.6000$$

Information Gain for Temperature:

$$IG_{temperature} = 0.8813 - 0.6000 = 0.2813$$

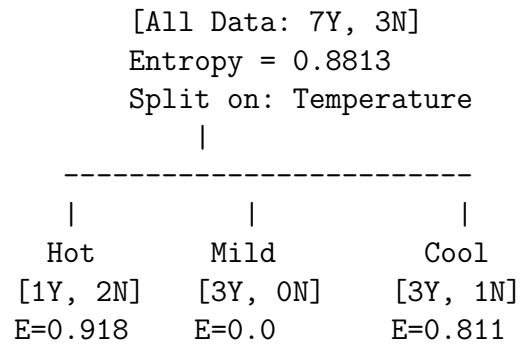
2.3 Step 3: Select Best Split

Comparing information gains:

- Weather: $IG = 0.1568$
- Temperature: $IG = 0.2813$

Result: Temperature has higher information gain, so we split on Temperature first.

2.4 Step 4: Build Tree Structure After First Split



The **Mild** branch is pure (all Yes), so it becomes a leaf node.

2.5 Step 5: Split Remaining Branches

2.5.1 Hot Branch (1Y, 2N)

Only Weather feature remains. Calculate entropy for Hot branch:

$$E_{hot} = -\frac{1}{3} \times \log_2(\frac{1}{3}) - \frac{2}{3} \times \log_2(\frac{2}{3}) = 0.9183$$

Split Hot branch on Weather:
Hot + Sunny (Persons 1,2):

- 2 instances: 0 Yes, 2 No
- $E = 0$ (pure)

Hot + Overcast (Person 3):

- 1 instance: 1 Yes, 0 No
- $E = 0$ (pure)

Hot + Rainy: No instances
Weighted entropy after split: 0 (both children pure)
Information Gain: $0.9183 - 0 = 0.9183$

2.5.2 Cool Branch (3Y, 1N)

Calculate entropy for Cool branch:

$$E_{cool} = -0.75 \times \log_2(0.75) - 0.25 \times \log_2(0.25) = 0.8113$$

Split Cool branch on Weather:
Cool + Sunny (Person 9):

- 1 instance: 1 Yes, 0 No
- $E = 0$ (pure)

Cool + Overcast (Person 7):

- 1 instance: 1 Yes, 0 No
- $E = 0$ (pure)

Cool + Rainy (Persons 5,6):

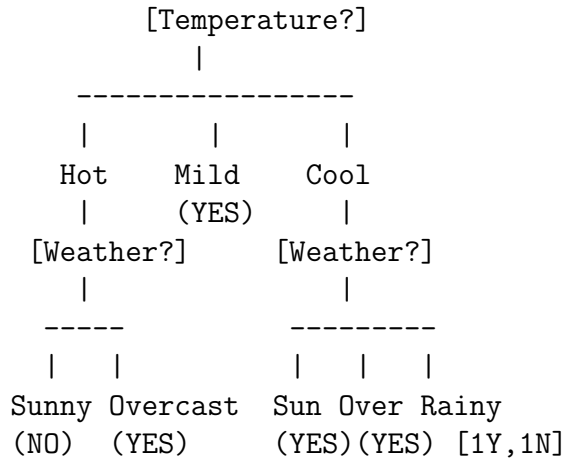
- 2 instances: 1 Yes, 1 No
- $E = 1.0$

Weighted entropy after split:

$$E_{after} = \frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{2}{4} \times 1.0 = 0.5$$

Information Gain: $0.8113 - 0.5 = 0.3113$

2.6 Step 6: Final Decision Tree using Entropy



For the mixed node **Cool + Rainy**, we predict the majority class (Yes).

3 Method 2: GINI Impurity

3.1 Step 1: Calculate Initial GINI Impurity

The GINI impurity formula is:

$$GINI(S) = 1 - \sum_{i=1}^{c} p_i^2$$

For root node:

$$GINI_{initial} = 1 - (0.7^2 + 0.3^2) = 1 - (0.49 + 0.09) = 0.42$$

3.2 Step 2: Evaluate Feature Splits using GINI

3.2.1 Option A: Split on Weather Feature

Sunny branch:

$$GINI_{sunny} = 1 - (0.5^2 + 0.5^2) = 1 - (0.25 + 0.25) = 0.5$$

Overcast branch:

$$GINI_{overcast} = 1 - (1^2 + 0^2) = 0$$

Rainy branch:

$$GINI_{rainy} = 1 - (0.75^2 + 0.25^2) = 1 - (0.5625 + 0.0625) = 0.375$$

Weighted GINI after Weather split:

$$GINI_{after-weather} = \frac{4}{10} \times 0.5 + \frac{2}{10} \times 0 + \frac{4}{10} \times 0.375 = 0.2 + 0 + 0.15 = 0.35$$

GINI Gain for Weather:

$$GG_{weather} = 0.42 - 0.35 = 0.07$$

3.2.2 Option B: Split on Temperature Feature

Hot branch:

$$GINI_{hot} = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) = 1 - \left(\frac{1}{9} + \frac{4}{9}\right) = 1 - \frac{5}{9} = 0.444$$

Mild branch:

$$GINI_{mild} = 0$$

Cool branch:

$$GINI_{cool} = 1 - (0.75^2 + 0.25^2) = 0.375$$

Weighted GINI after Temperature split:

$$GINI_{after-temp} = \frac{3}{10} \times 0.444 + \frac{3}{10} \times 0 + \frac{4}{10} \times 0.375 = 0.1332 + 0 + 0.15 = 0.2832$$

GINI Gain for Temperature:

$$GG_{temperature} = 0.42 - 0.2832 = 0.1368$$

3.3 Step 3: Select Best Split using GINI

Comparing GINI gains:

- Weather: $GG = 0.07$
- Temperature: $GG = 0.1368$

Result: Temperature has higher GINI gain, so we split on Temperature first (same as entropy method).

3.4 Step 4: Remaining Splits using GINI

3.4.1 Hot Branch (1Y, 2N)

Initial GINI: $GINI_{hot} = 0.444$

Split on Weather:

Hot + Sunny: $GINI = 0$ (pure) **Hot + Overcast**: $GINI = 0$ (pure)

Weighted GINI after split: 0

GINI Gain: $0.444 - 0 = 0.444$

3.4.2 Cool Branch (3Y, 1N)

Initial GINI: $GINI_{cool} = 0.375$

Split on Weather:

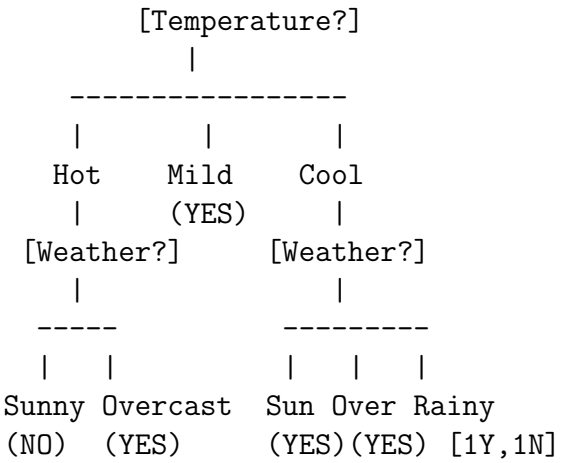
Cool + Sunny: $GINI = 0$ **Cool + Overcast**: $GINI = 0$ **Cool + Rainy**: $GINI = 1 - (0.5^2 + 0.5^2) = 0.5$

Weighted GINI after split: $\frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{2}{4} \times 0.5 = 0.25$

GINI Gain: $0.375 - 0.25 = 0.125$

3.5 Step 5: Final Decision Tree using GINI

The tree structure is identical to the entropy method:



4 Comparison and Analysis

4.1 Key Observations

1. **Same Tree Structure**: Both entropy and GINI methods produced identical tree structures for this dataset.
2. **Different Gain Values**:
   - Information Gain (Entropy): Weather = 0.1568, Temperature = 0.2813
   - GINI Gain: Weather = 0.07, Temperature = 0.1368
3. **Relative Performance**: While absolute gain values differ, the relative ordering of features remains the same - Temperature is preferred over Weather in both methods.
4. **Computational Efficiency**: GINI calculations are computationally cheaper as they don't require logarithmic computations.

4.2 Mathematical Properties

Property	**Entropy**	**GINI**
Range	$[0, \log_2 c]$	$[0, 1 - \frac{1}{c}]$
Max for binary	1.0	0.5
Computation	Requires log	Simple squares
Sensitivity	Higher	Lower

4.3 Final Prediction Rules

Regardless of the method used, the final decision rules are:

1. **IF Temperature = Mild THEN Enjoy = Yes**
2. **IF Temperature = Hot AND Weather = Sunny THEN Enjoy = No**
3. **IF Temperature = Hot AND Weather = Overcast THEN Enjoy = Yes**

4. **IF Temperature = Cool AND Weather = Sunny THEN Enjoy = Yes**

5. **IF Temperature = Cool AND Weather = Overcast THEN Enjoy = Yes**

6. **IF Temperature = Cool AND Weather = Rainy THEN Enjoy = Yes** (majority)

5 Conclusion

This complete worked example demonstrates that while entropy and GINI impurity use different mathematical foundations, they often lead to similar decision trees in practice. The choice between them typically depends on computational considerations rather than performance differences. Both methods successfully identified Temperature as the most informative feature for the initial split and produced interpretable decision rules for predicting outdoor activity enjoyment.