Classification and Regression Trees

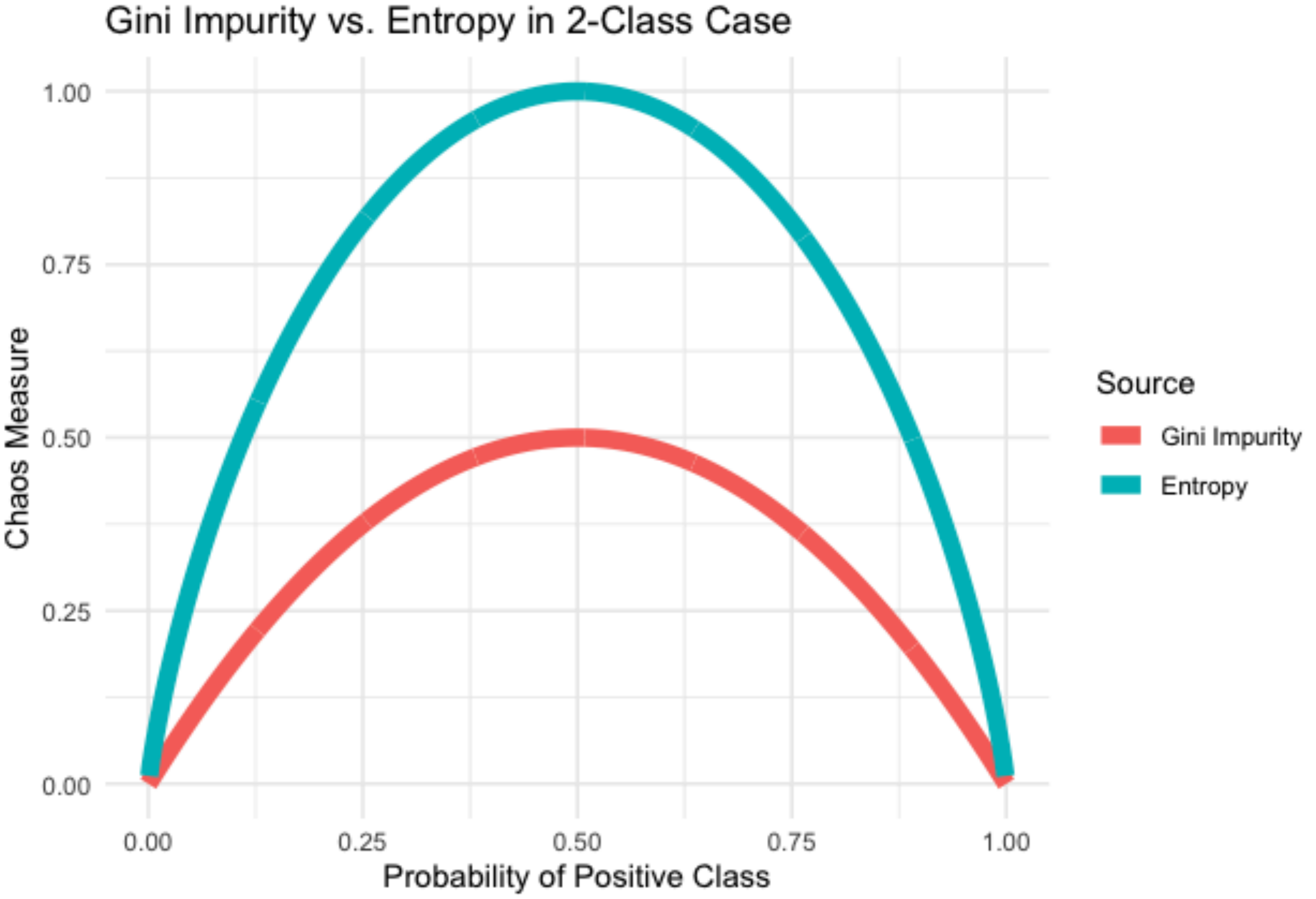
1. Start with an empty decision tree (undivided feature space)
2. Choose the ‘optimal’ predictor on which to split,
3. choose the ‘optimal’ threshold value for splitting by applying a **splitting criterion**
4. Recurse on on each new node until stopping condition is met

Splitting Criteria

Gini Index: $\mathrm{GI} = 1 - \sum_{i=1}^{n} p_i^2$

Entropy: $H = - \sum_{i}^{n} p_i \log(p_i)$

Goal: split where GI or *H* is minimzed


Gini Impurity vs. Entropy in 2-Class Case
1.00
0.75
0.50
0.25
0.00
Chaos Measure
0.00
0.25
0.50
0.75
1.00
Probability of Positive Class
Source
Gini Impurity
Entropy

Classification and Regression Trees

1. Start with an empty decision tree (undivided feature space)
2. Choose the 'optimal' predictor on which to split,
3. choose the 'optimal' threshold value for splitting by applying a **splitting criterion**
4. Recurse on on each new node until stopping condition is met

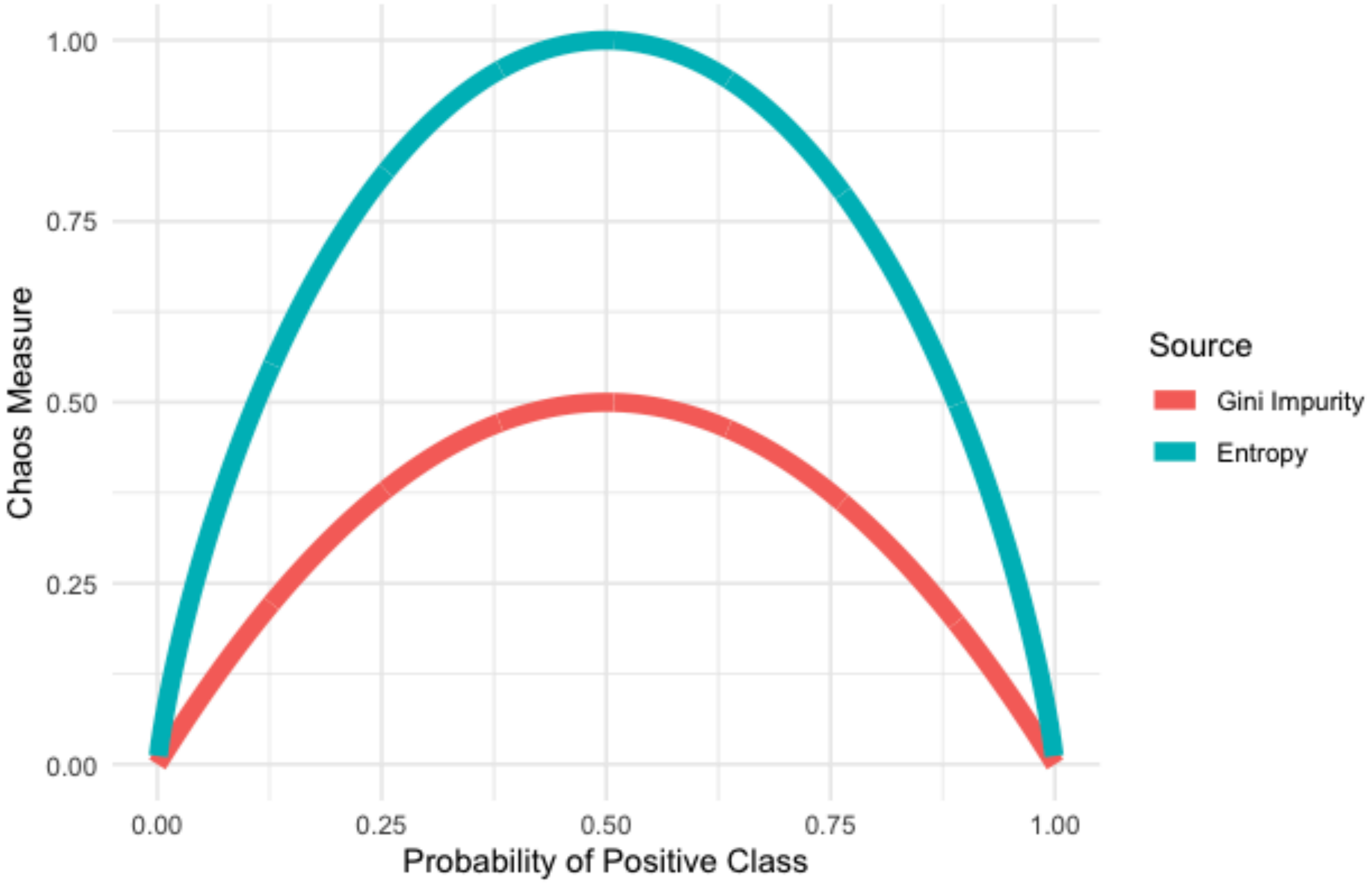
Splitting Criteria

Gini Index: $$\text{GI} = 1 - \sum_{i=1}^{n} p_i^2$$

Entropy: $$H = -\sum_{i}^{n} p_i \log(p_i)$$

Goal: split where GI or *H* is minimzed

Gini Impurity vs. Entropy in 2-Class Case



Example

cats	house	ho	children	income
1	0	1	1	34
0	1	0	1	58.3
1	1	1	0	71.5
0	0	0	1	74.9
0	0	0	1	75.3
1	0	0	1	75.6
0	0	0	1	81
1	1	1	0	82.3
1	1	1	0	85.6
1	1	1	1	95.4

$$\mathrm{GI} = 1 - \sum_{i=1}^{n} p_i^2$$