Classification Trees

Quarto

Quarto enables you to weave together content and executable code into a finished presentation. To learn more about Quarto presentations see https://quarto.org/docs/presentations/.

Bullets

When you click the **Render** button a document will be generated that includes:

- Content authored with markdown
- Output from executable code

Code

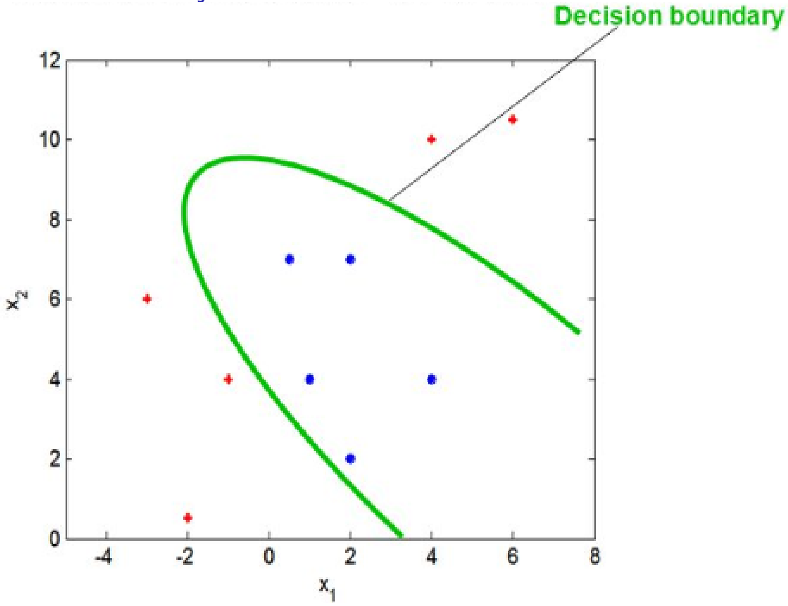
When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

Reading Materials

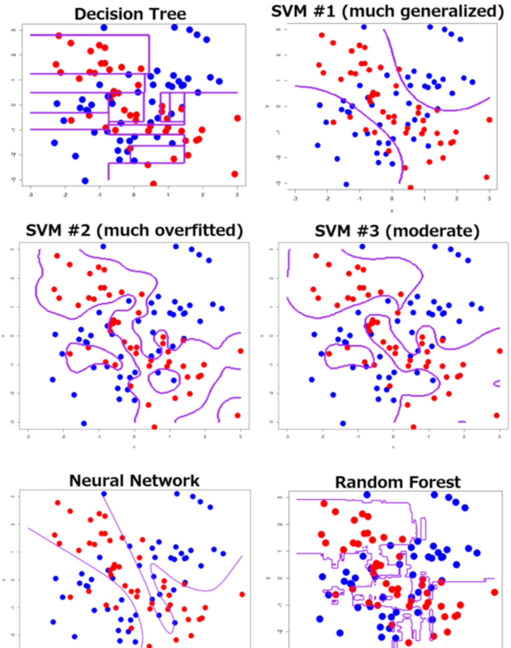
- Max Kuhn. Chapter 14. Section 14.1

Decision Boundary in Classification



Classification is a process of finding the **decision boundary** that

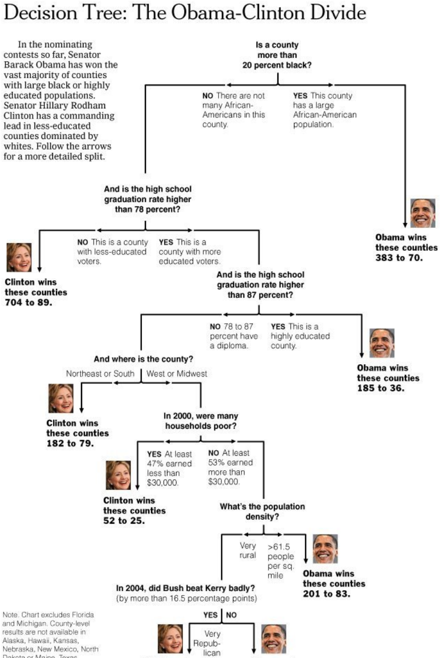
Decision Boundary in Classification



Decision Tree

- Decision Tree for classification is **Classification Tree**
- Decision Tree for Regression is **Regression Tree**

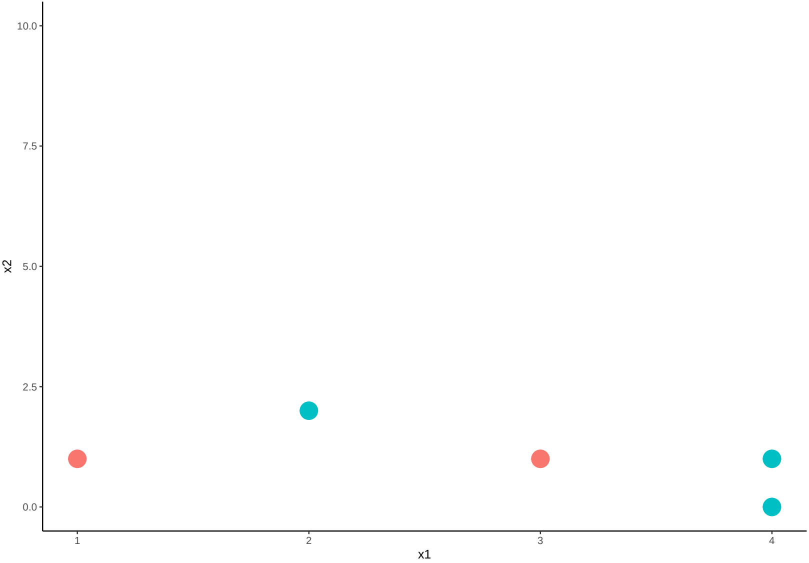
Example of Classification Tree



Classification Tree

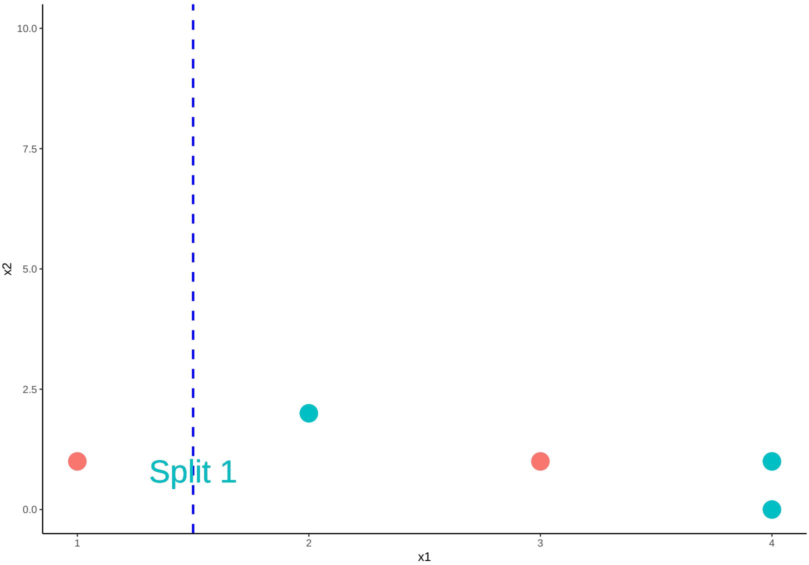
- In two dimension, classification Tree's decision boundary is a collection of horizontal and vertical line

Data

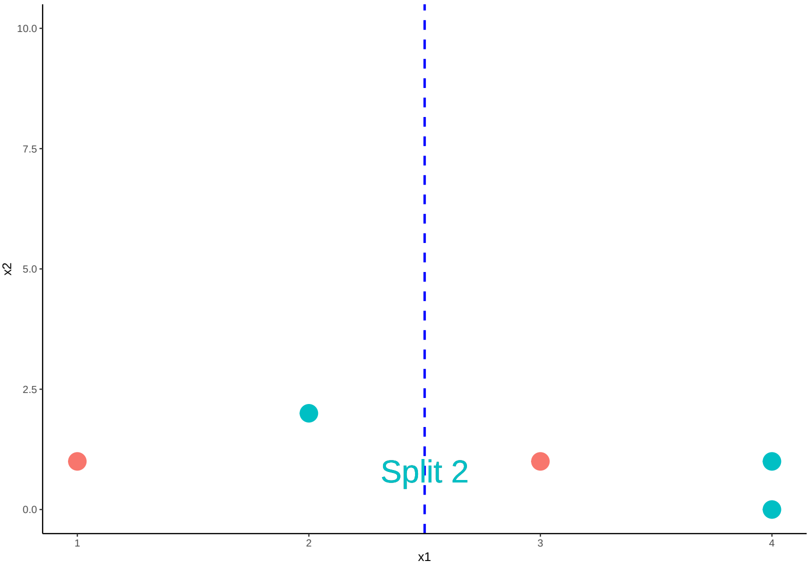


- The tree starts by a vertical or horizontal line that **best** seperate the data

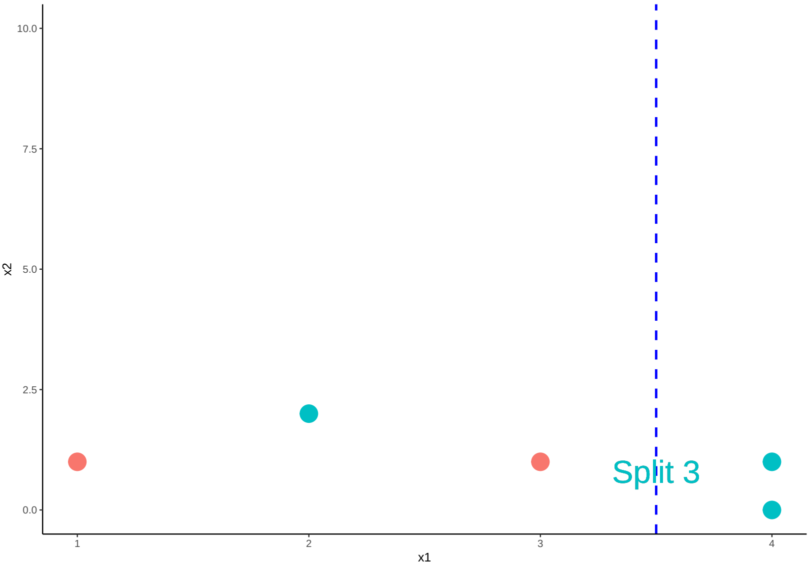
One way to seperate the reds and greens



One way to seperate the reds and greens



One way to seperate the reds and greens



Question

- **Question**: Which is the best split?

Partial Answer

- It looks like Split 1 and 3 are better than Split 2 since it misclassifies less
- Which is the better split between Split 1 and Split 3?
- We need to find a way to measure *how good a split is*

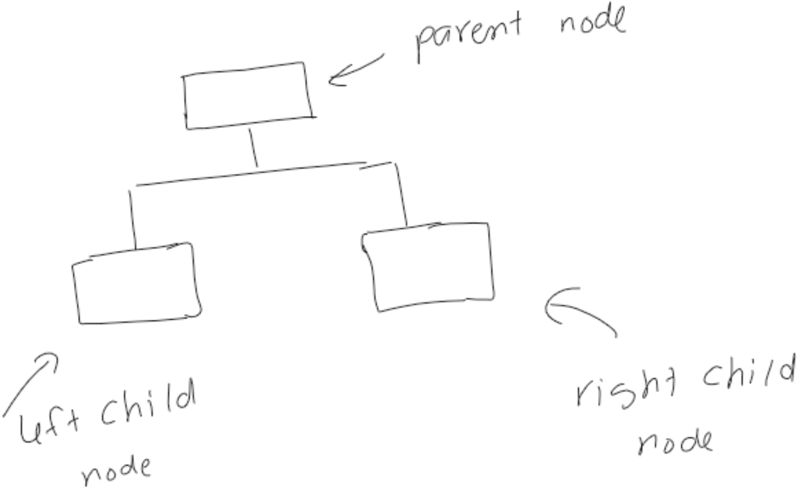
Impurity Measure

- The impurity of a node (**a node = a subset of the data or the original data**) measure how uncertain the node is.
- For example, node A with 50% reds and 50% greens would be more uncertained than node B with 90% reds and 10% greens. Thus, node A has greater impurity than node B.
- More uncertained = Greater impurity

Impurity Measure

- A split that *gains* more impurity is the **better split**!

Impurity Gain



$$IG = I_{parent} - \frac{N_{left}}{N} I_{left} - \frac{N_{right}}{N} I_{right}$$

Impurity Measure

- Impurity can be measured by: classification error, Gini Index, and Entropy.

Impurity Measure

- Let p_0 and p_1 be the proportion of class 0 and class 1 in a node.

$$
\begin{aligned}
\text{By Classification Error: } & I = min\{p_0, p_1\} \\
\text{By Gini Index: } & I = 1 - p_0^2 - p_1^2 \\
\text{By Entropy: } & I = -p_0 \log_2(p_0) - p_1 \log_2(p_1)
\end{aligned}
$$

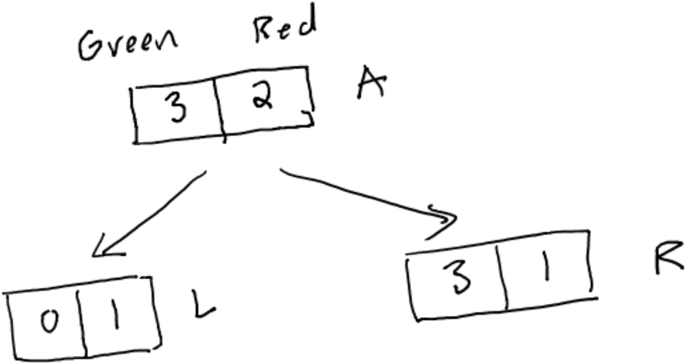
Calculation

- Let's calculate the impurity gain of the three splits to decide which split is the best

IG By Classification Error

- Let **green** and **red** be class 0 and class 1, respectively.

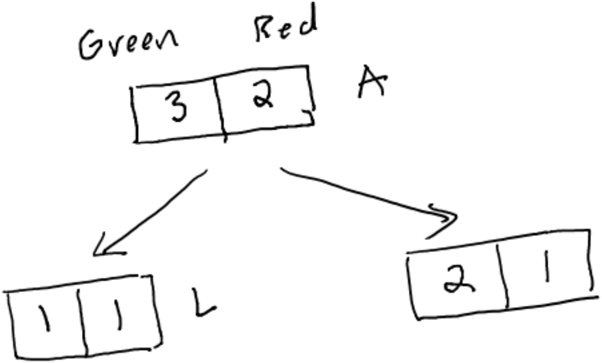
For Split 1: $N = 5, N_{left} = 1, N_{right} = 4$



- Node *parent*, A: $p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$. Thus, $I_A = \min(\frac{2}{5}, \frac{3}{5}) = \frac{2}{5}$
- Node *child left*, L: $p_0 = \frac{0}{1} = 0, p_1 = \frac{1}{1} = 1$. Thus, $I_L = \min(0, 1) = 0$
- Node *child right*, R: $p_0 = \frac{3}{\ }, p_1 = \frac{1}{\ }$. Thus,

IG By Classification Error

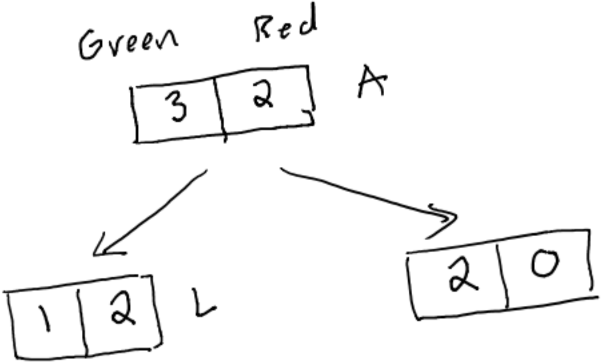
For Split 2: $N = 5, N_{left} = 2, N_{right} = 3$



- Node *parent*, A: $p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$. Thus, $I_A = \min(\frac{2}{5}, \frac{3}{5}) = \frac{2}{5}$
- Node *child left*, L: $p_0 = \frac{1}{2}, p_1 = \frac{1}{2}$. Thus, $I_L = \frac{1}{2}$
- Node *child right*, R: $p_0 = \frac{2}{3}, p_1 = \frac{1}{3}$. Thus, $I_R = \min(\frac{2}{3}, \frac{1}{3}) = \frac{1}{3}$
- Impurity Gain of Split 2:

IG By Classification Error

For Split 3: $N = 5, N_{left} = 3, N_{right} = 2$



- Node *parent*, A: $p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$. Thus, $I_A = \text{min}(\frac{2}{5}, \frac{3}{5}) = \frac{2}{5}$
- Node *child left*, L: $p_0 = \frac{1}{3}, p_1 = \frac{2}{3}$. Thus, $I_A = \text{min}(\frac{1}{3}, \frac{2}{3}) = \frac{1}{3}$
- Node *child right*, R: $p_0 = \frac{2}{2}, p_1 = \frac{0}{2}$. Thus, $I_R = \text{min}(1, 0) = 0$

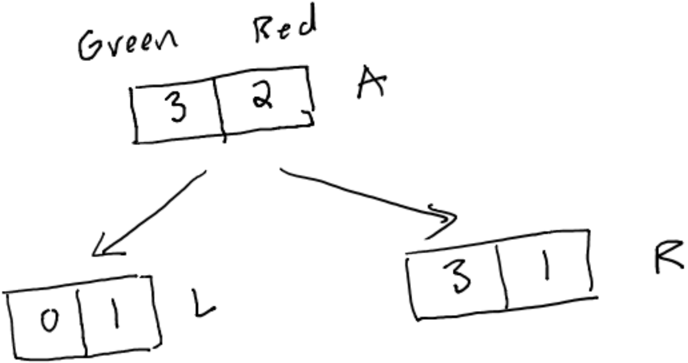
Comparing IG By Classification Error

	IG
Split 1	0.2
Split 2	0
Split 3	0.2

- By classification error, Split 1 and Split 3 are tie as the best because they have the same impurity gain.

IG By Gini Index

For Split 1: $N = 5, N_{left} = 1, N_{right} = 4$

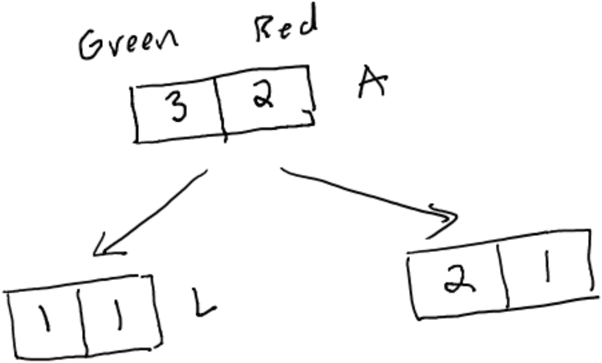


- Node *parent*, A: $p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$. Thus, $I_A = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$
- Node *child left*, L: $p_0 = \frac{0}{1} = 0, p_1 = \frac{1}{1} = 1$. Thus,

$$I_L = 1 - 0^2 - 1^2 = 0$$

IG By Gini Index

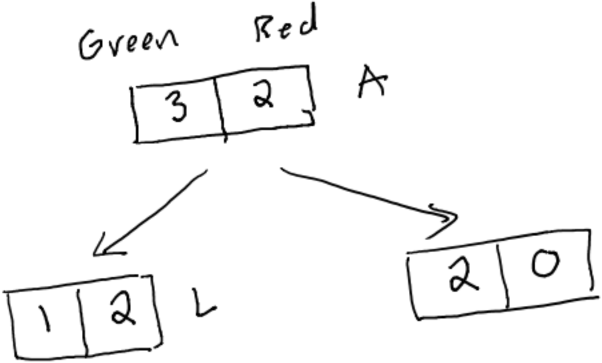
For Split 2: $N = 5, N_{left} = 2, N_{right} = 3$



- Node *parent*, A: $p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$. Thus, $I_A = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$
- Node *child left*, L: $p_0 = \frac{1}{2}, p_1 = \frac{1}{2}$. Thus, $I_L = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$
- Node *child right*, R: $p_0 = \frac{2}{3}, p_1 = \frac{1}{3}$. Thus,

IG By Gini Index

For Split 3: $N = 5, N_{left} = 3, N_{right} = 2$



- Node *parent*, A: $I_A = 0.48$
- Node *child left*, L: $p_0 = \frac{1}{3}, p_1 = \frac{2}{3}$. Thus, $I_A = 1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = 0.44$
- Node *child right*, R: $p_0 = \frac{2}{2}, p_1 = \frac{0}{2}$. Thus, $I_R = 1 - 0^2 - 1^2 = 0$

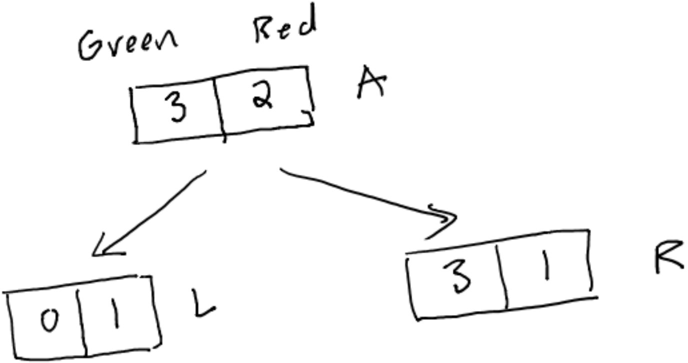
Comparing IG By Gini Index

	IG
Split 1	0.18
Split 2	0.016
Split 3	0.216

- By Gini Index, Split 3 is the best because it has the greatest impurity gain.

IG By Entropy

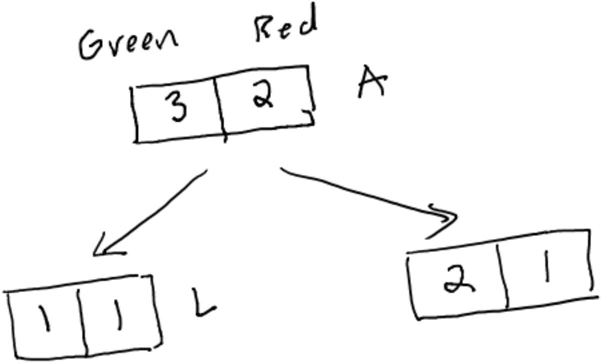
For Split 1: $N = 5, N_{left} = 1, N_{right} = 4$



- Node *parent*, A: $p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$. Thus, $I_A = -log_2(\frac{2}{5}) - log_2(\frac{3}{5}) = 0.971$
- Node *child left*, L: $p_0 = \frac{0}{1} = 0, p_1 = \frac{1}{1} = 1$. Thus, $I_L = 0$
- Node *child right*, R: $p_0 = \frac{3}{4}, p_1 = \frac{1}{4}$. Thus,

IG By Entropy

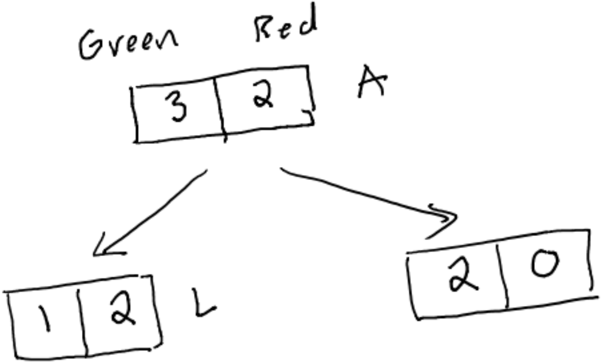
For Split 2: $N = 5, N_{left} = 2, N_{right} = 3$



- Node *parent*, A: $p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$. Thus, $I_A = 0.971$
- Node *child left*, L: $p_0 = \frac{1}{2}, p_1 = \frac{1}{2}$. Thus, $I_L = -log_1(\frac{1}{2}) - log_2(\frac{1}{2}) = 1$
- Node *child right*, R: $p_0 = \frac{2}{3}, p_1 = \frac{1}{3}$. Thus, $I_R = -log_2(\frac{2}{3}) - log_2(\frac{1}{3}) = 0.918$

IG By Entropy

For Split 3: $N = 5, N_{left} = 3, N_{right} = 2$



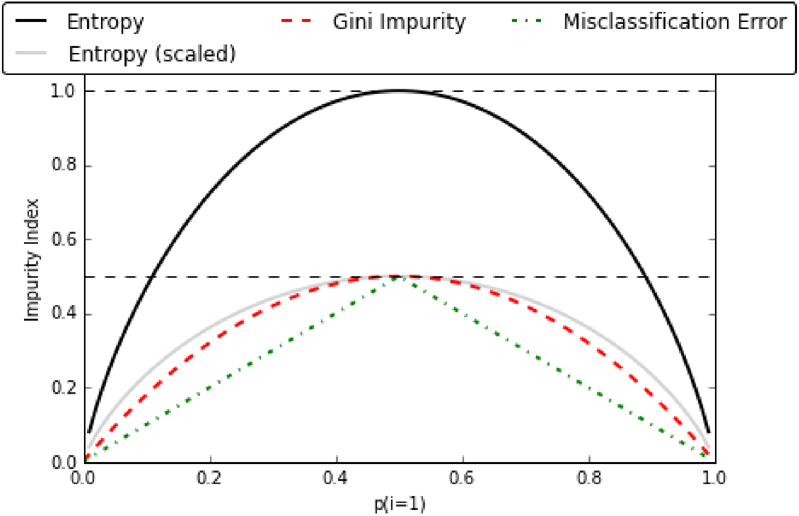
- Node *parent*, A: $I_A = 0.971$
- Node *child left*, L: $p_0 = \frac{1}{3}, p_1 = \frac{2}{3}$. Thus, $I_A = -log_2(\frac{1}{3}) - log_2(\frac{2}{3}) = 0.918$
- Node *child right*, R: $p_0 = \frac{2}{2}, p_1 = \frac{0}{2}$. Thus, $I_R = 0$
- Impurity Gain of Split 3:

Comparing IG By Entropy

	IG
Split 1	0.322
Split 2	0.02
Split 3	0.42

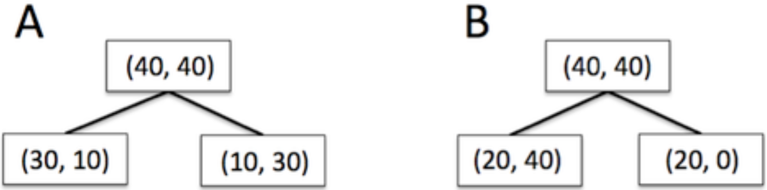
- By Gini Index, Split 3 is the best because it has the greatest impurity gain.

Comparing Impurity Measures



- Relation between impurity and the class probabilities. All impurity measures are maximized at $p_1 = 1/2$ and minimized at $p_1 = 0$ and $p_1 = 1$.

Another Example



- Which split is better?

Decide the best split using Chi-Square test of Independence

- Besides impurity gain, one can use the Chi-square, χ^2, test of independence to decide the best split.

Review of Chi-Square test of Independence

- Let X and Y be two categorical variables.
- We want to test if X and Y are independent/associated
  - H_0: X and Y are independent
  - H_α : X and Y are dependent
- Test statistic:

$$\sum \frac{(e_i - o_i)^2}{e_i} \sim \chi^2 \text{ distribution with degree of freedom} (n-1)(m-1)$$

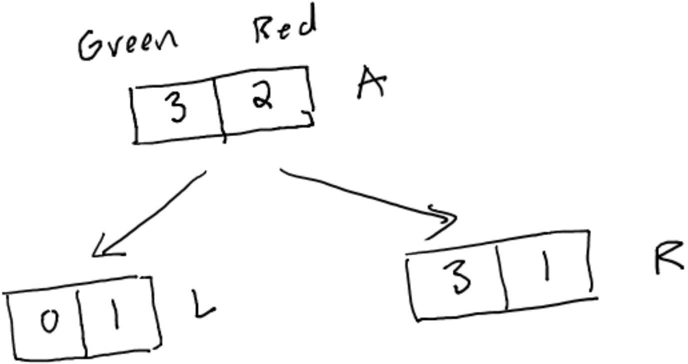
Review of Chi-Square test of Independence

- In our context, the greater the χ^2 value, the smaller the $p - value$
- The smaller the $p - value$, the more dependent the two variables are. Thus the better the split is.
- Therefore, we look for the split with the **greatest** χ^2 **value.**

Applying to Our Example

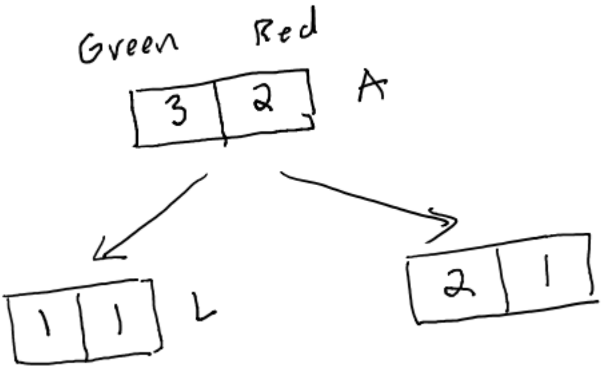
- We will calculate the χ^2 values of the three splits.
- The best split is the split with the greatest χ^2 value.

Split 1



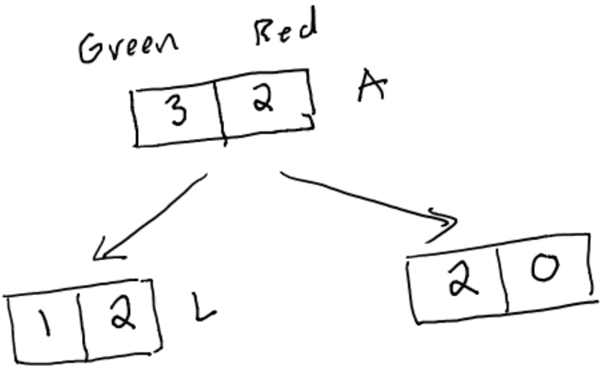
	Greens	Reds	
Left Branch	0 (Cell 1)	1 (Cell 2)	1
Right Branch	3 (Cell 3)	1 (Cell 4)	4
	3	2	

Split 2



	Greens	Reds	
Left Branch	1 (Cell 1)	1 (Cell 2)	2
Right Branch	2 (Cell 3)	1 (Cell 4)	3
	3	2	

Split 3



	Greens	Reds	
Left Branch	1 (Cell 1)	2 (Cell 2)	3
Right Branch	2 (Cell 3)	0 (Cell 4)	2
	3	2	

Comparing the three splits

	χ^2
Split 1	1.875
Split 2	0.139
Split 3	2.222

- Split 3 is the best because it has the greatest χ^2!