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

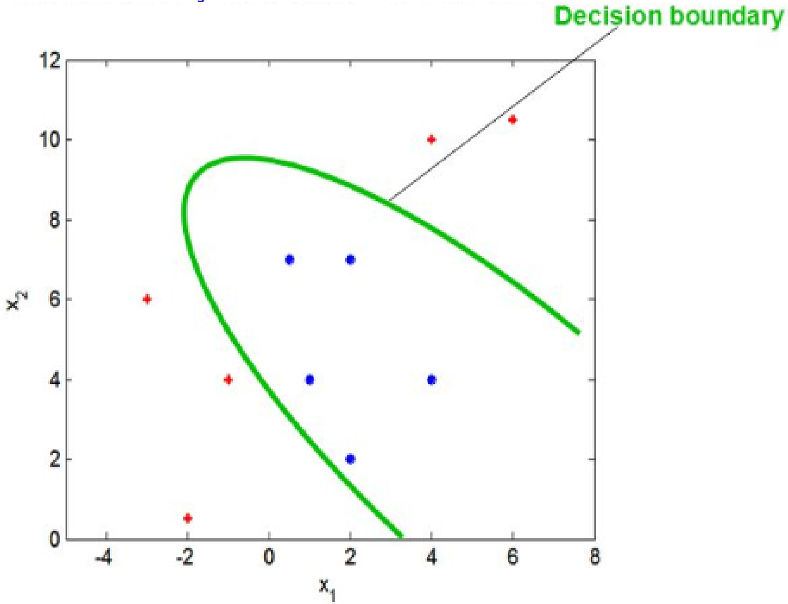
When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

Reading Materials

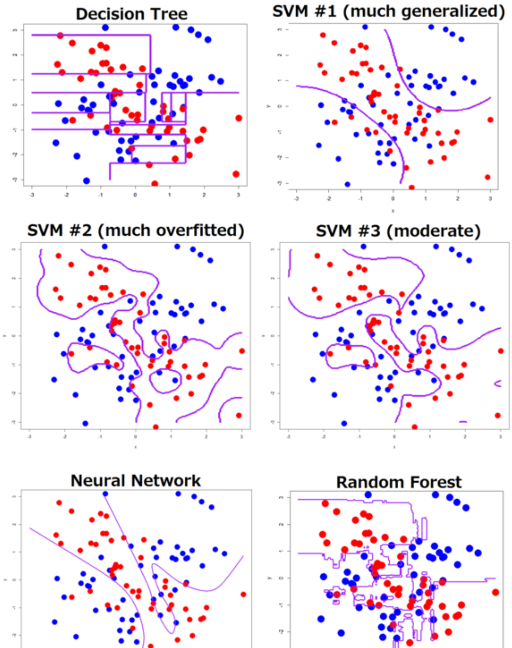
- Max Kuhn. Chapter 14. Section 14.1

Decision Boundary in Classification



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

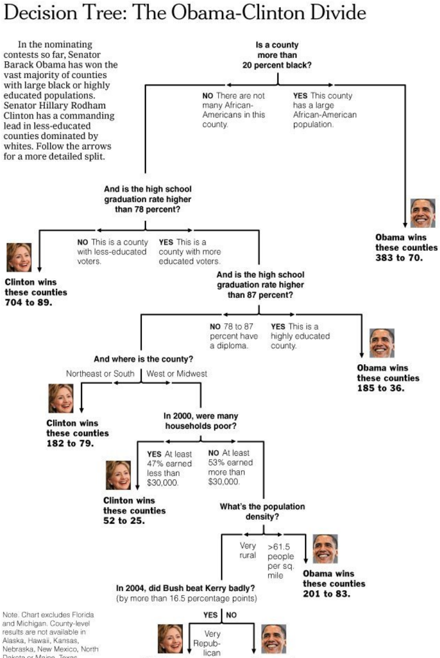
Decision Boundary in Classification



Decision Tree

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

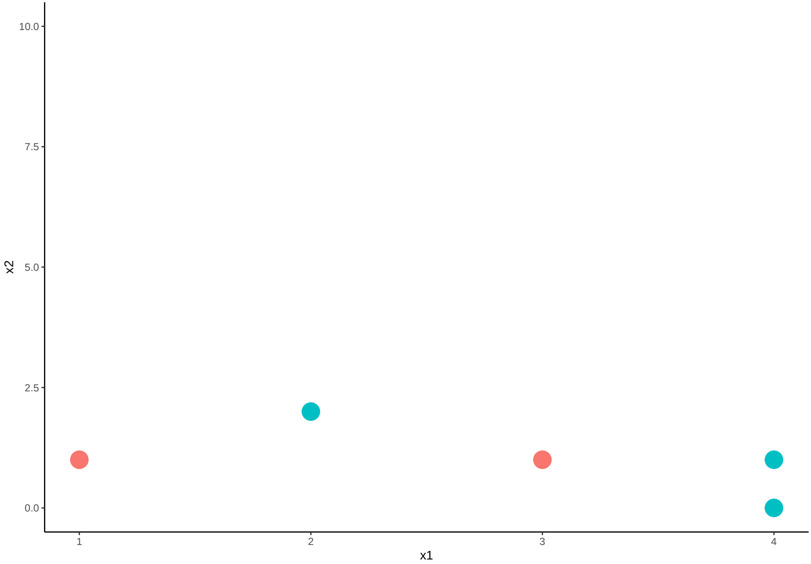
Example of Classification Tree



Classification Tree

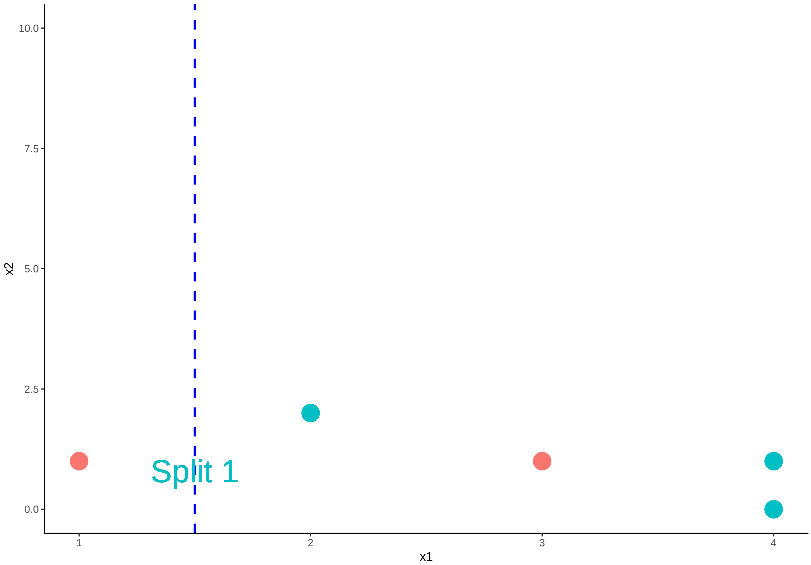
- In two dimension, classification Tree's decision boundary is a collection of horizontal and vertical line

Data

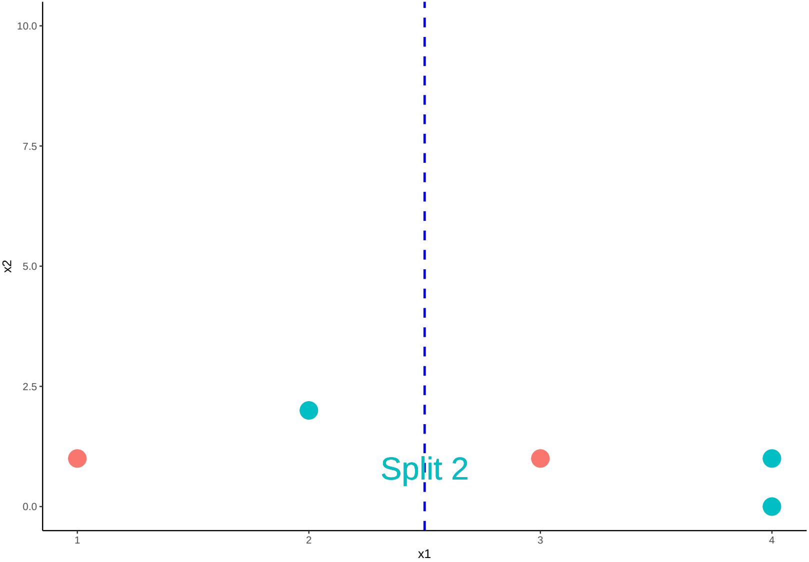


- The tree starts by a vertical or horizontal line that **best** seperate the data

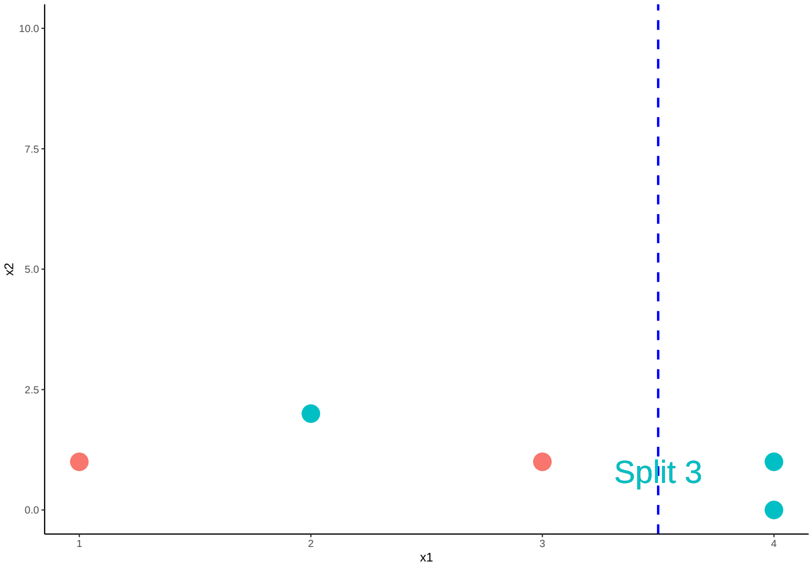
One way to seperate the reds and greens



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*

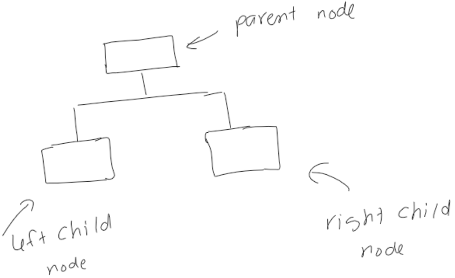
Impurity Measure

- 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



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

- IG is Impurity Gain of the split
- N_{left} and N_{right} are the number of points in the left child node and right child node, respectively.
- $N_{left} + N_{right} = N$

Impurity Measure

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

Impurity Measure

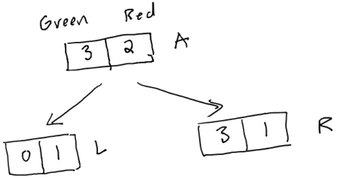
- Let p_0 and p_1 be the proportion of class 0 and class 1 in a node.

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

Calculation

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

IG By Classification Error



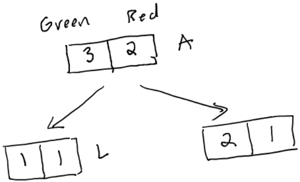
- Let **green** and **red** be class 0 and class 1, respectively.

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: $p_0 = \frac{3}{4}, p_1 = \frac{1}{4}$. Thus, $I_R = \min(\frac{3}{4}, \frac{1}{4}) = \frac{1}{4}$
- Impurity Gain of Split 1:

$$IG = \frac{2}{5} - \frac{1}{5} \cdot 0 - \frac{4}{5} \cdot \frac{1}{4} = 0.2$$

IG By Classification Error



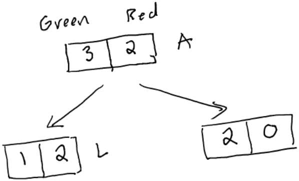
For Split 2: $N = 5, N_{left} = 2, N_{right} = 3$

- 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:

$$IG = \frac{2}{5} - \frac{2}{5} \cdot \frac{1}{2} - \frac{3}{5} \cdot \frac{1}{3} = 0$$

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

$$IG = \frac{2}{5} - \frac{3}{5} \cdot \frac{1}{3} - \frac{2}{5} \cdot 0 = 0.2$$

Comparing IG By Classification Error

Split	IG
1	0.2
2	0
3	0.2

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