

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

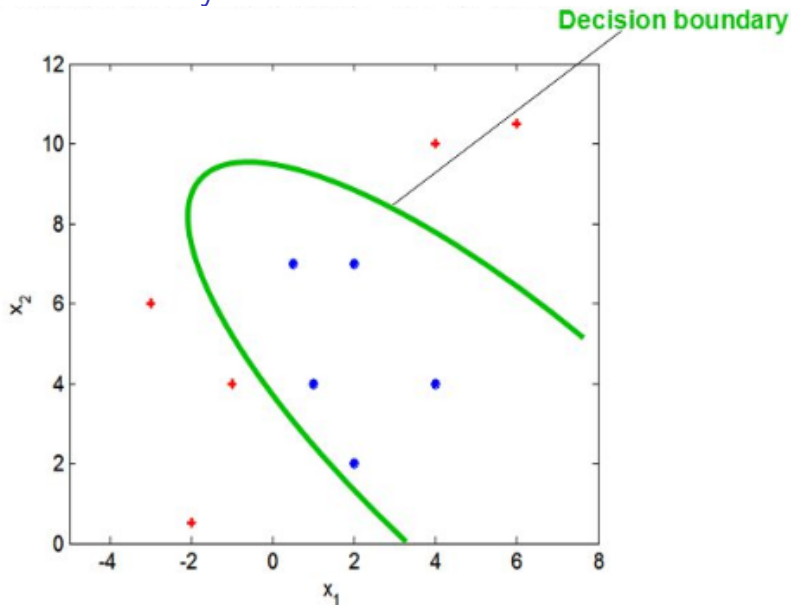
When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

Reading Materials

- ▶ Max Kuhn. Chapter 14. Section 14.1

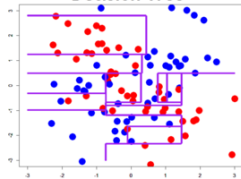
Decision Boundary in Classification



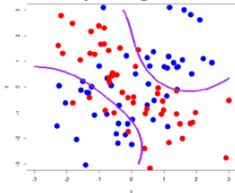
Classification is a process of finding the **decision boundary** that

Decision Boundary in Classification

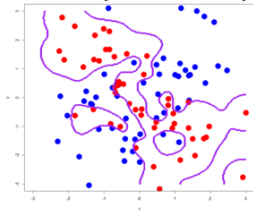
Decision Tree



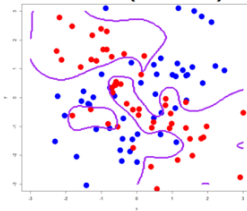
SVM #1 (much generalized)



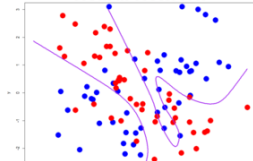
SVM #2 (much overfitted)



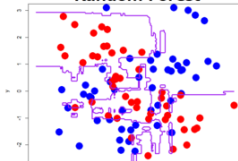
SVM #3 (moderate)



Neural Network



Random Forest



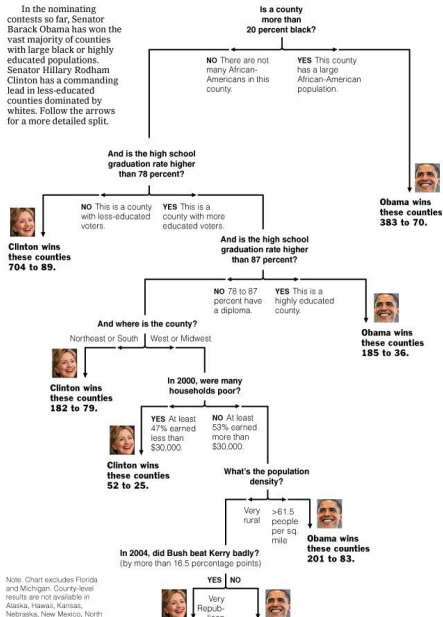
Decision Tree

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

Example of Classification Tree

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.

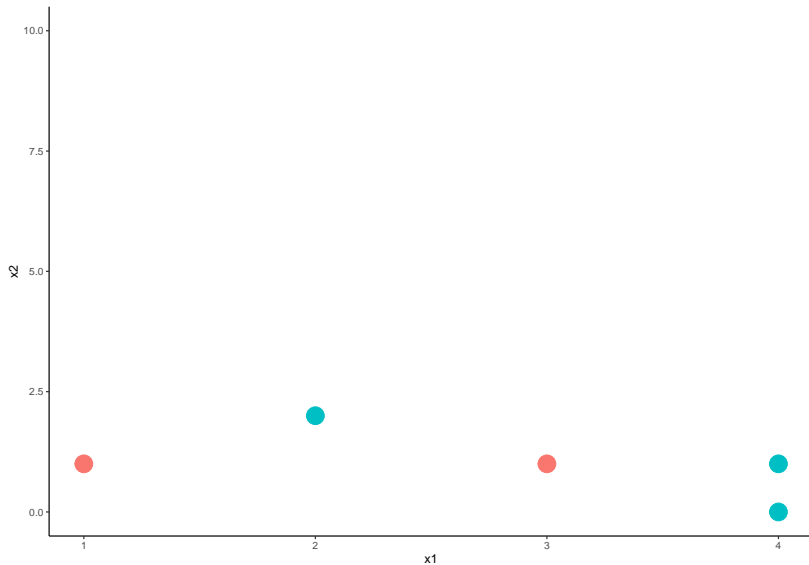


Note: Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota, or Maine.

Classification Tree

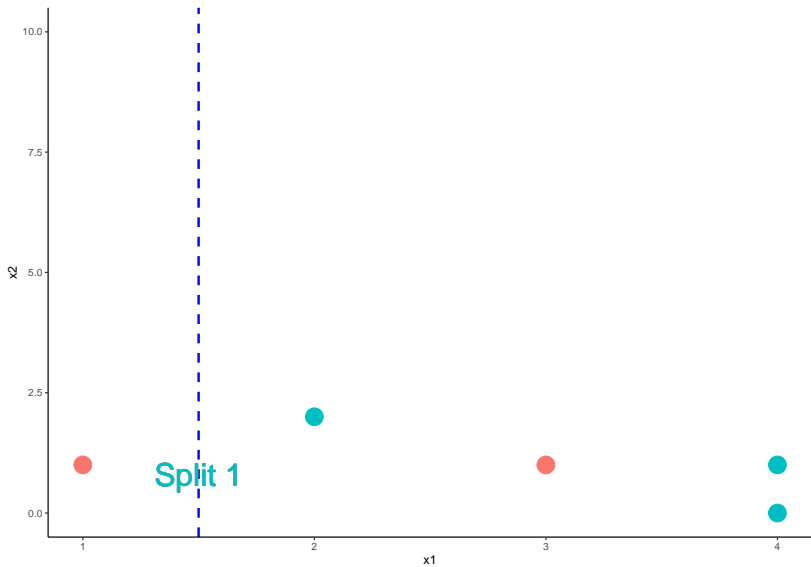
- ▶ In two dimension, classification Tree's decision boundary is a collection of horizontal and vertical line

Data

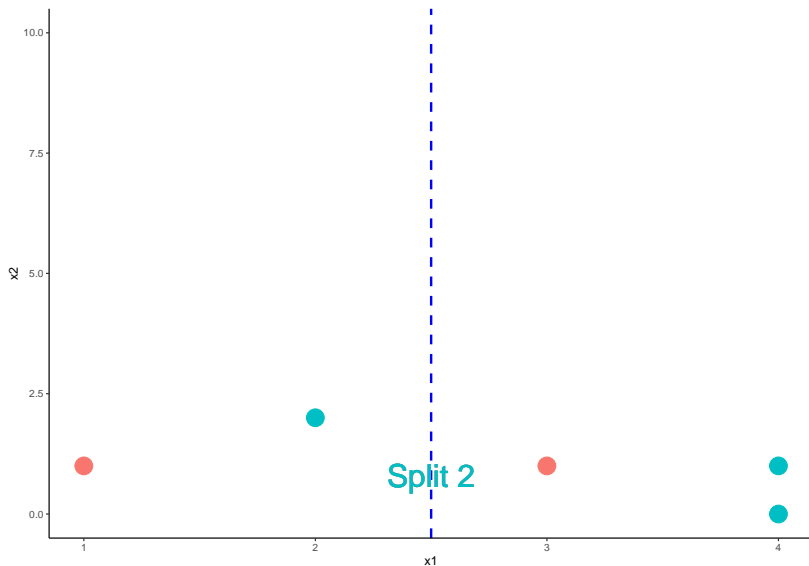


- ▶ The tree starts by a vertical or horizontal line that **best** separate the data

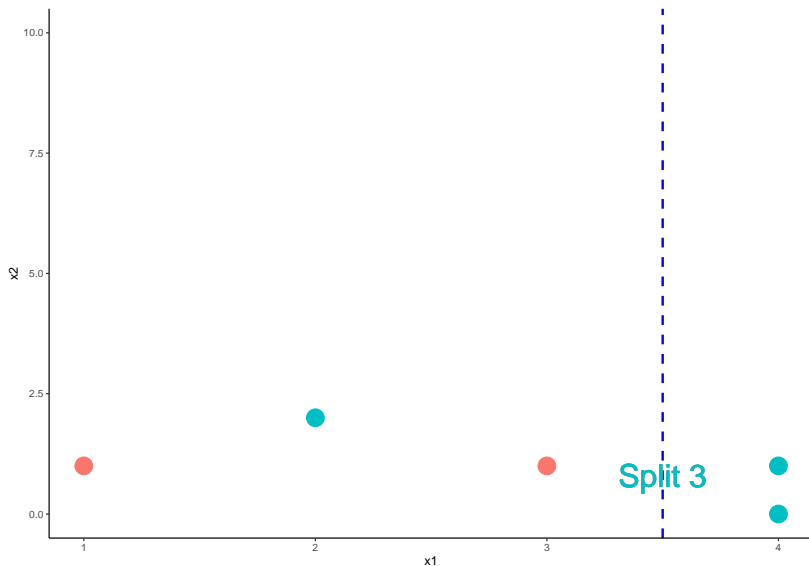
One way to separate the reds and greens



One way to separate the reds and greens



One way to separate 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*

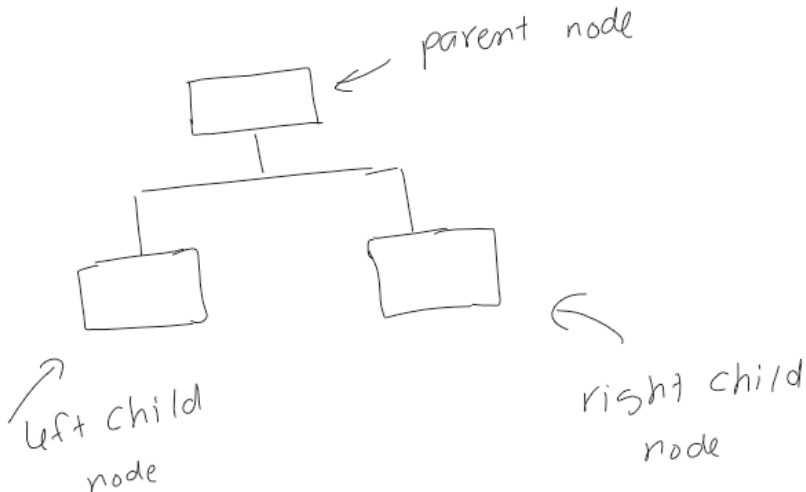
Impurity Measure

- ▶ 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 uncertain than node B with 90% reds and 10% greens. Thus, node A has greater impurity than node B.
- ▶ More uncertain = 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.

By Classification Error: $I = \min\{p_0, p_1\}$

By Gini Index: $I = 1 - p_0^2 - p_1^2$

By Entropy: $I = -p_0 \log_2(p_0) - p_1 \log_2(p_1)$