Decision Trees

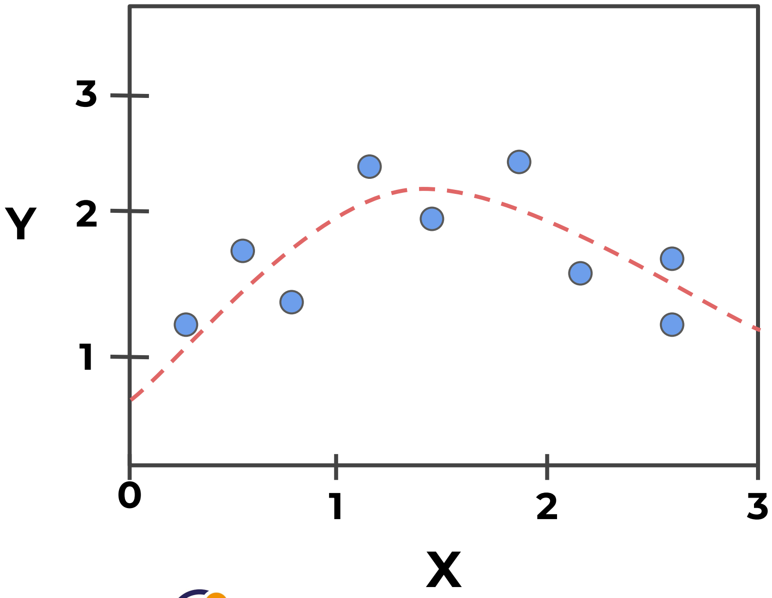
Regression Trees

Tree Based Methods

- Rely on the ability to **split** data based on **information** from features.
- This means we need a mathematical definition of **information** and the ability to measure it.

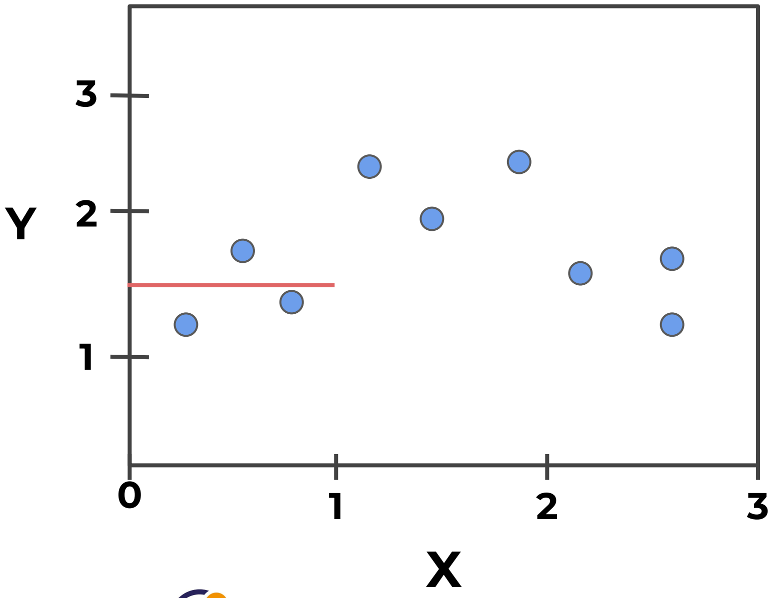
Tree Based Methods

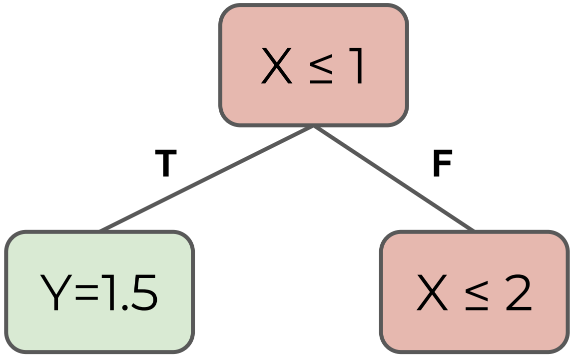
- 1963: Piecewise-constant regression tree



Tree Based Methods

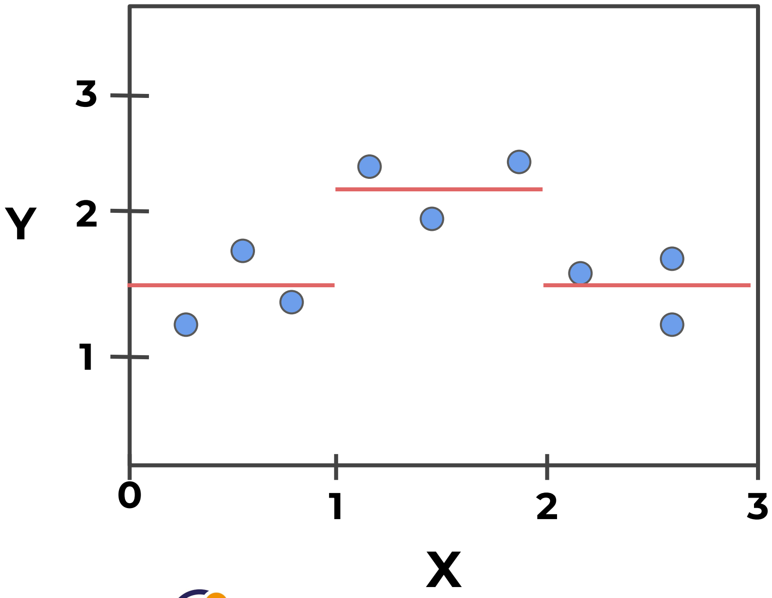
- 1963: Piecewise-constant regression tree

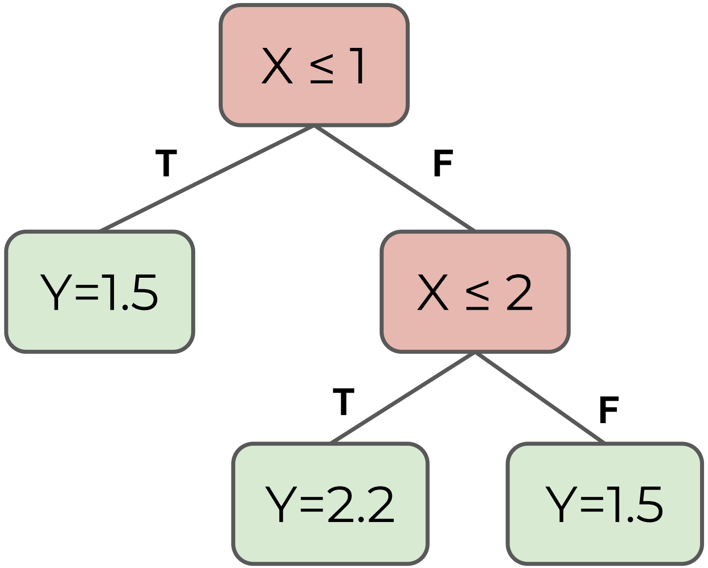




Tree Based Methods

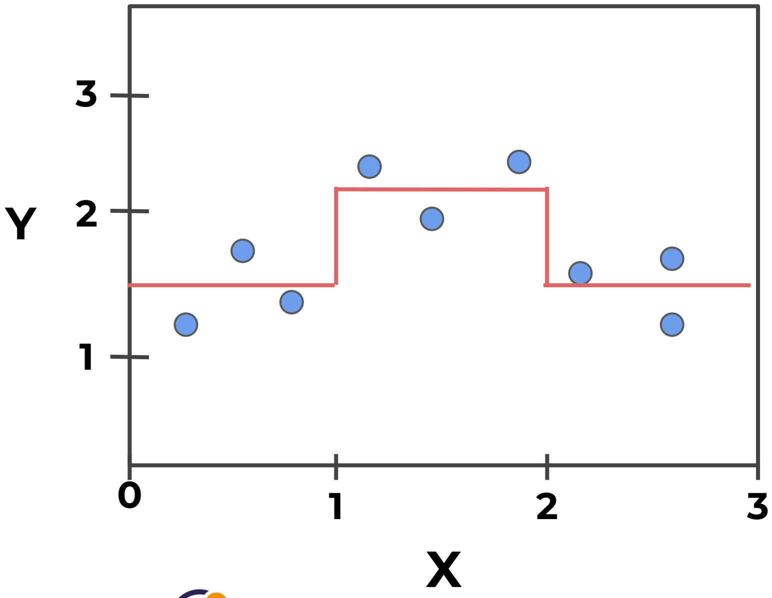
- 1963: Piecewise-constant regression tree

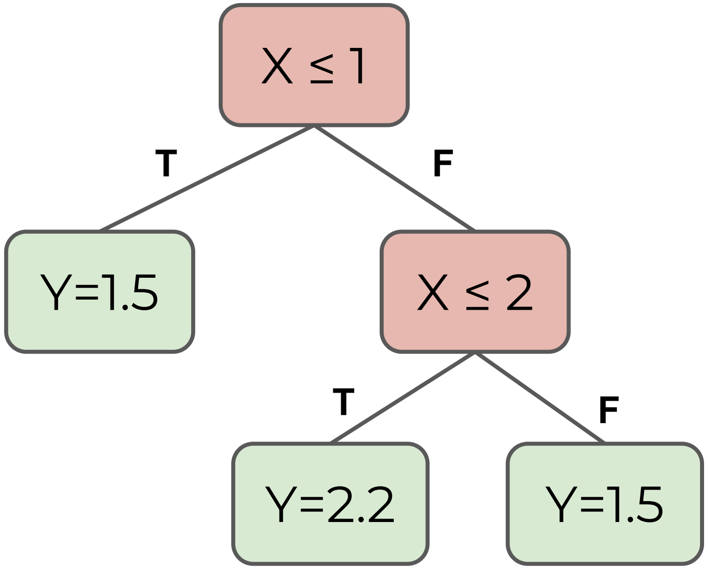




Tree Based Methods

- 1963: Piecewise-constant regression tree





Decision Trees

Gini Impurity

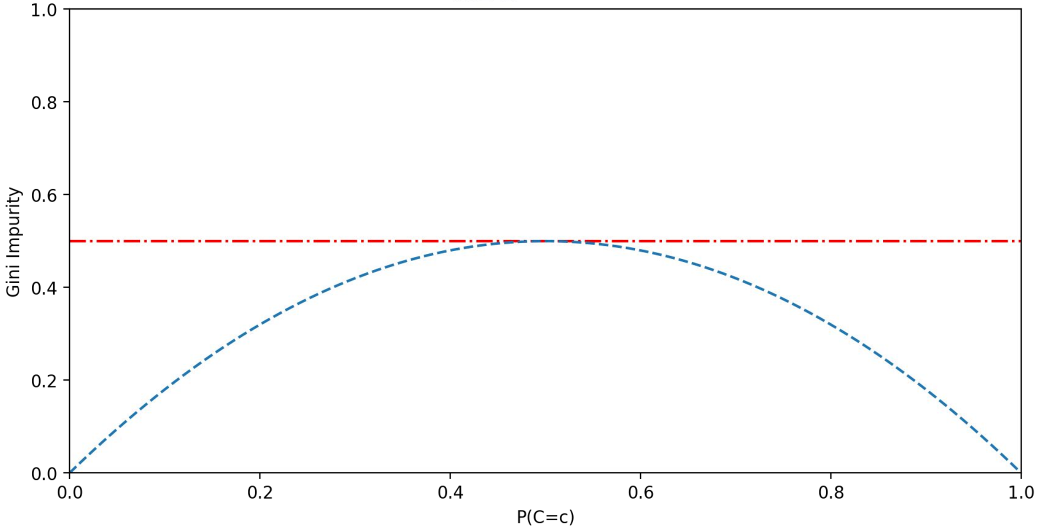
Gini Impurity

- **Gini impurity** is a mathematical measurement of how “pure” the information in a data set is.
- In regards to classification, we can think of this as a measurement of class uniformity.

Gini Impurity

- Gini Impurity for Classification:

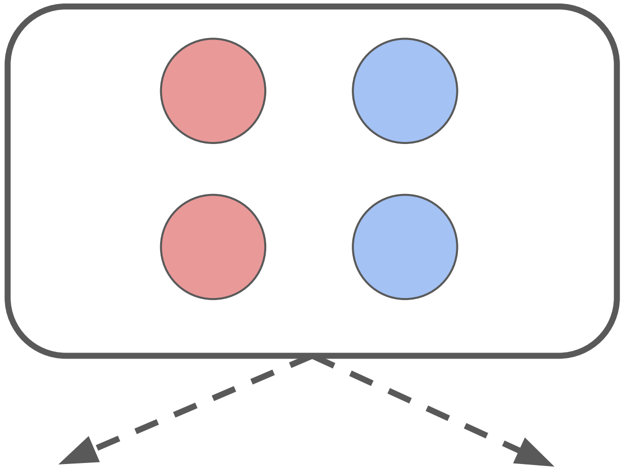
$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Gini Impurity

- Gini Impurity for Classification:

$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Class Red
(2/4)(1 - 2/4) = 0.25

+

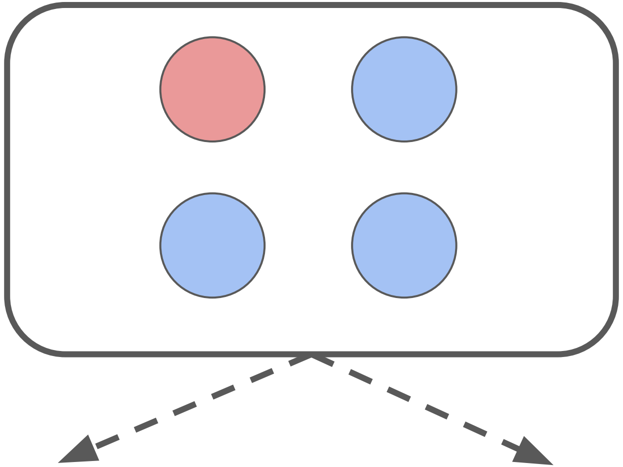
Class Blue
(2/4)(1 - 2/4) = 0.25

Gini Impurity
0.25 + 0.25 = 0.5

Gini Impurity

- Data is more “pure” (less impurity)

$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Class Red
(1/4)(1 - 1/4) = 0.1875

+

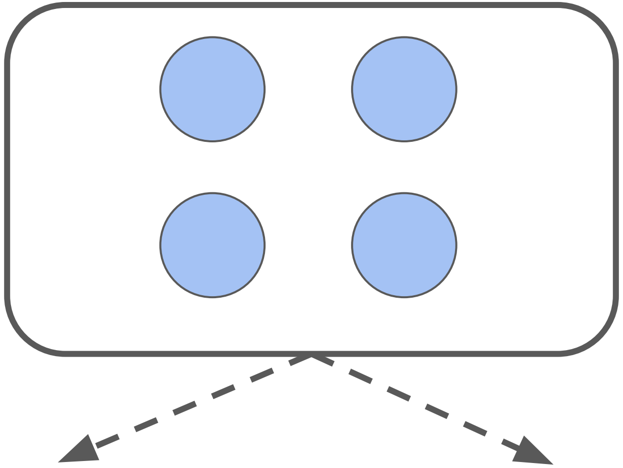
Class Blue (3/4)(1 - 3/4) =0.1875

Gini Impurity
0.1875+0.1875 = 0.375

Gini Impurity

- Data is completely "pure" (no impurity)

$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



Class Red
(0/4)(1 - 0/4) = 0

+

Class Blue
(4/4)(1 - 4/4) = 0

Gini Impurity
0 + 0 = 0

Gini Impurity

- If the goal of a decision tree is to separate out classes, we can use **gini impurity** to decide on data split values.
- We want to **minimize** the gini impurity at leaf nodes.
- Minimized impurity at leaf nodes means we are separating classes effectively!

Decision Trees

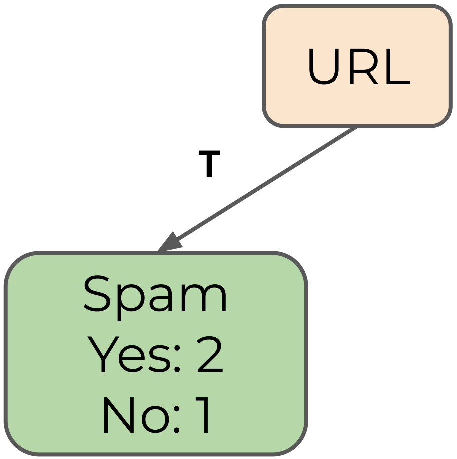
- Create a decision tree to predict spam.

X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No

Decision Trees

- Predict if email is spam if it contains a URL:

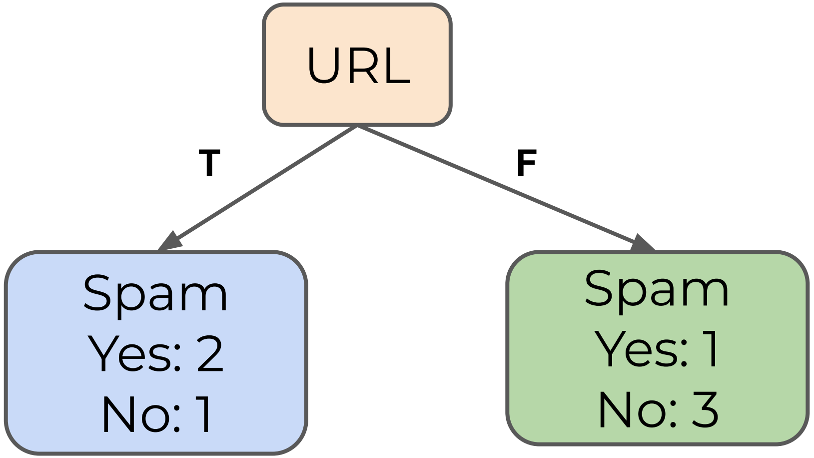
X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No



Decision Trees

- Predict if email is spam if it contains a URL:

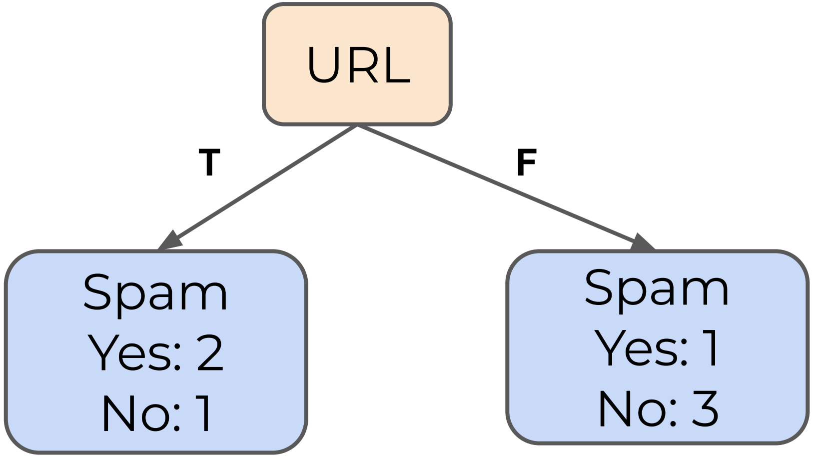
X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No



Decision Trees

- Recall the gini impurity formula:

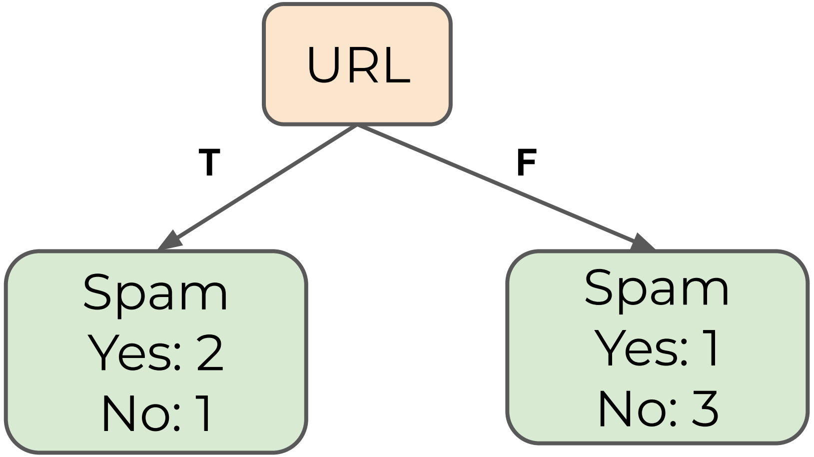
X - URL Link	Y-Spam
Yes	Yes
Yes	Yes
No	No
No	No
No	Yes
No	No
Yes	No



$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Decision Trees

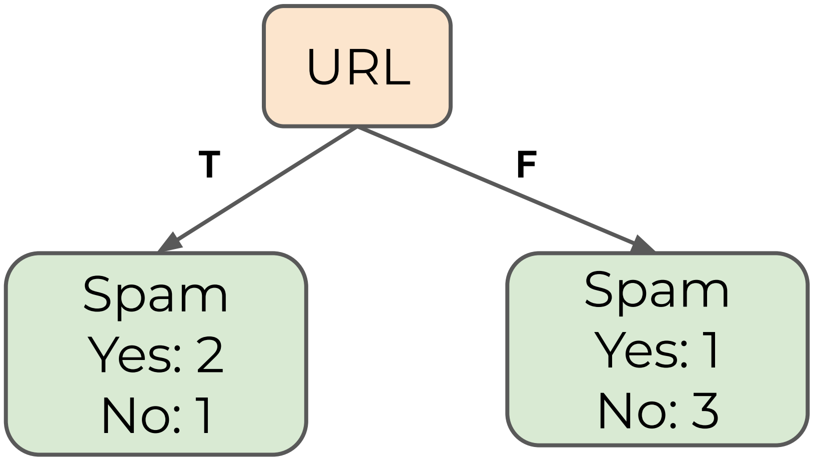
- Treat Yes Spam and No Spam as **c** classes:
- Left Leaf Node:
  - $(2/3)(1-2/3) + (1/3)(1-1/3)$
  - Left Leaf Gini=0.44
- Right Leaf Node:
  - $(1/4)(1-1/4) + (3/4)(1-3/4)$
  - Right Leaf Gini=0.375



$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Decision Trees

- Total Emails: (2+1) + (1+3) = 7
- Left Leaf Gini=0.44
- Right Leaf Gini=0.375
- Left Emails: 3
- Right Emails: 4
- (3/7)*0.44 + (4/7)*0.375
- Gini Impurity: 0.403



$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$

Decision Trees

Continuous Features: Binning

Decision Trees

- Let's calculate the feature gini impurity:

X - Words in Email	Y-Spam
10	Yes
40	No
20	Yes
50	No
30	No

Decision Trees

- First sort data:

X - Words in Email	Y-Spam
10	Yes
40	No
20	Yes
50	No
30	No

Decision Trees

- First sort data:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

Decision Trees

* Calculate potential split values for node:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

Decision Trees

- Use averages between rows as values:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

15

25

35

45

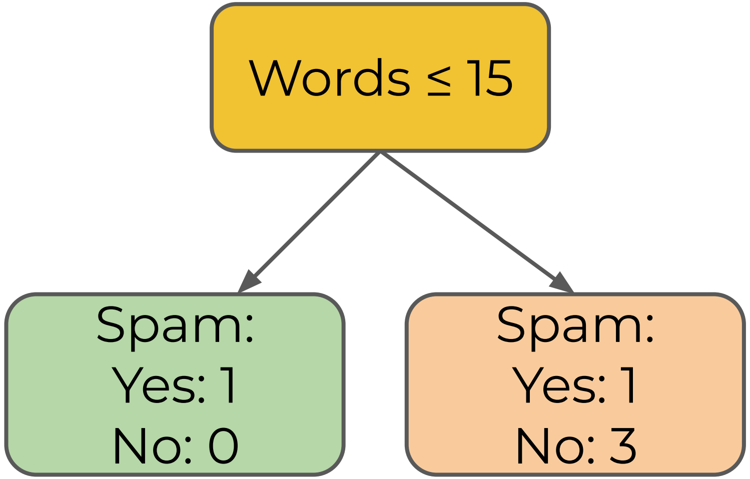
Words ≤ N

Decision Trees

- Calculate gini impurity for each split:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

$$G(Q) = \sum_{c \in C} p_c(1 - p_c)$$



$$G(Q) = (1/5)(0+0) + (4/5)((1/4)(1-1/4)+(3/4)(1-3/4)$$

$$= 0.3$$

Decision Trees

- Repeat for all possible splits:

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

- 15 → Gini=0.3
- 25 → Gini=0
- 35 → Gini=0.26
- 45 → Gini=0.4

Decision Trees

- Choose lowest impurity split value

X - Words in Email	Y-Spam
10	Yes
20	Yes
30	No
40	No
50	No

25

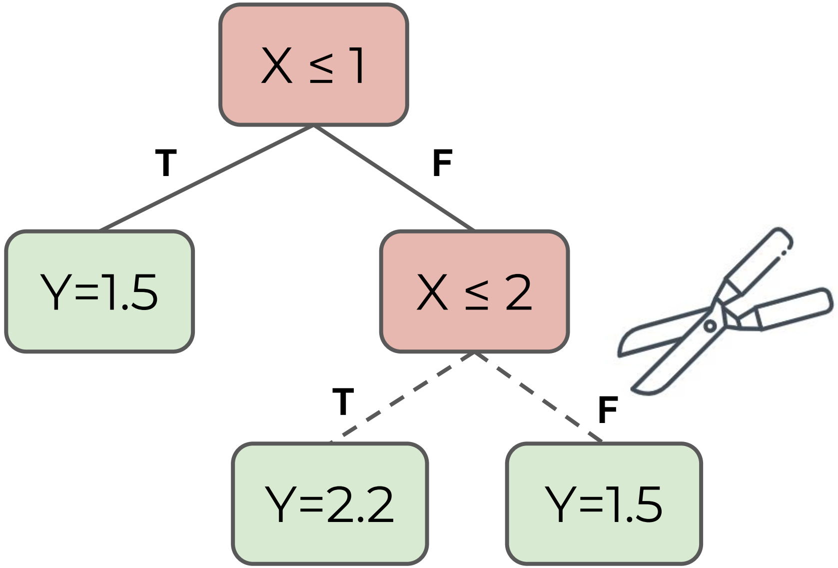
Gini=0

Decision Trees

Dealing with overfitting: pruning (automatic) and maximum depth (hyperparameter)

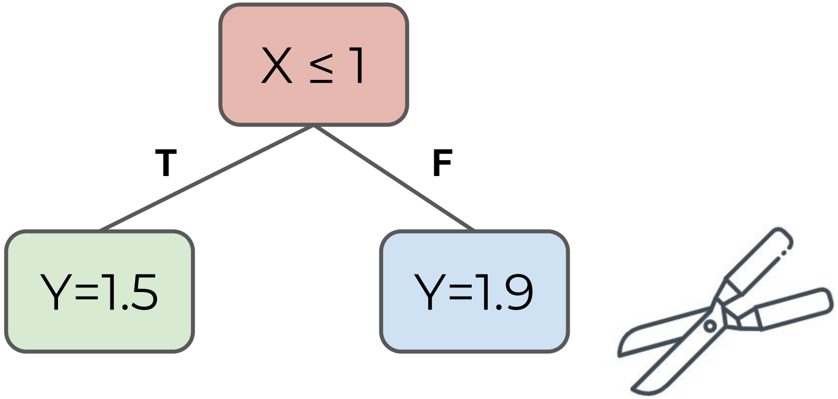
Decision Trees

- Pruning:



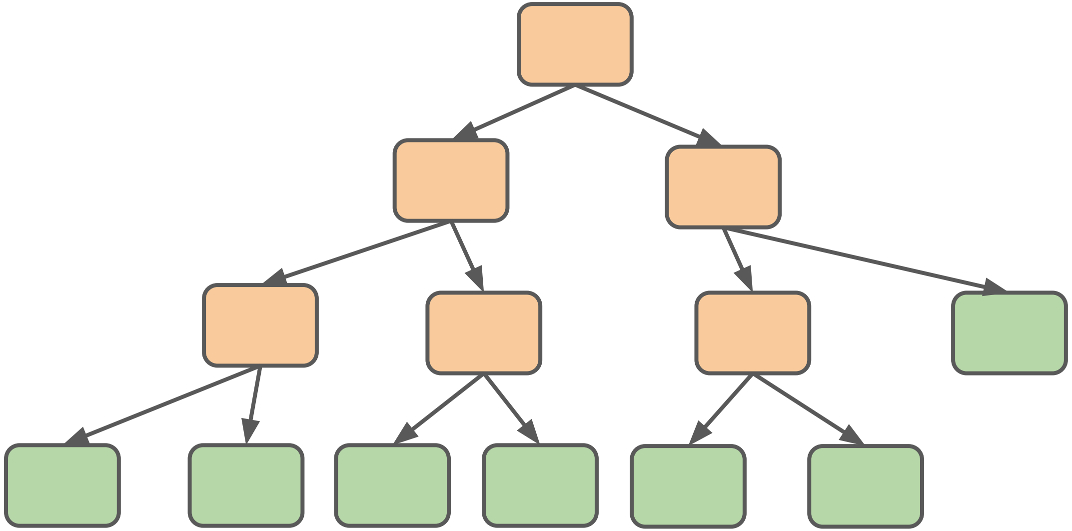
Decision Trees

- Pruning:



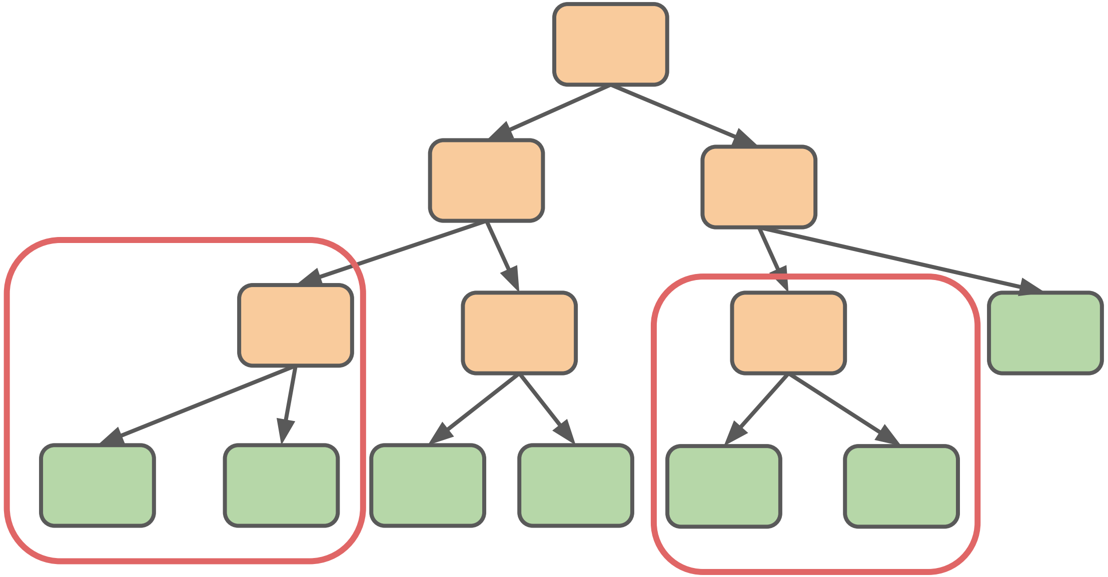
Decision Trees

- A large overfitted tree:



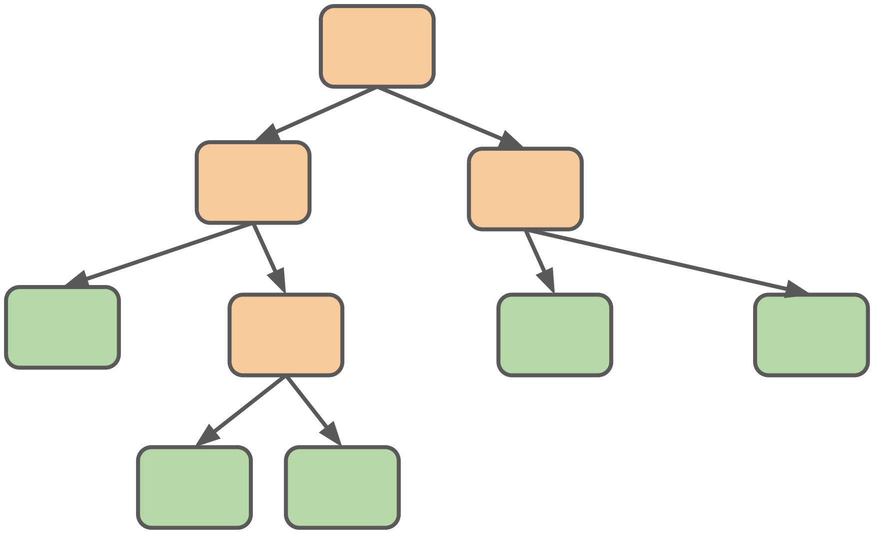
Decision Trees

- Pruning: minimum gini impurity



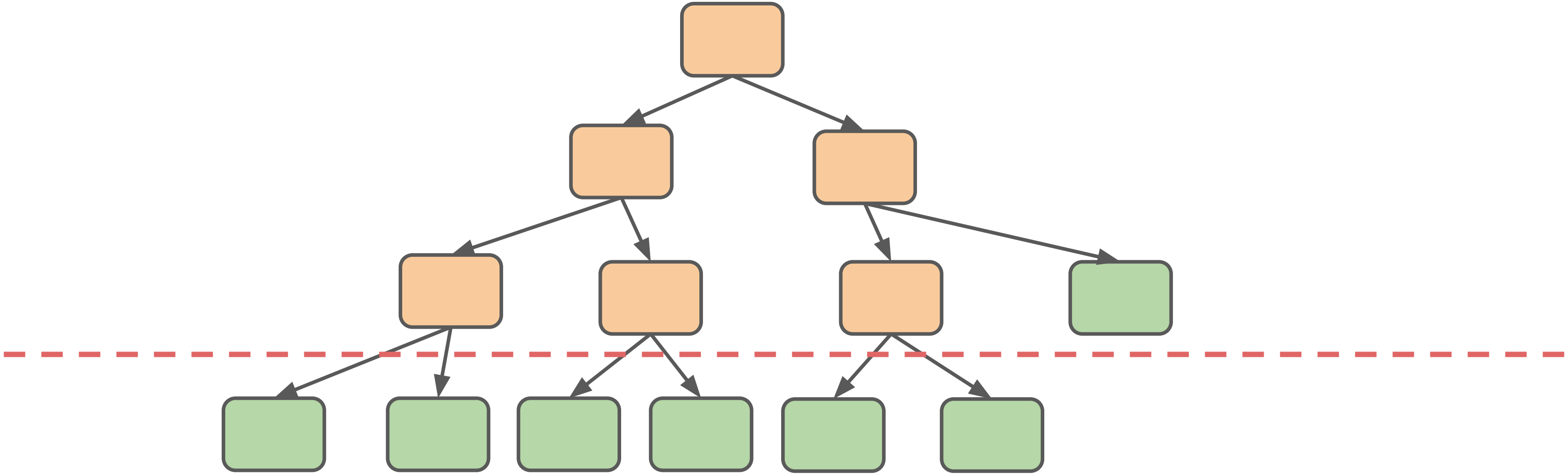
Decision Trees

- Pruning: minimum gini impurity



Decision Trees

- We can also mandate a max depth:



PIERIAN DATA

Decision Trees

- We can also mandate a max depth: