

Classification Trees

Classification Trees

author: Son Nguyen

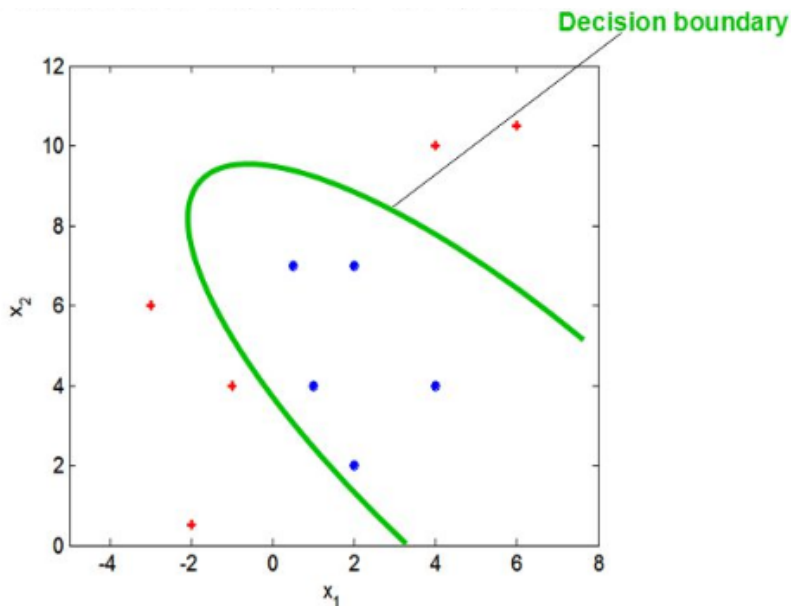
Reading Materials

- ▶ Max Kuhn. Chapter 14. Section 14.1

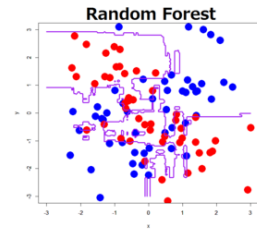
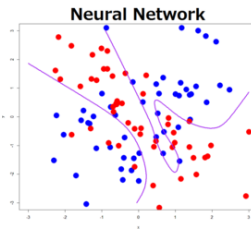
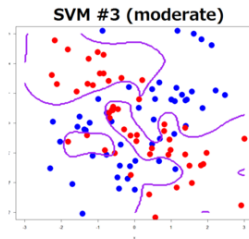
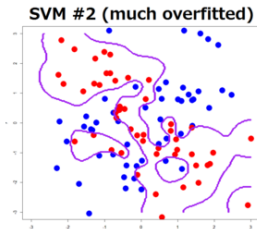
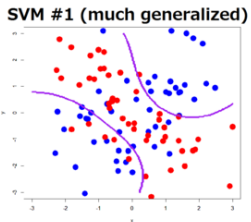
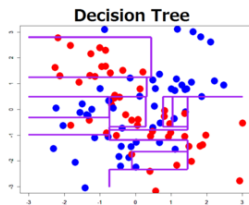
Decision Boundary in Classification

Classification is a process of finding the **decision boundary** that best separates two classes

Decision Boundary in Classification



Decision Boundary in Classification



► SVM = Support Vector Machine

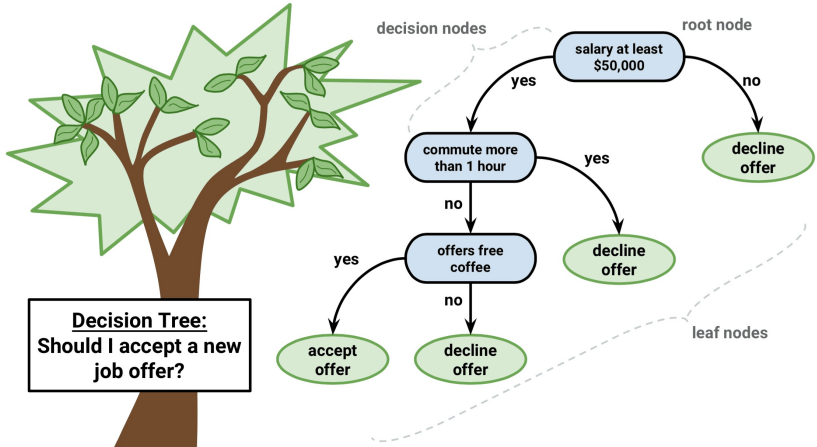
Decision Tree

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

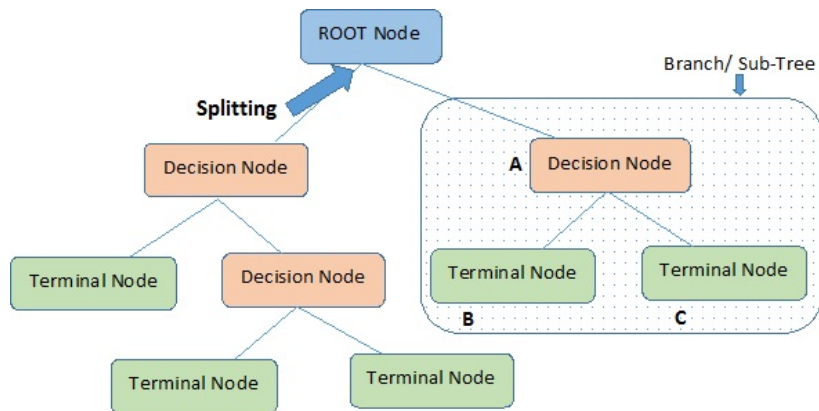
Example of Classification Tree

Link

Example of Classification Tree

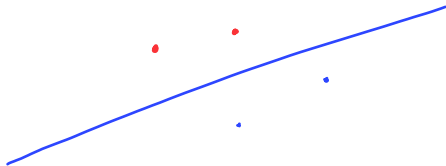


Example of Classification Tree

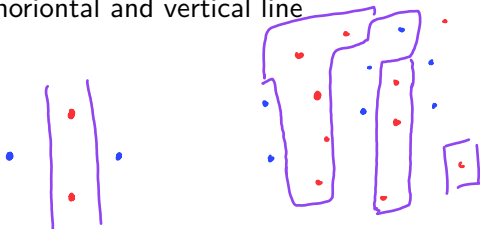


Note:- A is parent node of B and C.

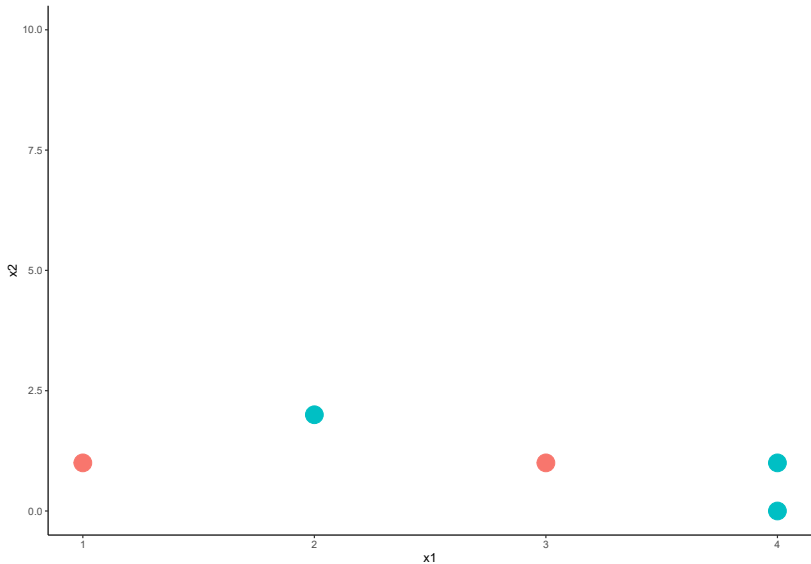
Classification Tree



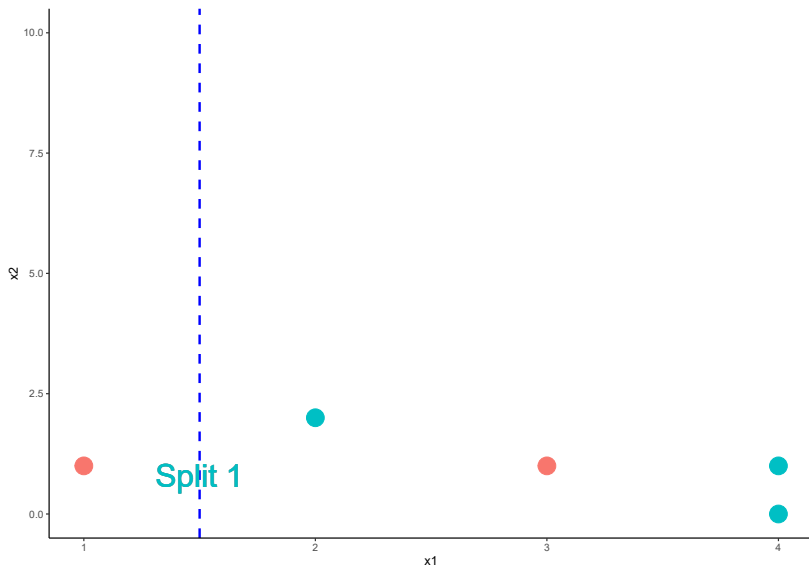
- In two dimension, classification Tree's decision boundary is a collection of horizontal and vertical line



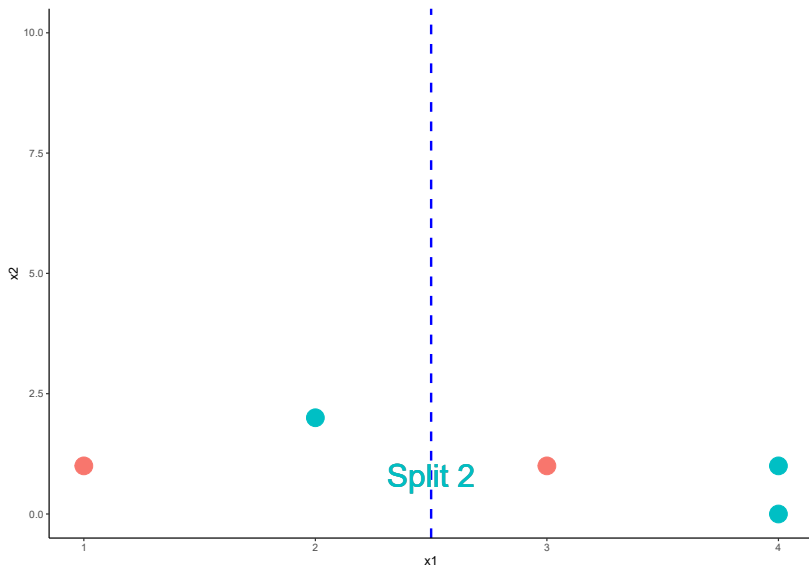
Find a vertical line that best separate **red** and **green**.



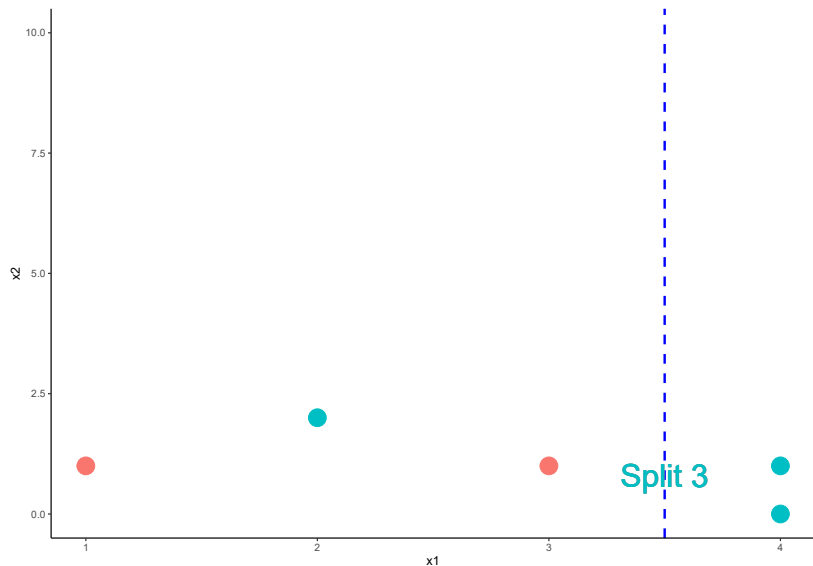
One way to separate the reds and greens



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

Partial Answer

- ▶ Which is the better split between Split 1 and Split 3?

Partial Answer

- ▶ We need to find a way to measure *how good a split is*

Impurity Measure

Green Red

3	2
---	---

 (high impurity)

- ▶ The impurity of a node (**a node = a subset of the data or the original data**) measure how uncertain the node is.

5	0
---	---

 (low impurity)

100	0
-----	---

60	59
----	----

 high impurity

90	49
----	----

Impurity Measure

- ▶ 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.

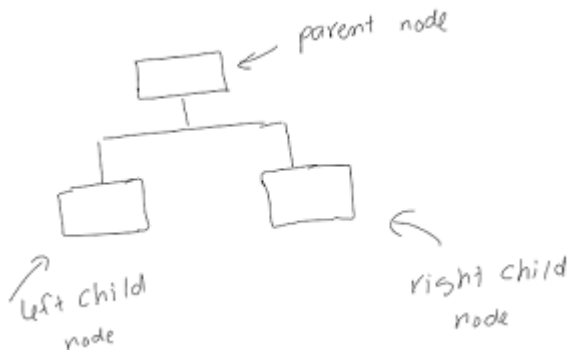
Impurity Measure

- ▶ More uncertain = Greater impurity

Children Impurity

- ▶ A split resulting smaller children impurity is a **better split**

Children Impurity ($I_{children}$)



$$I_{children} = \frac{N_{left}}{N} I_{left} + \frac{N_{right}}{N} I_{right}$$

- ▶ N_{left} and N_{right} are the number of points in the left child node and right child node, respectively.
- ▶ $N_{left} + N_{right} = N$

Spit 1 :

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.

① By Classification Error: $I = \min\{p_0, p_1\}$

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

③ By Entropy: $I = -p_0 \log_2(p_0) - p_1 \log_2(p_1)$

Example :

p_0	p_1
60	40

$$p_0 = \frac{60}{100} = .6, \quad p_1 = \frac{40}{100} = .4$$

① I by classification

$$\begin{aligned} I &= \min(p_0, p_1) \\ &= \min(.6, .4) = .4 \end{aligned}$$

② I (by Gini Index)

$$\begin{aligned} I &= 1 - p_0^2 - p_1^2 = 1 - .6^2 - .4^2 \\ &= .48 \end{aligned}$$

③ I (by Entropy)

$$= -p_0 \log_2(p_0) - p_1 \log_2(p_1) = .921$$

p_0	p_1
90	10

$$p_0 = \frac{90}{100} = .9, \quad p_1 = \frac{10}{100} = .1$$

①

$$I = \min(.9, .1) = .1$$

②

$$I = 1 - .9^2 - .1^2 = .18$$

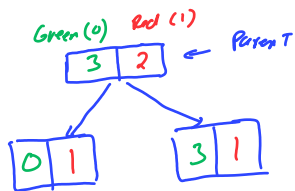
③

$$\begin{aligned} I &= -.9 \log_2(.9) - .1 \log_2(.1) \\ &= .469 \end{aligned}$$

Calculation

- ▶ Let's calculate the Children Impurity ($I_{children}$) of the three splits to decide which split is the best

Split 1

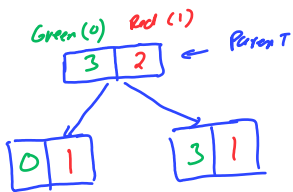


By Gini Index,

$$I_{\text{Parent}} = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = .48$$

$$I_{\text{children}} =$$

Split 1 :



$$I_{\text{children}} = \frac{N_{\text{left}}}{N} I_{\text{left}} + \frac{N_{\text{right}}}{N} I_{\text{right}}$$

$$N_{\text{left}} = 0 + 1 = 1$$

$$N_{\text{right}} = 3 + 1 = 4$$

$$N = 3 + 2 = 5$$

B-1 Gini Index,

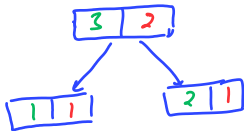
$$I_{\text{Parent}} = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \boxed{.48}$$

$$I_{\text{children}} = \frac{1}{5} \cdot I_{\text{left}} + \frac{4}{5} \cdot I_{\text{right}}$$

$$I_{\text{left}} = 1 - 1^2 - 0^2 = 0, \quad I_{\text{right}} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = .375$$

$$\Rightarrow I_{\text{children}} = \frac{1}{5} \cdot 0 + \frac{4}{5} \cdot .375 = \boxed{.3}$$

Split 2 :



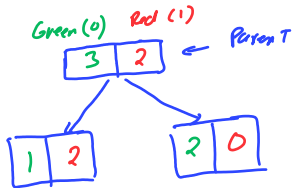
$$N_{\text{left}} = 2, \quad N_{\text{right}} = 3, \quad N = 5$$

$$I_{\text{children}} = \frac{2}{5} \cdot I_{\text{left}} + \frac{3}{5} \cdot I_{\text{right}}$$

$$= \frac{2}{5} \left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 \right] + \frac{3}{5} \left[1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \right]$$

$$= 0.467$$

Split 3 :



$$N_{\text{left}} = 3, \quad N_{\text{right}} = 2$$

$$\begin{aligned} I_{\text{children}} &= \frac{3}{5} \left[1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \right] + \frac{2}{5} \cdot [1 - 1] \\ &= .267 \end{aligned}$$

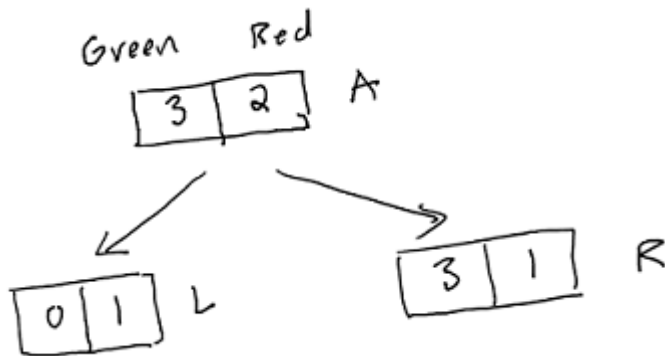
	I_{children}	I_{Gain}	$I_{\text{Gain}} = I_{\text{Parent}} - I_{\text{Children}}$
Split 1	.3	$.48 - .3 = .18$	
Split 2	.467	$.48 - .467 = .013$	
Split 3	.267	$.48 - .267 = .213$	

Split 3 has the max. I_{Gain} . Therefore Split 3 is the best split by Gini-Index

Split 1: Impurity by Classification Error

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

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



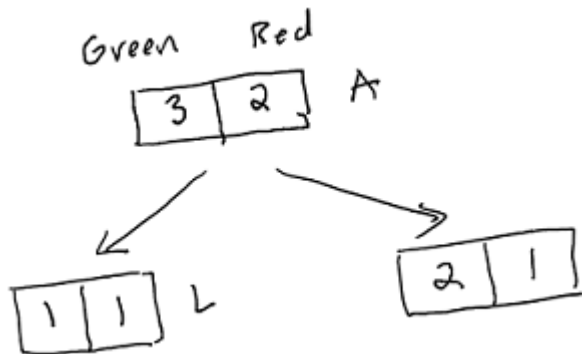
Split 1: Impurity by Classification Error

- ▶ 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, $I_R = \min(\frac{3}{4}, \frac{1}{4}) = \frac{1}{4}$
- ▶ Children Impurity of Split 1:

$$\begin{aligned}I_{children} &= \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R \\&= \frac{1}{5} \cdot 0 + \frac{4}{5} \cdot \frac{1}{4} = 0.2\end{aligned}$$

Split 2: Impurity by Classification Error

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



Split 2: Impurity by Classification Error

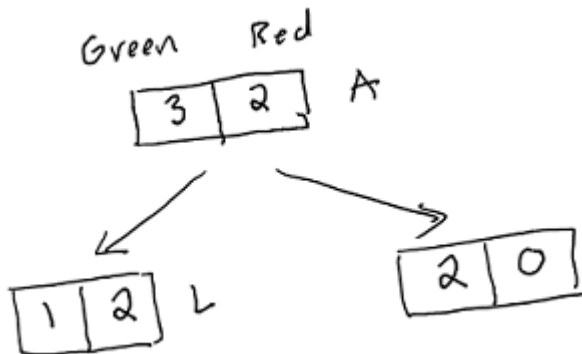
- ▶ Node *child left*, L: $p_0 = \frac{1}{2}, p_1 = \frac{1}{2}$. Thus, $I_L = \frac{1}{2}$
- ▶ Node *child right*, R: $p_0 = \frac{2}{3}, p_1 = \frac{1}{3}$. Thus, $I_R = \min(\frac{2}{3}, \frac{1}{3}) = \frac{1}{3}$
- ▶ Children Impurity of Split 2:

$$I_{children} = \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R$$

$$= \frac{2}{5} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{1}{3} = 0.4$$

Split 3: Impurity by Classification Error

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



Split 3: Impurity by Classification Error

- ▶ 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, $I_R = \min(1, 0) = 0$
- ▶ Children Impurity of Split 3:

$$I_{children} = \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R$$

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

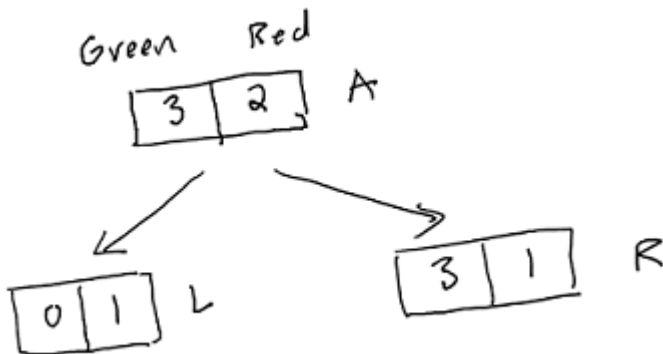
Comparing Impurity by Classification Error

	$I_{children}$
Split 1	0.2
Split 2	0.4
Split 3	0.2

- By classification error, Split 1 and Split 3 are tie as the best because they have the same Children Impurity ($I_{children}$).

Split 1: Impurity by Gini Index

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



Split 1: Impurity by Gini Index

- ▶ Node *child left*, L: $p_0 = \frac{0}{1} = 0, p_1 = \frac{1}{1} = 1$. Thus,

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

- ▶ Node *child right*, R: $p_0 = \frac{3}{4}, p_1 = \frac{1}{4}$. Thus,

$$I_R = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

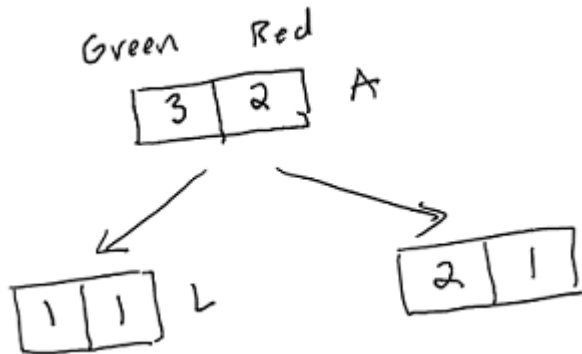
- ▶ Children Impurity of Split 1:

$$I_{children} = \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R$$

$$= \frac{1}{5} \cdot 0 + \frac{4}{5} \cdot 0.375 = 0.3$$

Split 2: Impurity by Gini Index

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



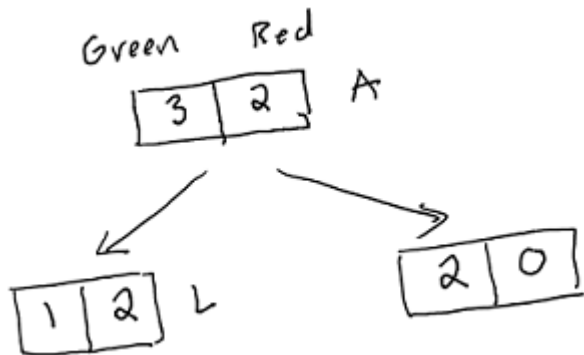
Split 2: Impurity by Gini Index

- ▶ 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,
 $I_R = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = 0.44$
- ▶ Children Impurity of Split 2:

$$\begin{aligned} I_{children} &= \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R \\ &= \frac{2}{5} \cdot \frac{1}{2} + \frac{3}{5} \cdot 0.44 = 0.464 \end{aligned}$$

Split 3: Impurity by Gini Index

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



Split 3: Impurity by Gini Index

- ▶ 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,
 $I_R = 1 - 0^2 - 1^2 = 0$
- ▶ Children Impurity of Split 3:

$$I_{children} = \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R$$

$$= \frac{3}{5} \cdot 0.44 + \frac{2}{5} \cdot 0 = \text{0.44} \cdot 264$$

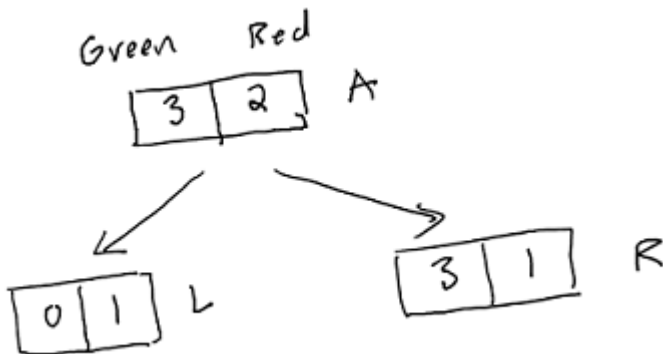
Comparing Impurity by Gini Index

	$I_{children}$
Split 1	0.3
Split 2	0.464
Split 3	0.464 .264

- By Gini Index, Split 3 is the best because it has the smallest Children Impurity ($I_{children}$).

Split 1: Impurity by Entropy

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



Split 1: Impurity by Entropy

- ▶ 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,

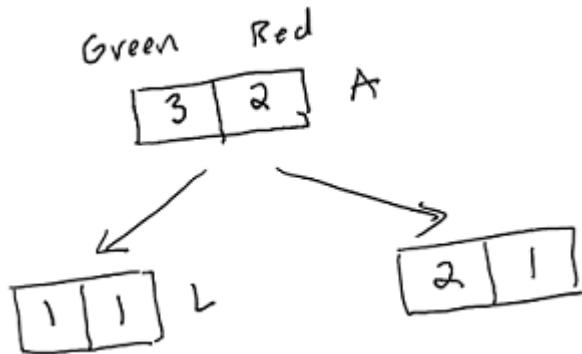
$$I_R = -\log_2\left(\frac{3}{4}\right) - \log_2\left(\frac{1}{4}\right) = 0.811$$

- ▶ Children Impurity of Split 1:

$$\begin{aligned} I_{\text{children}} &= \frac{N_{\text{left}}}{N} I_L + \frac{N_{\text{right}}}{N} I_R \\ &= \frac{1}{5} \cdot 0 + \frac{4}{5} \cdot 0.811 = 0.649 \end{aligned}$$

Split 2: Impurity by Entropy

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



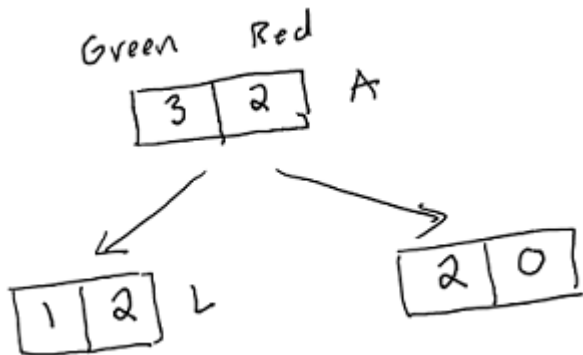
Split 2: Impurity by Entropy

- ▶ Node *child left*, L: $p_0 = \frac{1}{2}, p_1 = \frac{1}{2}$. Thus,
 $I_L = -\log_2(\frac{1}{2}) - \log_2(\frac{1}{2}) = 1$
- ▶ Node *child right*, R: $p_0 = \frac{2}{3}, p_1 = \frac{1}{3}$. Thus,
 $I_R = -\log_2(\frac{2}{3}) - \log_2(\frac{1}{3}) = 0.918$
- ▶ Children Impurity of Split 2:

$$\begin{aligned}I_{children} &= \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R \\&= \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot 0.918 = 0.951\end{aligned}$$

Split 3: Impurity by Entropy

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



Split 3: Impurity by Entropy

- ▶ 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, $I_R = 0$
- ▶ Children Impurity of Split 3:

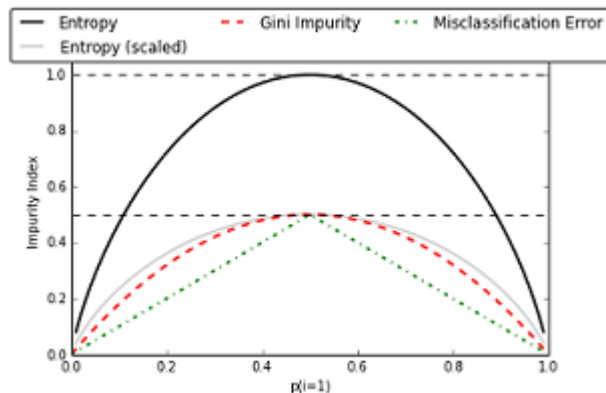
$$\begin{aligned}I_{children} &= \frac{N_{left}}{N} I_L + \frac{N_{right}}{N} I_R \\&= \frac{3}{5} \cdot 0.918 + \frac{2}{5} \cdot 0 = 0.551\end{aligned}$$

Comparing Impurity by Entropy

	$I_{children}$
Split 1	0.649
Split 2	0.951
Split 3	0.551

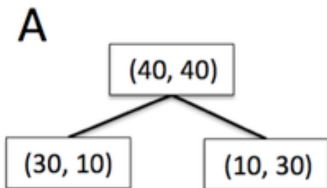
- By Gini Index, Split 3 is the best because it has the smallest Children Impurity ($I_{children}$).

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 (Extra credit)



- Which split is better?

Split 1

color

Green

Red

left

1

2

3

right

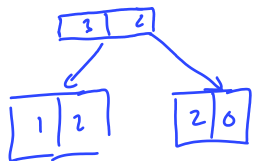
2

0

2

3

2



Split 2

color

Green

Red

left

1

1

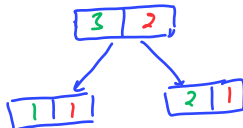
right

2

1

3

2



we want the split where the branch variable and the color variable are more dependent on each other (so that when you know the branch, you tend to know the color).

⇒ we can use the χ^2 test for independence.

Decide the best split using Chi-Square test of Independence

- ▶ Besides Children Impurity, 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_a : X and Y are dependent
- ▶ Test statistic:

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

small p-value support H_a

p-value ↗

⇒ we seek for the split with lowest p-value.

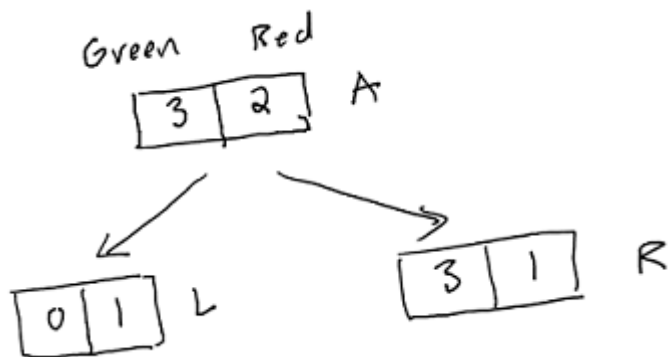
Review of Chi-Square test of Independence

- ▶ 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.**
or *Smallest p-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	Total
Left Branch	0	1	1
Right Branch	3	1	4
Total	3	2	

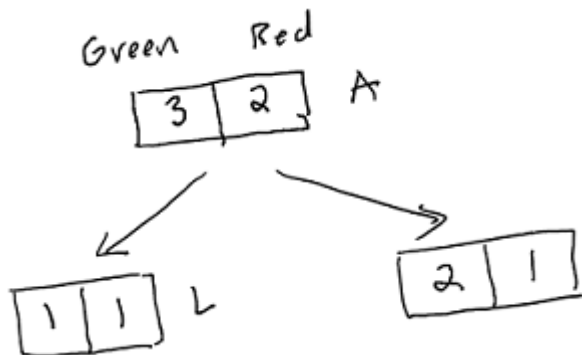
Split 1

$$\chi^2 = \frac{(e_1 - o_1)^2}{e_1} + \frac{(e_2 - o_2)^2}{e_2} + \frac{(e_3 - o_3)^2}{e_3} + \frac{(e_4 - o_4)^2}{e_4}$$

- ▶ $i = 1$ (Cell 1): $e_1 = \frac{1 \cdot 3}{5}$, $o_1 = 0$
- ▶ $i = 2$ (Cell 2): $e_2 = \frac{1 \cdot 2}{5}$, $o_2 = 1$
- ▶ $i = 3$ (Cell 3): $e_3 = \frac{3 \cdot 4}{5}$, $o_3 = 3$
- ▶ $i = 4$ (Cell 4): $e_4 = \frac{2 \cdot 4}{5}$, $o_4 = 1$
- ▶ Plug in, we have:

$$\chi^2 = 1.875$$

Split 2



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

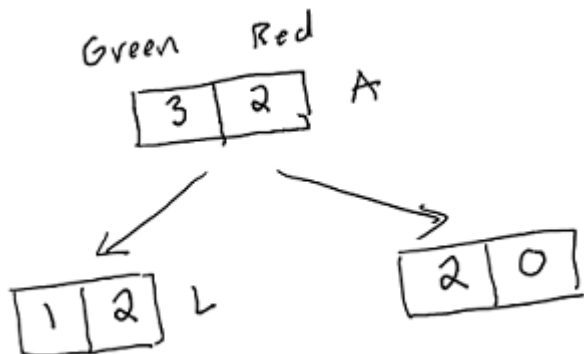
Split 2

$$\chi^2 = \frac{(e_1 - o_1)^2}{e_1} + \frac{(e_2 - o_2)^2}{e_2} + \frac{(e_3 - o_3)^2}{e_3} + \frac{(e_4 - o_4)^2}{e_4}$$

- ▶ $i = 1$ (Cell 1): $e_1 = \frac{2 \cdot 3}{5}$, $o_1 = 1$
- ▶ $i = 2$ (Cell 2): $e_2 = \frac{2 \cdot 2}{5}$, $o_2 = 1$
- ▶ $i = 3$ (Cell 3): $e_3 = \frac{3 \cdot 3}{5}$, $o_3 = 2$
- ▶ $i = 4$ (Cell 4): $e_4 = \frac{3 \cdot 2}{5}$, $o_4 = 1$
- ▶ Plug in, we have:

$$\chi^2 = 0.139$$

Split 3



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

Split 3


$$\chi^2 = \frac{(e_1 - o_1)^2}{e_1} + \frac{(e_2 - o_2)^2}{e_2} + \frac{(e_3 - o_3)^2}{e_3} + \frac{(e_4 - o_4)^2}{e_4}$$

- ▶ (Cell 1): $e_1 = \frac{2 \cdot 3}{5}$, $o_1 = 1$
- ▶ (Cell 2): $e_2 = \frac{2 \cdot 2}{5}$, $o_2 = 2$
- ▶ (Cell 3): $e_3 = \frac{3 \cdot 3}{5}$, $o_3 = 2$
- ▶ (Cell 4): $e_4 = \frac{3 \cdot 2}{5}$, $o_4 = 0$
- ▶ Plug in, we have:

$$\chi^2 = 2.222$$

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 !
(smallest p-value)

Logworth

- ▶ The quality of the split can be measured by **Logworth**
- ▶ Formula:

$$\text{logworth} = -\log(p_{\text{value}})$$

- ▶ The greater the logworth, the better the split

Logworth

	χ^2	\leftrightarrow p-value	\leftrightarrow logworth
Split 1	1.875	0.114	0.943
Split 2	0.139	0.998	0.0008
Split 3	2.222	0.088	1.055

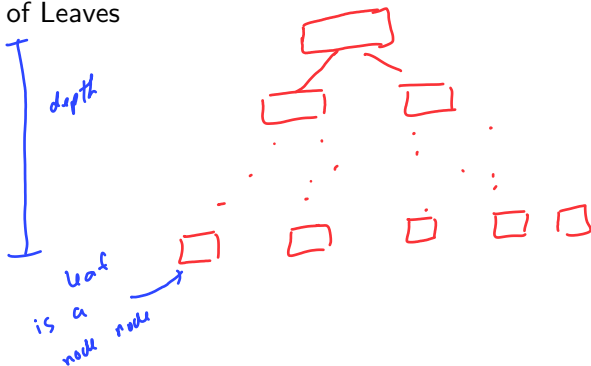
- ▶ Greatest χ^2 = Lowest *p-value* = Greatest logworth = Best Split
- ▶ Split 3 is the best split!

What happens after the first split?

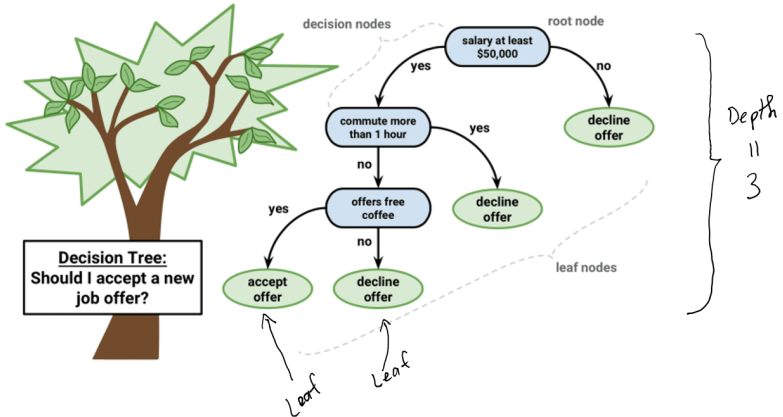
- ▶ After the first split, the data are divided into two subsets.
- ▶ The splitting process is repeated for each subset.
- ▶ The process ends when a stopping criteria is satisfied

Stopping Criteria

- ▶ Minimum Leaf Size: The minimum of observations in the leaves
- ▶ Maximum Number of Leaves
- ▶ Maximum Depth
- ▶ Others



Stopping Criteria

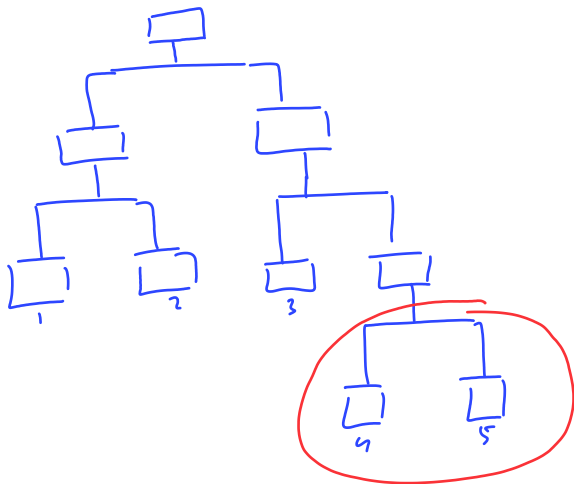


Decision Tree Algorithm - How to grow a tree

- ▶ Step 1: Calculate the Children Impurity or p – *value* of all possible splits at all variables
- ▶ Step 2: Select the split that give the minimum Children Impurity or lowest p – *value* to split the data into two subdata D_1 and D_2
- ▶ Repeat *Step 1* and *Step 2* to both D_1 and D_2 .
- ▶ Until a stopping criteria is satisfied

Complexity of Decision Tree

- ▶ A complexity of a tree can be measured by the number of leaves the tree has
- ▶ The more leaves a tree has, the more complex the tree is.
- ▶ A complex tree may be **overfitted**, i.e. having low training error but high testing error.



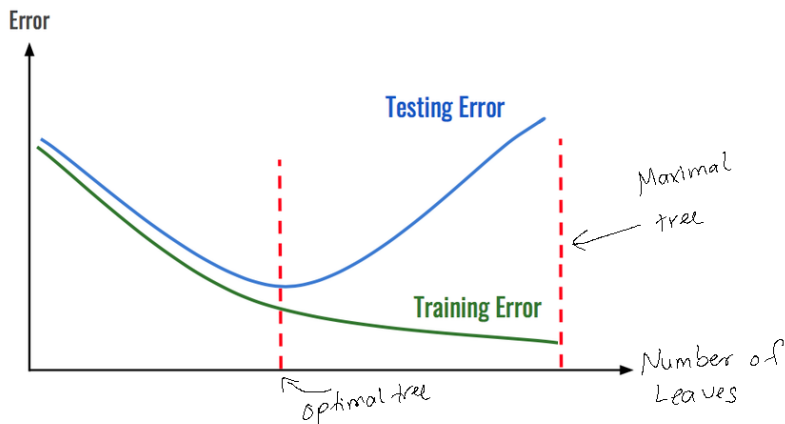
Pruning a tree

- ▶ For any given data, one can construct a tree that achieves 0 misclassification on training data
- ▶ After growing the tree one needs to prune it to avoid overfitted

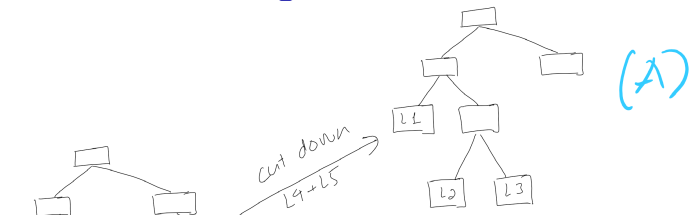
Pruning a tree

- ▶ The tree with maximum number of leaves is called the **maximal tree** (still satisfied the stopping rule)
- ▶ From the **maximal tree**, leaves are cut down, one by one, to obtain all possible subtrees
- ▶ The subtree with lowest error on validation data, is the **optimal tree**

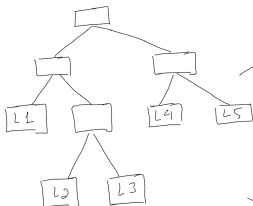
Maximal vs Optimal Tree



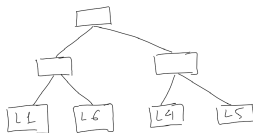
Example of Tree Pruning



(A)



Cut down
 $L2 + L3$



(B)

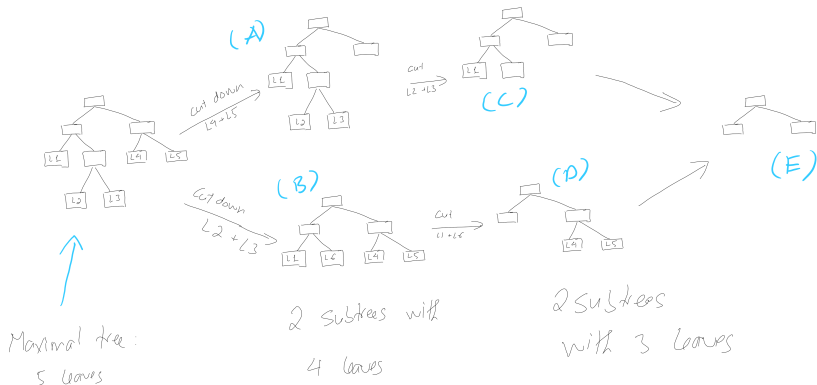
Maximal tree:

5 leaves

2 subtrees with

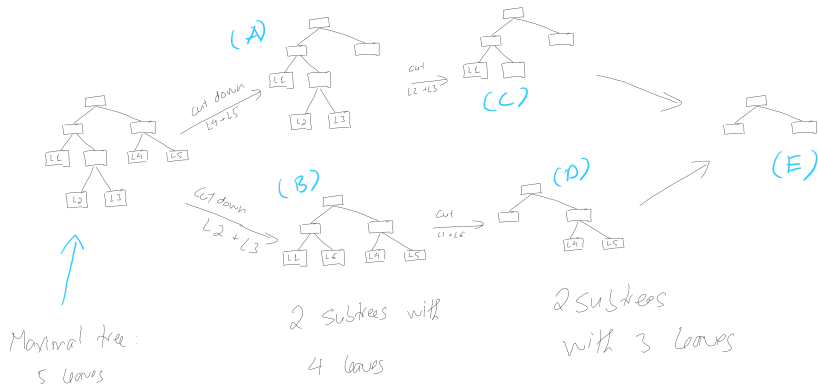
4 leaves

Example of Tree Pruning



- ▶ All the subtrees A, B, C, D, and E will be validated with the validation data to find the **optimal tree**
- ▶ The **optimal tree** could be the **maximal tree**!

Question



- What if both B and C give the lowest error on the validation data? Which tree should be selected as the final model?

