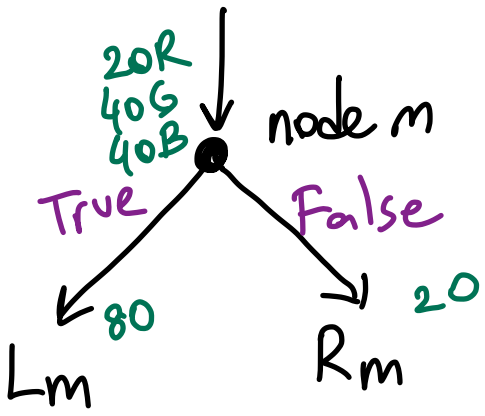
Univariate Decision Trees



$$L_m = \{ x \mid x_j > w_{m0} \}$$

threshold

↳ feature index

$$R_m = \{ x \mid x_j \leq w_{m0} \}$$

N_m = # of data points that reach node m

$N_m = 100$

$N_{m,c}$ = # of data points that reach node m from class #c

$K = 3$ $N_{m,1} = 20$ $N_{m,2} = 40$ $N_{m,3} = 40$

$N_{m,s}$ = # of data points that reach node m and takes split #s

$S = 2$ $N_{m,1} = 80$ $N_{m,2} = 20$

$$P_{mc} = \hat{P}(y=c \mid \underline{\underline{x_m}}) = \frac{N_{m,c}}{N_m}$$

$P_{m1} = \frac{20}{100}$

$P_{m2} = \frac{40}{100}$

$P_{m3} = \frac{40}{100}$

impurity of a node

$$I_m = -\sum_{c=1}^{N} P_{mc} \cdot \log_2 (P_{mc})$$

impurity of a split

$$I'_m = \sum_{s=1}^{S} \left[\frac{N_{m,s}}{N_m} \left[-\sum_{c=1}^{K} P_{msc} \log_2 (P_{msc}) \right] \right]$$

K → # of classes

weight of child node

impurity of child node

$$\frac{80}{100} I_m(L_m) + \frac{20}{100} I_m(R_m)$$

Entropy:

$$-p\log_2(p) - (1-p)\log_2(1-p)$$

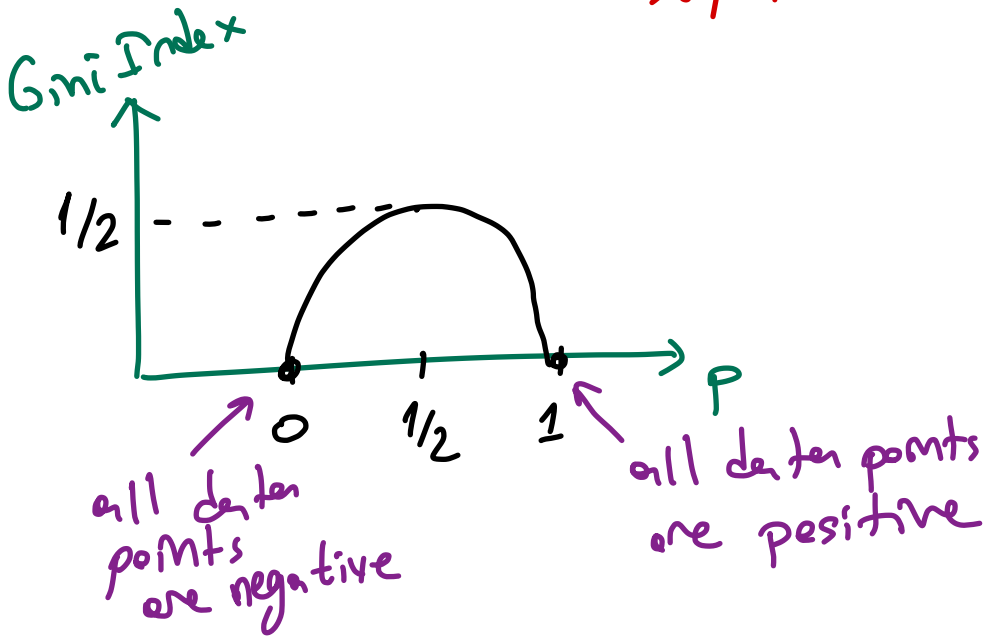
p = ratio of positive data points $\rightarrow \frac{N_+}{N_+ + N_-}$

$1-p$ = ratio of negative data points $\rightarrow \frac{N_-}{N_+ + N_-}$

$0\log_2(0) \equiv 0$

Gini Index:

$$2 \cdot p \cdot (1-p)$$



Misclassification Error:

60% 40%

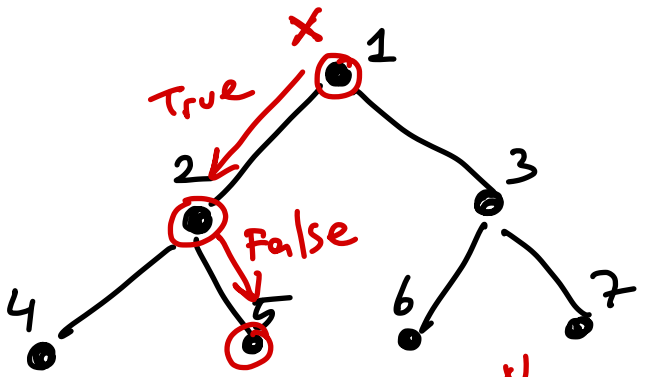
$$1 - \max(p, 1-p) \quad \text{OR} \quad \min(p, 1-p)$$

multiclass classification $\Rightarrow 1 - \max(p_1, p_2, \ldots, p_K)$

Classification accuracy when majority label is used

Regression Trees:

$$b_m(\underline{x}) = \begin{cases} 1 & \text{if } x \in \chi_m \text{ (x reaches node m)} \\ 0 & \text{otherwise} \end{cases}$$



$b_1(x) = 1$ $b_2(x) = 1$ $b_3(x) = 0$

$b_4(x) = 0$ $b_5(x) = 1$ $b_6(x) = 0$

$b_7(x) = 0$

$$N_m = \sum_{i=1}^{N} b_m(x_i)$$

error of a node.

$$E_m = \frac{1}{N_m} \sum_{i=1}^{N} \left[(y_i - g_m)^2 b_m(x_i) \right]$$

y_i: true value; g_m: predicted value at node m.

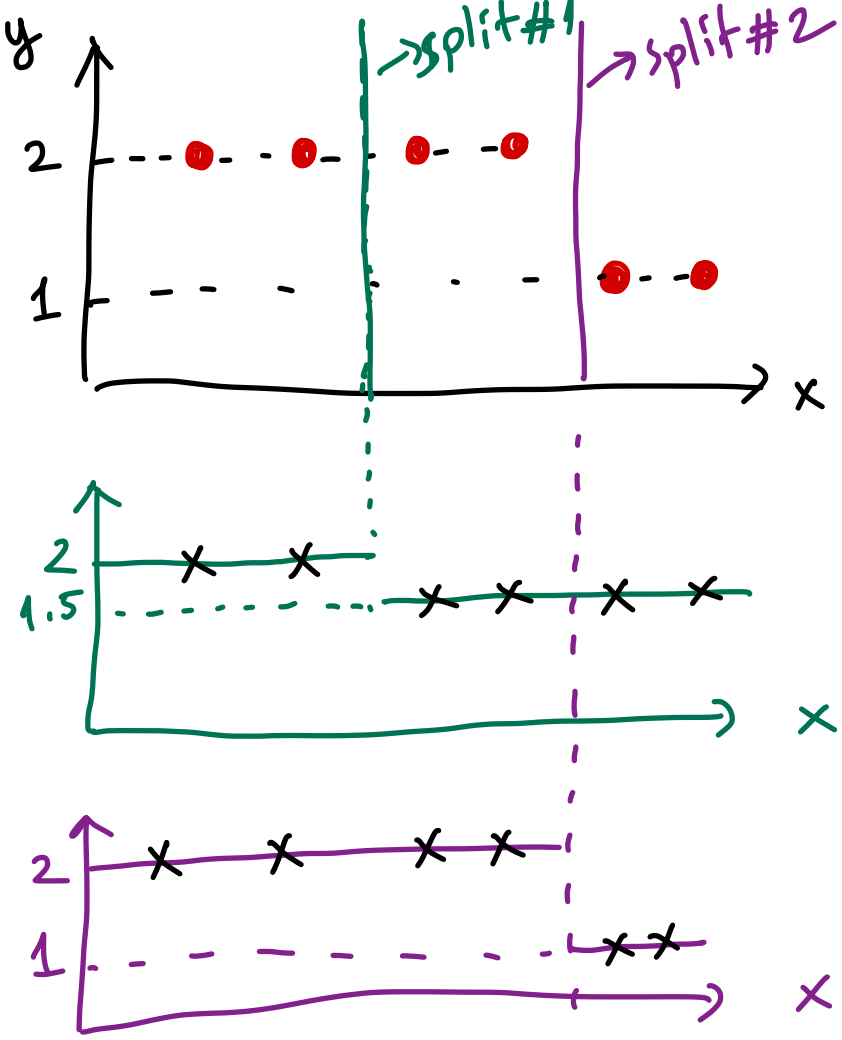
N_m: # of data points that reach to node m.

$$g_m = \frac{\sum_{i=1}^{N} [y_i b_m(x_i)]}{\sum_{i=1}^{N} b_m(x_i)}$$

average response (sample mean)

error of a split

$$E_m' = \frac{1}{N_m} \sum_{s=1}^{S} \sum_{i=1}^{N} \left[(y_i - g_{ms})^2 b_{ms}(x_i) \right]$$



$$E(S_1) = \frac{1}{6}\left[(2-2)^2+(2-2)^2+(2-1.5)^2+(2-1.5)^2+(1-1.5)^2+(1-1.5)^2\right] = 1/6$$

$$E(S_2) = \frac{1}{6}\left[(2-2)^2+(2-2)^2+(2-2)^2+(2-2)^2+(1-1)^2+(1-1)^2\right] = 0$$

$w = [w_1, w_2, \dots, w_D]$

$$[0\ 0 \dots 1 \dots 0\ 0] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_{D-1} \\ x_D \end{bmatrix} - w_{m0} > 0$$

Multivariate Decision Trees

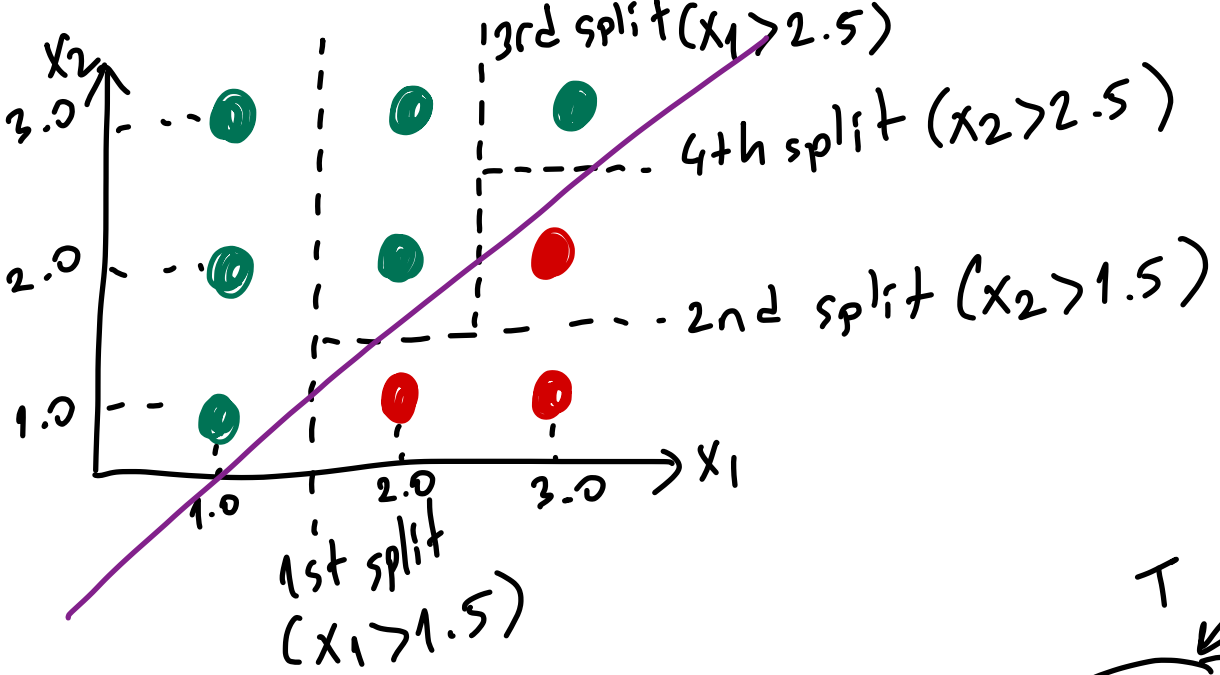
$f_m(x)$: $x_j > w_{m0}$

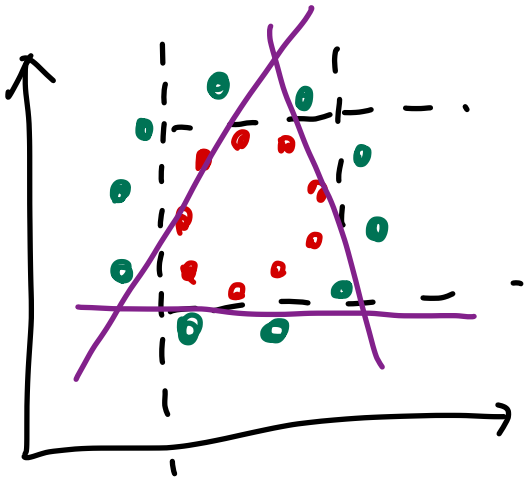
$\Rightarrow x_j - w_{m0} > 0$ ⇐ univariate split

$f_m(x)$: $w_m^T . x + w_{m0} > 0$ (threshold)

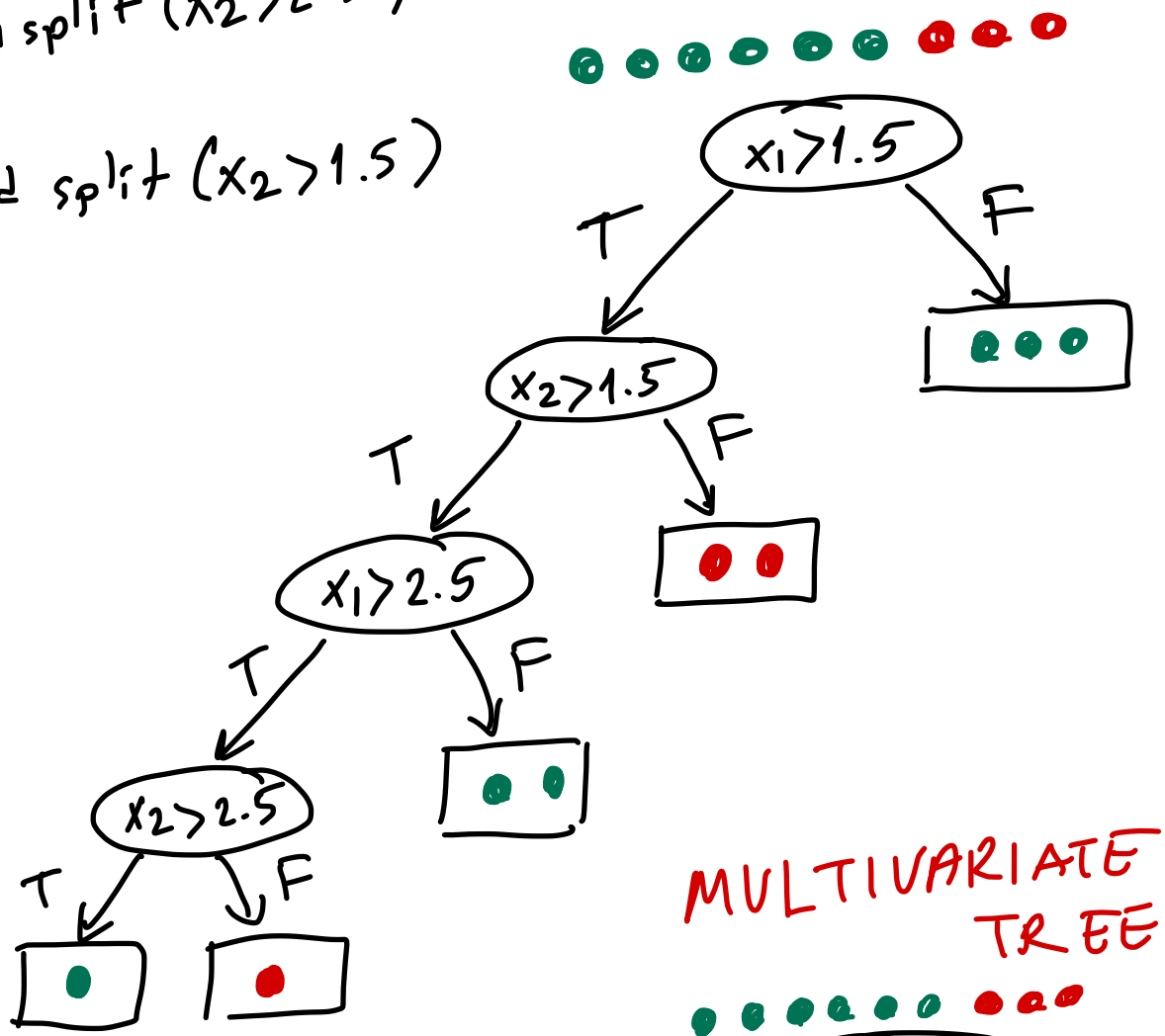
⇐ multivariate split.

w_m: arbitrary vector in $\mathbb{R}^D$

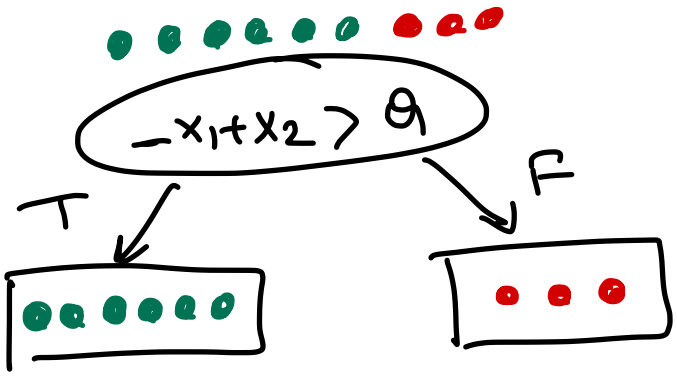




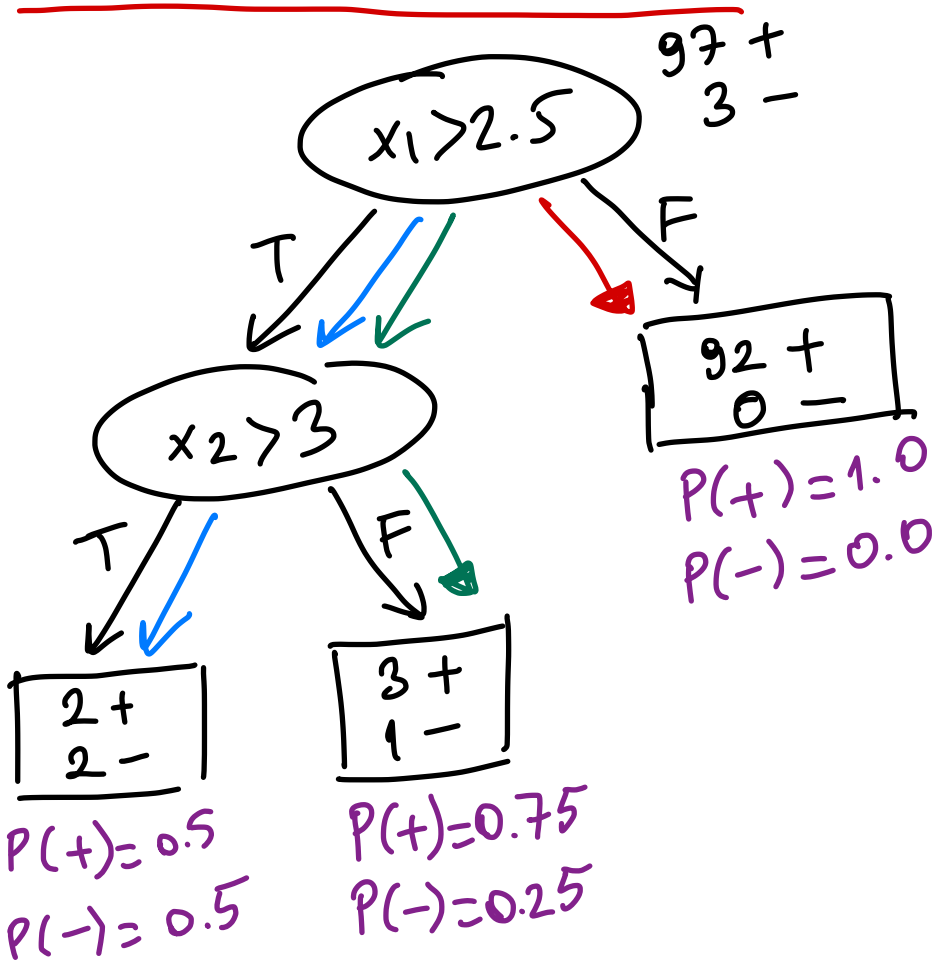
UNIVARIATE TREE



MULTIVARIATE TREE



RULE EXTRACTION

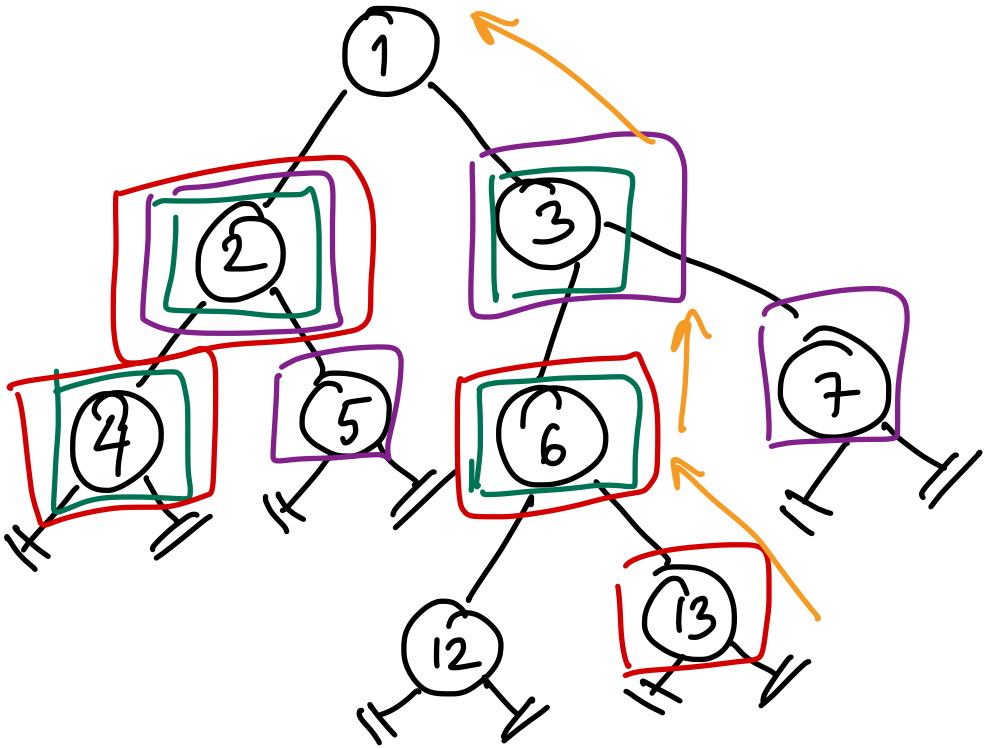


- extract one rule set for each terminal node.

Path 1: $x_1 \leq 2.5$

Path 2: $x_1 > 2.5$ AND $x_2 > 3$

Path 3: $x_1 > 2.5$ AND $x_2 \leq 3$



left child = 2 * parent

6 = 2 * 3

4 = 2 * 2

right child = 2 * parent + 1

7 = 2 * 3 + 1

5 = 2 * 2 + 1

$\lfloor 2.5 \rfloor = 2$

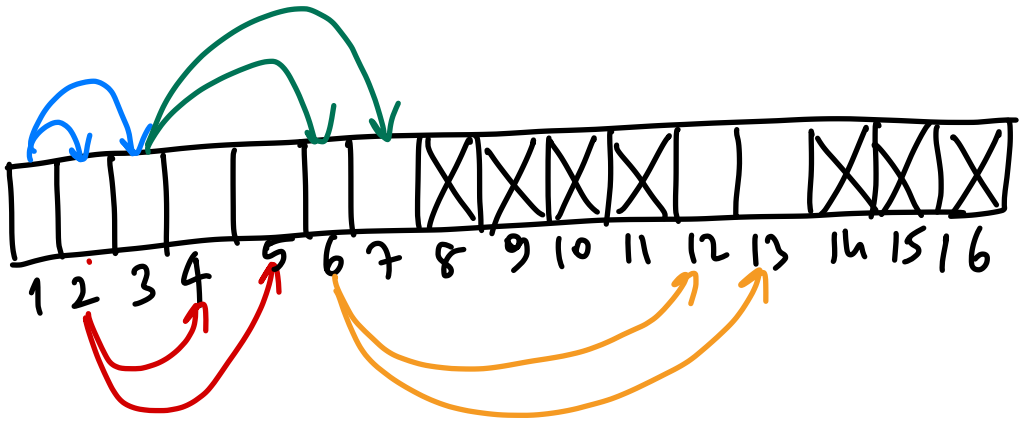
$\lfloor 3 \rfloor = 3$

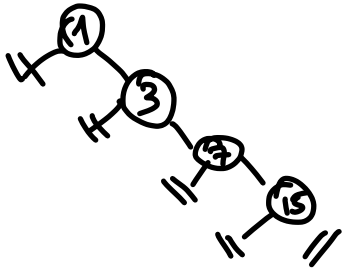
parent = $\lfloor$ child node/2 $\rfloor$

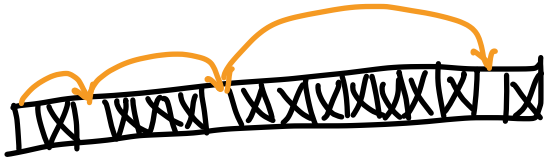
↑ floor function

6 = $\lfloor 13/2 \rfloor = \lfloor 6.5 \rfloor$

2 = $\lfloor 4/2 \rfloor = \lfloor 2 \rfloor$







Terminal nodes ⇒ 4, 5, 7, 12, 13

Path 1 4 → 2 → 1

Path 2 5 → 2 → 1

Path 3 7 → 3 → 1

Path 4 12 → 6 → 3 → 1

Path 5 13 → 6 → 3 → 1 ✓