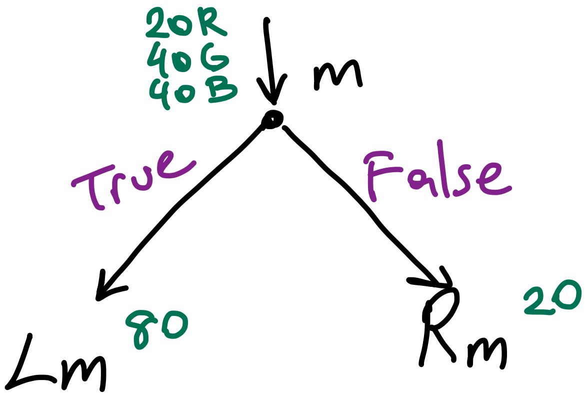
Univariate Decision Trees



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

threshold

feature index

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

$N_m = 100 \Leftarrow N_m$ = # of data points that reach node m

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

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

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

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

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

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

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

$s = 2$

$N_{m,1} = 80$ $\quad N_{m,2} = 20$

impurity of a node

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

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

impurity of a split

$$I_m' = \sum_{s=1}^{S} \left[\underbrace{\frac{N_{m,s}}{N_m}}_{\text{weight of a child node}} \cdot \underbrace{\left[-\sum_{c=1}^{K} P_{msc} \cdot \log_2(P_{msc}) \right]}_{\text{impurity of that child node.}} \right]$$

Entropy: $-p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p)$

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

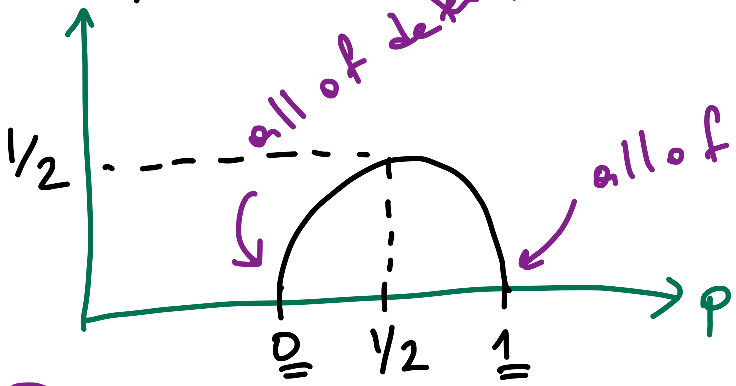
$1-p$ = ratio of negative data points $\rightarrow \dfrac{N_-}{N_+ + N_-}$

$0 \log_2 0 \equiv 0$

Gini Index: $2 \cdot p \cdot (1-p)$

all of data points are negative.

all of data points are +



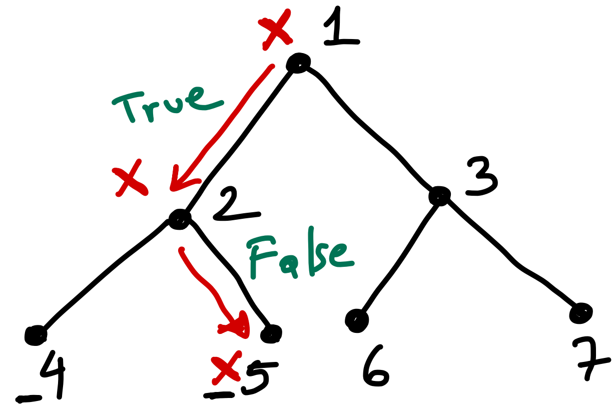
Misclassification Error:

$1 - \max(p, 1-p)$ OR $\min(p, 1-p)$

multiclass classification $\Rightarrow$ $1 - \underbrace{\max(p_1, p_2, \ldots, p_k)}_{\text{classification accuracy when majority label is used.}}$

Regression Trees:

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



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

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

$b_7(x) = 0$

error for nodes

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

$y_i \to$ true value

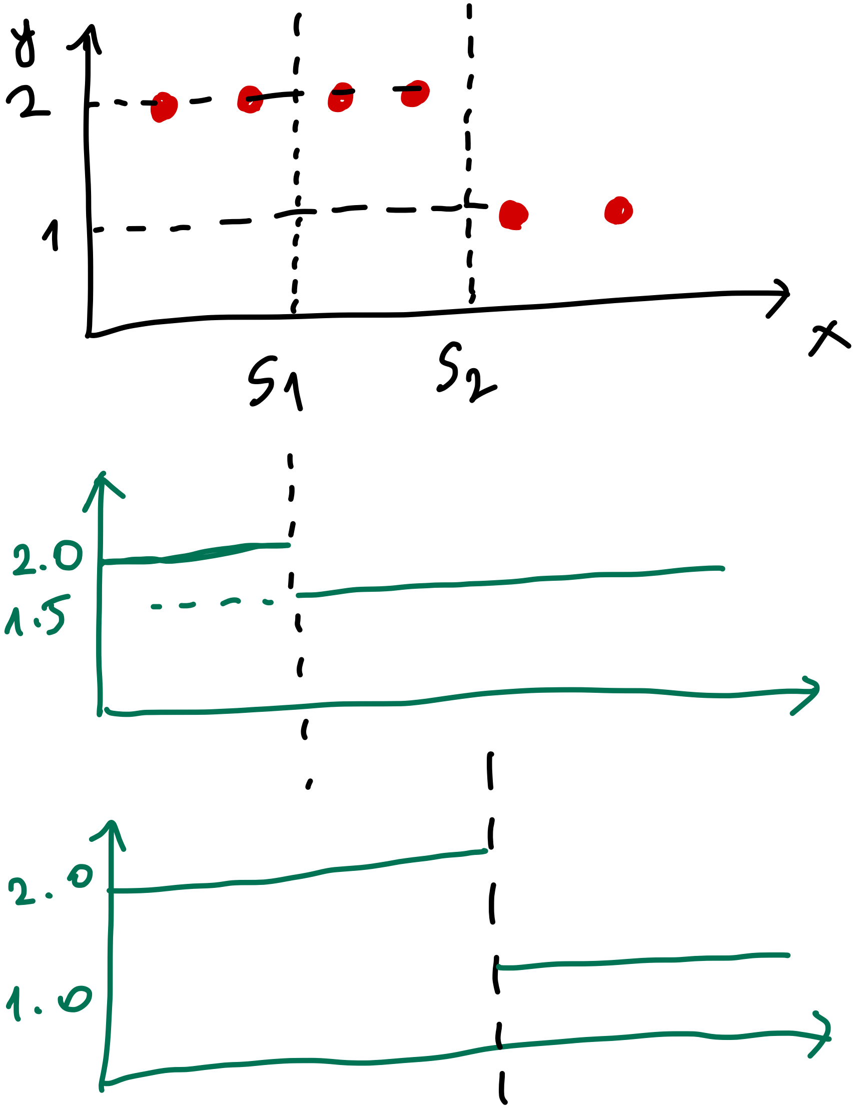
$g_m \to$ predicted value at node m

$N_m \to$ # of data points that reach to node m

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

error for splits

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



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

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

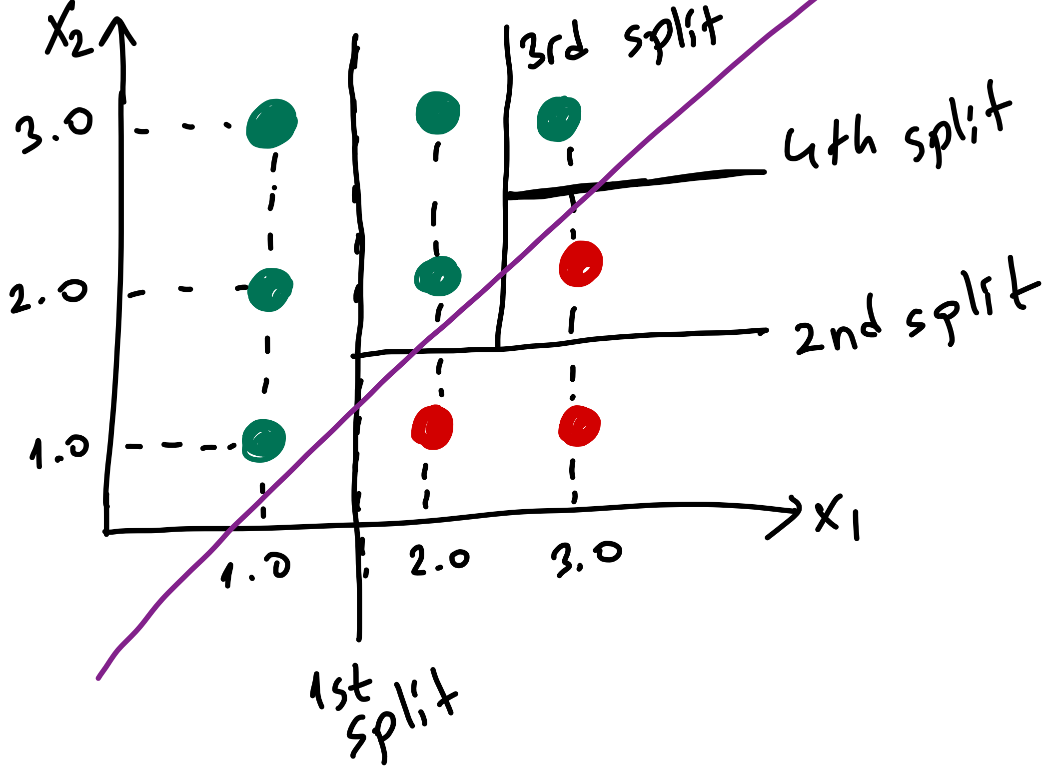
$$\begin{bmatrix} w_1 & w_2 & w_3 & \cdots & w_j & \cdots & w_D \end{bmatrix}$$

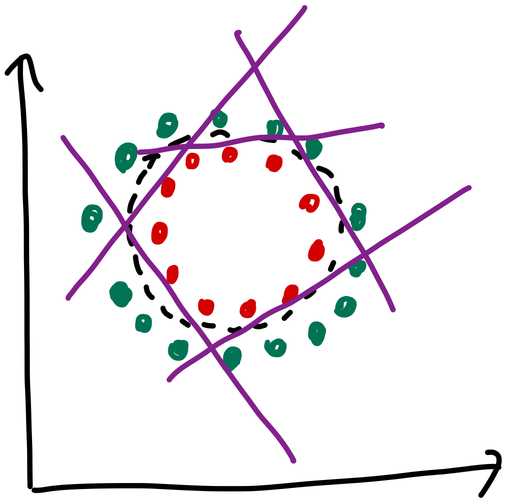
$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_D \end{bmatrix} - w_{m0} > 0$$

Multivariate Decision Trees:

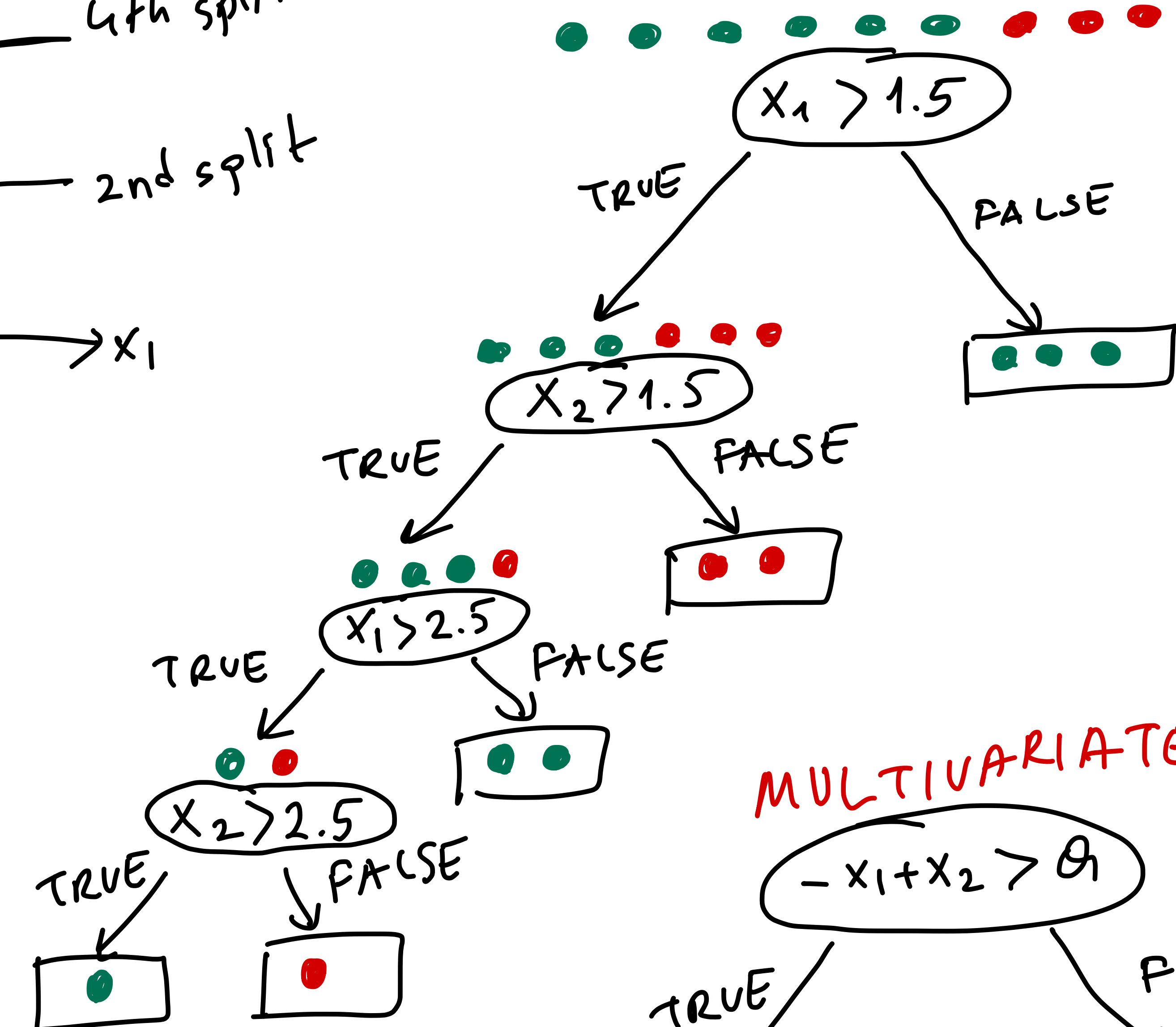
$f_m(x): \quad x_j > w_{m0} \Rightarrow \underbrace{x_j - w_{m0}} > 0$ ⇐ univariate split

$f_m(x): \quad w_m^T \cdot x + w_{m0} > 0$ ⇐ multivariate split

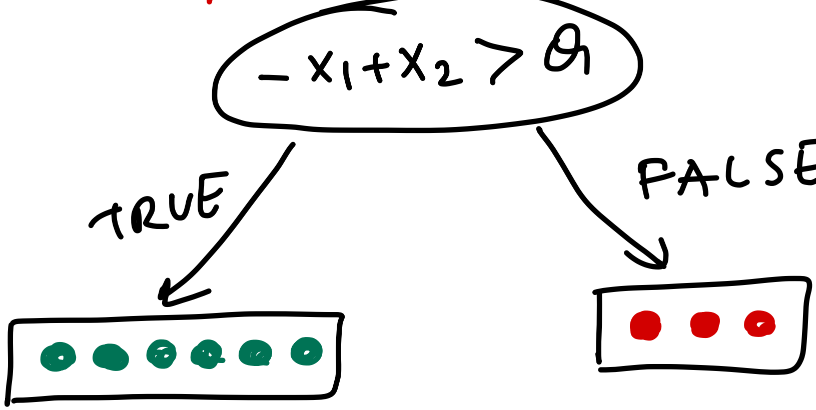




UNIVARIATE TREE

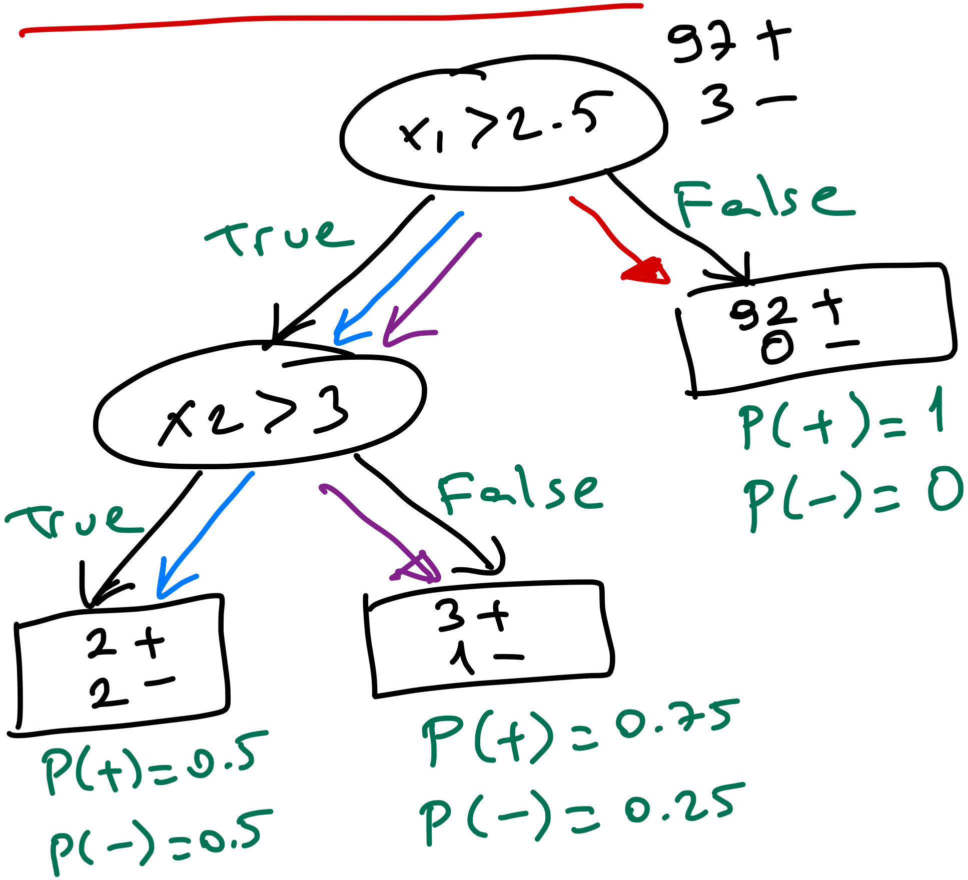


MULTIVARIATE TREE



$$\hat{w}_1 x_1 + \hat{w}_2 x_2 + \hat{w}_{m0} > 0$$

Rule Extraction!



- extract one rule for each terminal node

Path 1: $x_1 \leq 2.5$

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

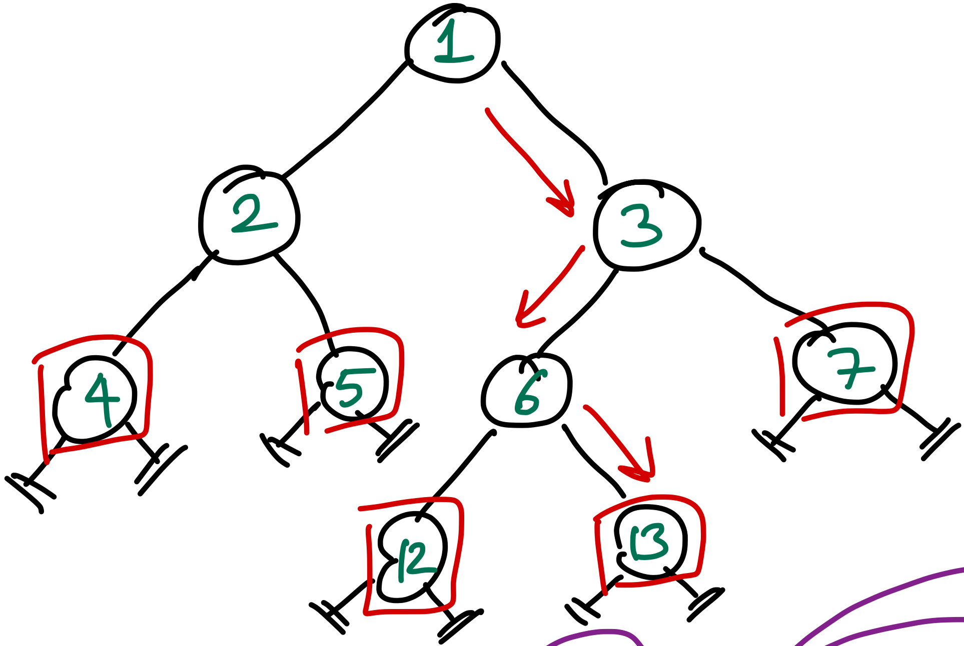
Path 3: $x_1 > 2.5$ & $x_2 \leq 3$

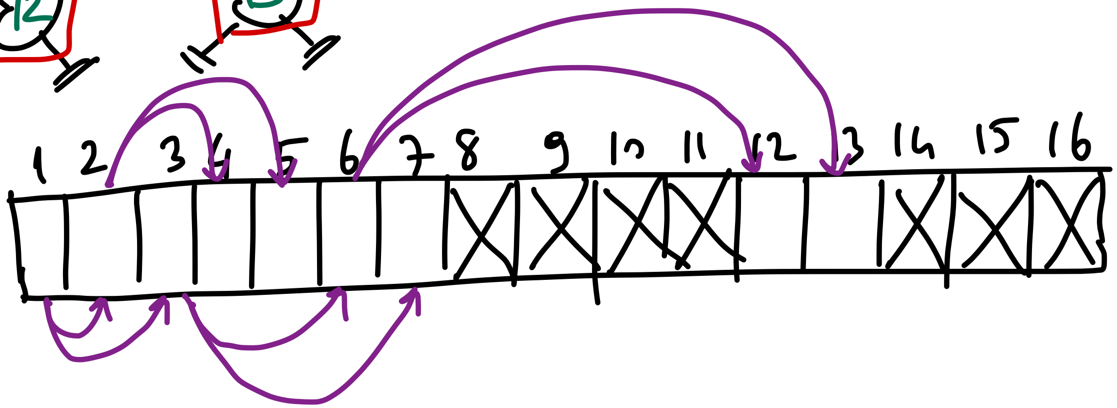
left child = 2 * parent

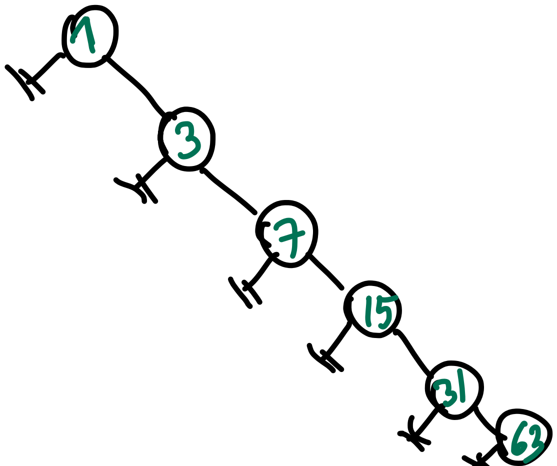
right child = 2 * parent + 1

parent = $\lfloor child/2 \rfloor$

floor function







13 ← 6 ← 3 ← 1

12 ← 6 ← 3 ← 1

7 ← 3 ← 1

5 ← 2 ← 1

4 ← 2 ← 1