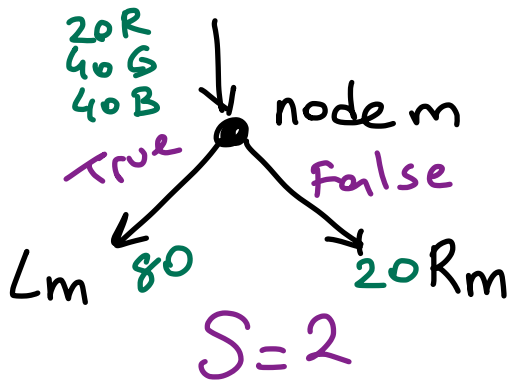
Univariate Decision Trees



$S = 2$

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

w_{m0} → threshold

x_j → feature index

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

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

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

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

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

$(20,40,40)\ N_{mc}$ = # of data points that reach node m from class c

$(80,20)\ N_{m,s}$ = # of data points that reach node m and take split s

$3 \Leftarrow K$ = # of classes

$$p_{mc} = \hat{P}(y = c \mid x_m) = \frac{N_{mc}}{N_m}$$

impurity of a node $\Leftarrow$

$$I_m = -\sum_{c=1}^{K} p_{mc} \log_2 (p_{mc}) \Leftarrow \text{Entropy}$$

impurity of a split $\Leftarrow$

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

weight of a child node × impurity of a node

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

Entropy:

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

$\frac{N+}{N+ + N-}$ ← p = ratio of positively labeled data points

$\frac{N-}{N+ + N-}$ ← $1-p$ = ratio of negatively labeled data points

$0 \log_2(0) \equiv 0$

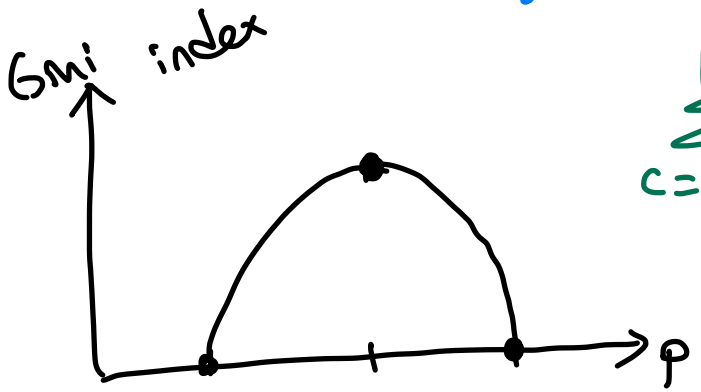
Gini Index:

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

$p=0 \Rightarrow 2 \cdot 0 \cdot (1-0)$

$p=0.5 \Rightarrow 1 \cdot \frac{1}{2} \cdot \frac{1}{2}$

$p=1.0 \Rightarrow 2 \cdot 1 \cdot (1-1)$



all data points are from (−) class

all data points are from (+) class

$$\sum_{c=1}^{K} p_c(1-p_c)$$

0% → 0 100% → 0 0% → 0

Misclassification Error:

60% 40%

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

classification accuracy

OR

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

misclassification rate

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

classification accuracy when majority label is used as prediction

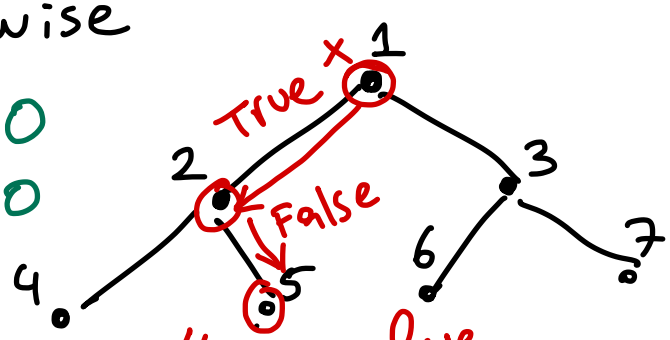
Regression Trees

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

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

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

$b_7(x) = 0$



path $\Rightarrow 1 - 2 - 5$

error of a node

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

y_i: observed / true value

g_m: predicted value at node m

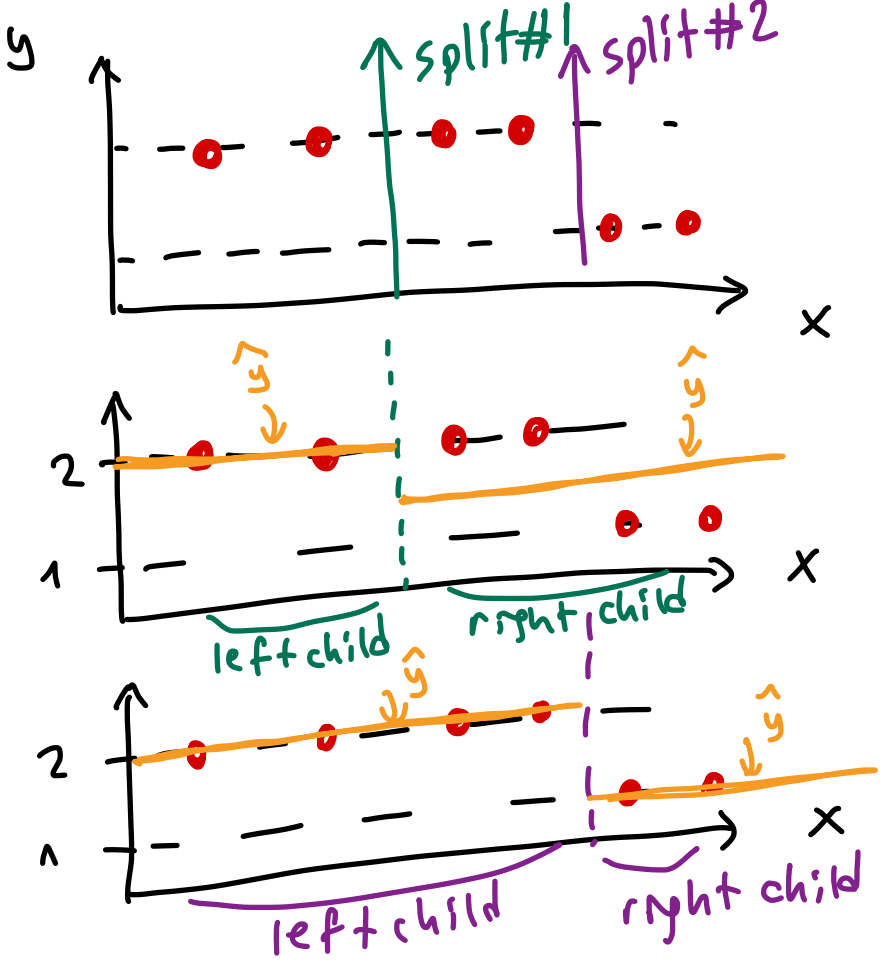
N_m: # of data points that reach node m

$$g_m = \frac{\sum_{i=1}^{N} \left[y_i \, b_m(x_i) \right]}{\sum_{i=1}^{N} b_m(x_i)}$$ } average response (sample mean)

error of a split

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

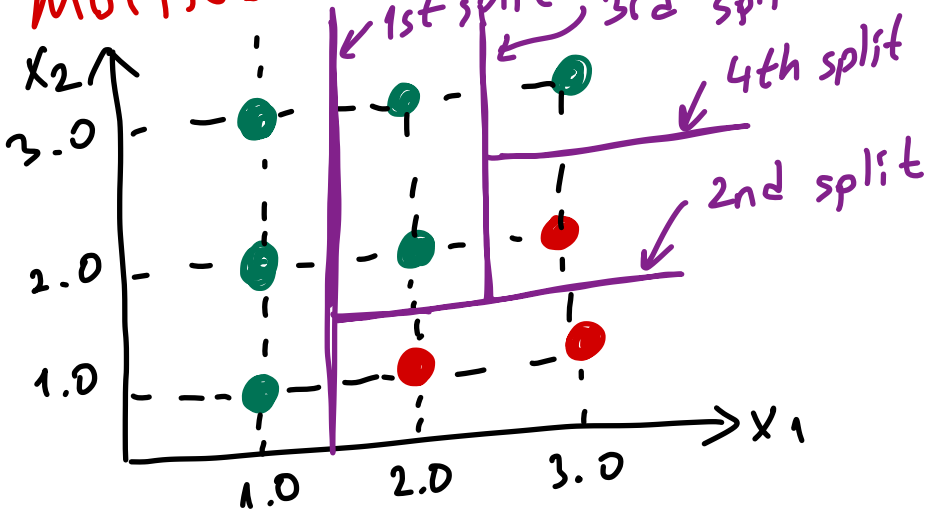
$$\sum_{s=1}^{S} \frac{N_{ms}}{N_m} \sum_{i=1}^{N} \frac{1}{N_{ms}} \left[\left(y_i - g_{ms} \right)^2 b_{ms}(x_i) \right]$$



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

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

Multivariate Decision Trees

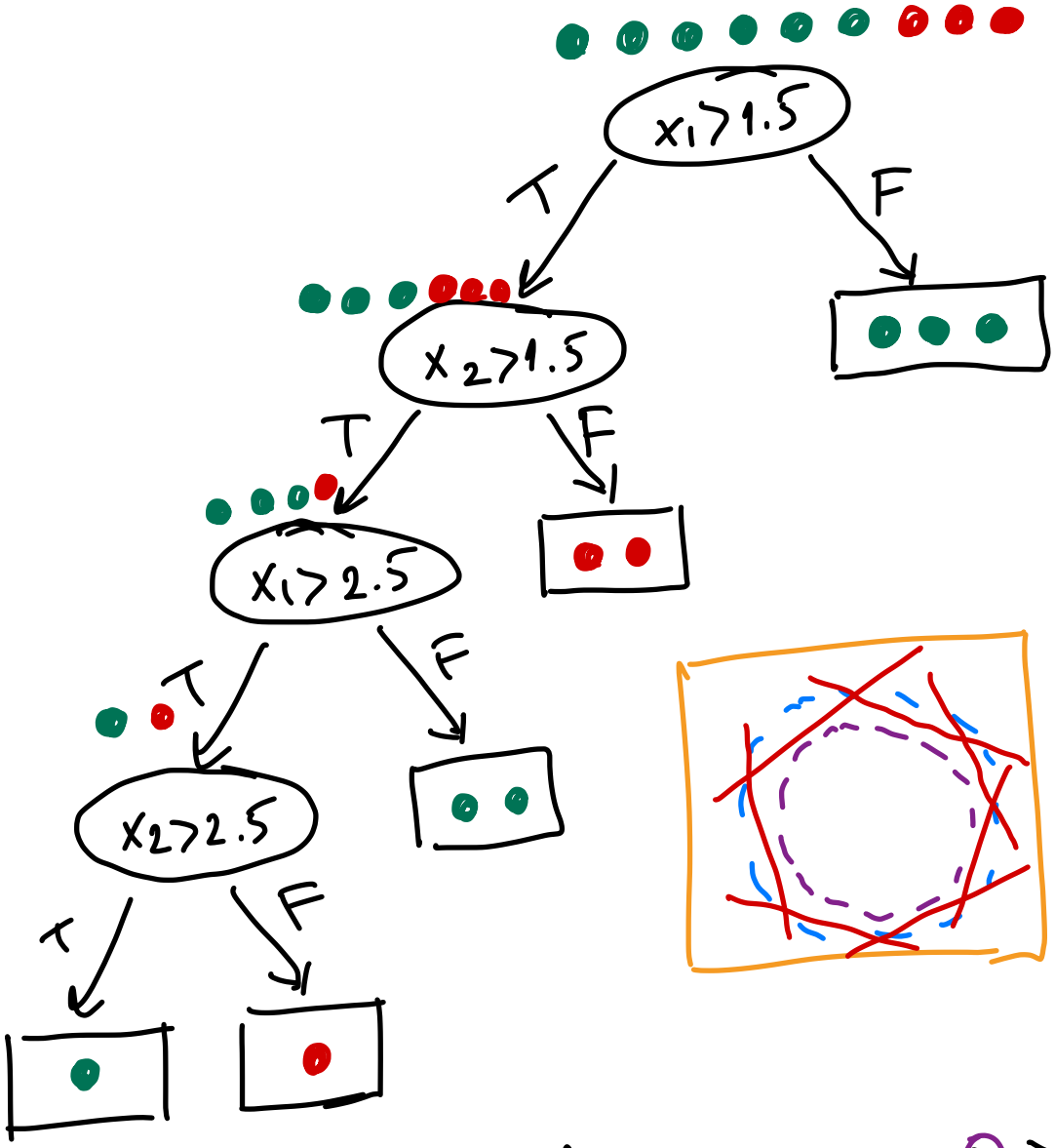


1st split $x_1 > 1.5$

2nd split $x_2 > 1.5$

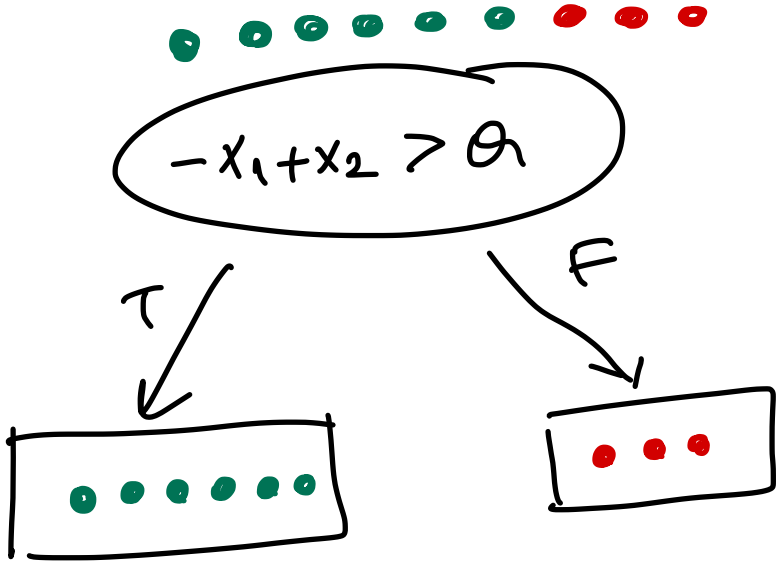
3rd split $x_1 > 2.5$

4th split $x_2 > 2.5$



$w_1 x_1 + w_2 x_2 + w_0 > 0$

$-1 \quad +1 \quad -\theta$

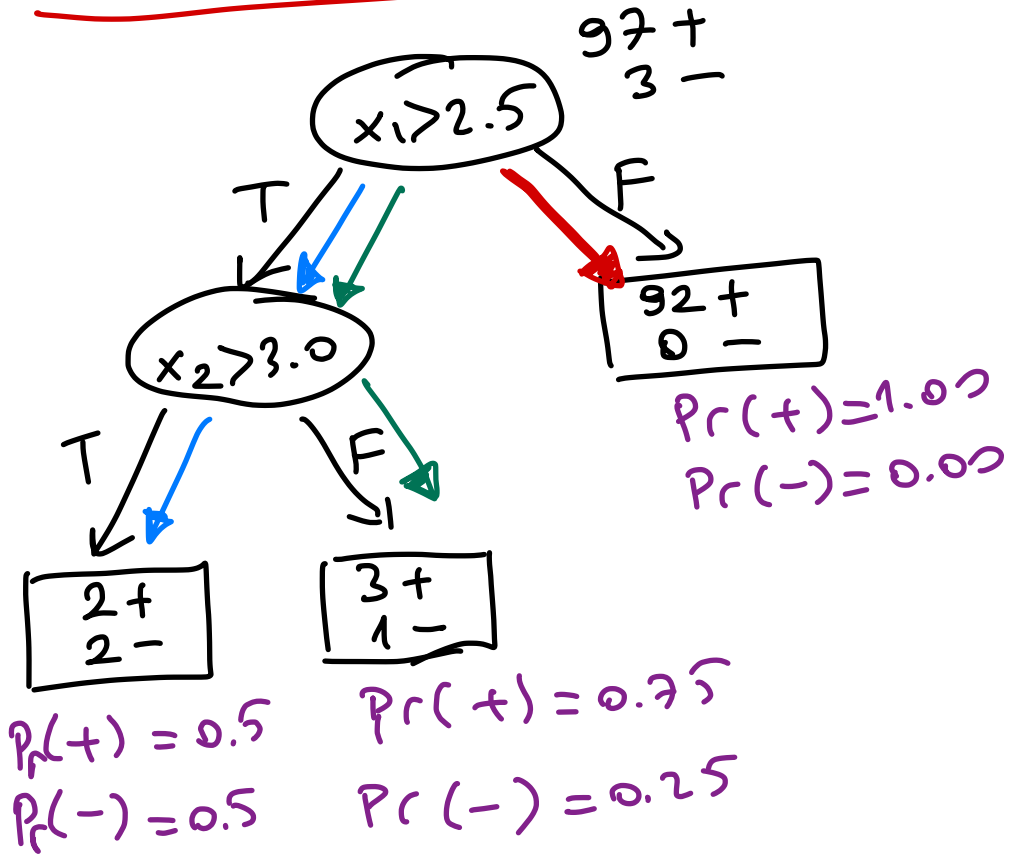


univariate $f_m(\mathbf{x})$: $x_j > w_{m0} \Rightarrow x_j - w_{m0} > 0$

$[0\ 0\ 1\ 0\ 0] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} - w_{m0} > 0$

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

RULE EXTRACTION

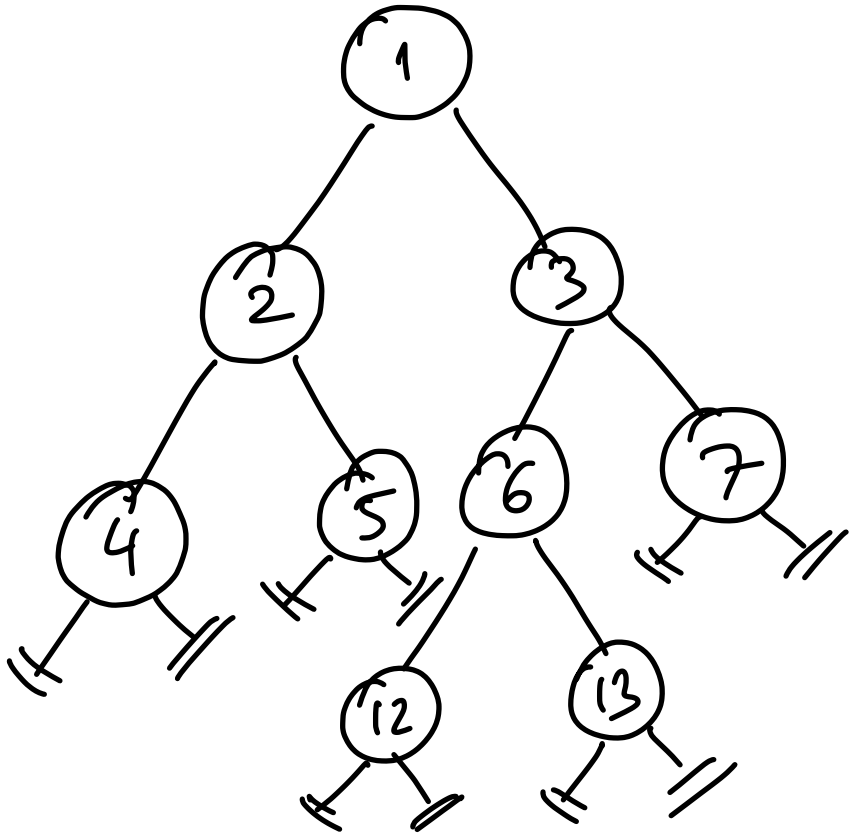


# of rules = # of leaves

Path 1 $x_1 > 2.5 \wedge x_2 > 3.0$?

Path 2 $x_1 > 2.5 \wedge x_2 \leq 3.0$ +

Path 3 $x_1 \leq 2.5$ +



left child = 2 * parent

right child = 2 * parent + 1

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

floor function