Univariate Decision Trees:

feature index

m $\quad j, w_{m0}$ threshold

20 R, 40 G, 40 B

$R_m = \{x \mid x_j \leq w_{m0}\}$ → 20

$L_m = \{x \mid x_j > w_{m0}\}$ → 80

N_m = # of data points that reach node m

$$\left.\begin{array}{l} 20 \; N_{m1} \\ 40 \; N_{m2} \\ \vdots \\ 40 \; N_{mK} \end{array}\right\} N_{mc} = \text{\# of data points that reach node } m \text{ from class } c$$

$$\left.\begin{array}{l} 80 \; N_{m1} \\ 20 \; N_{m2} \\ \vdots \\ N_{mS} \end{array}\right\} N_{ms} = \text{\# of data points that reach node } m \text{ and take split } s$$

flow from class c

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

total flow

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

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

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

$$I_m = -\sum_{c=1}^{K} p_{mc} \cdot \log_2(p_{mc}) \quad \text{impurity of node } m.$$

$$I_m' = -\sum_{s=1}^{S} \left[\underbrace{\frac{N_{ms}}{N_m}}_{\text{weight of child node}} \underbrace{\left[\sum_{c=1}^{K} p_{msc} \cdot \log_2(p_{msc}) \right]}_{L_m,\ R_m} \right] \quad \text{impurity of the split at node } m.$$

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

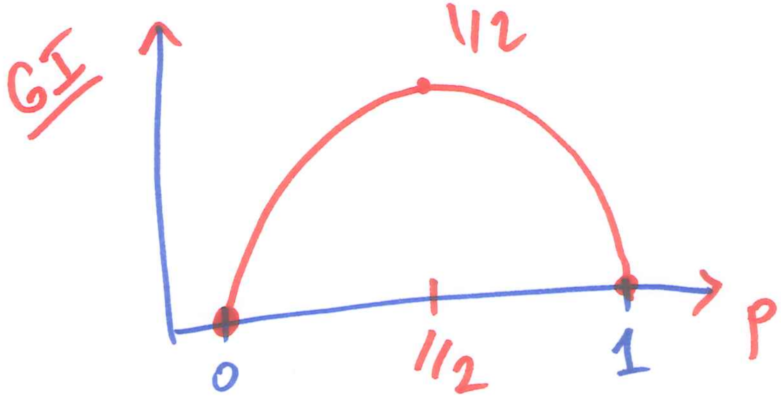
Entropy: $-p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p)$ N_+ N_-

p: ratio of positive data points

$1-p$: ratio of negative data points

$$p = \frac{N_+}{N_+ + N_-}$$

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



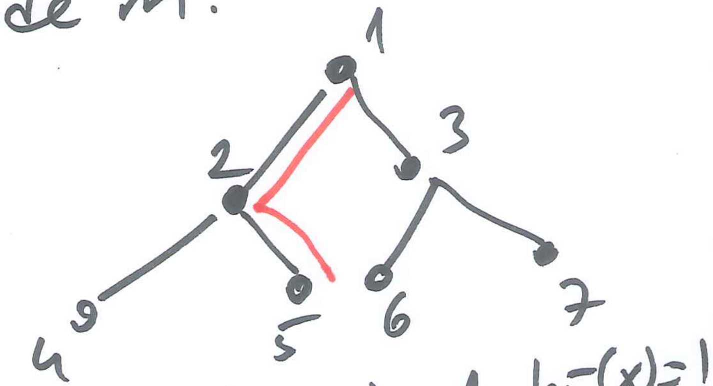
p_1	p_2	p_3
0.3	0.6	0.1

Misclassification Error: $1 - \max(p, 1-p)$ or $\min(p, 1-p)$

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

Regression Trees:

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



$b_1(x) = 1$ $b_2(x) = 1$ $b_5(x) = 1$

$b_3(x) = 0$ $b_4(x) = 0$ $b_6(x) = 0$

$b_7(x) = 0$

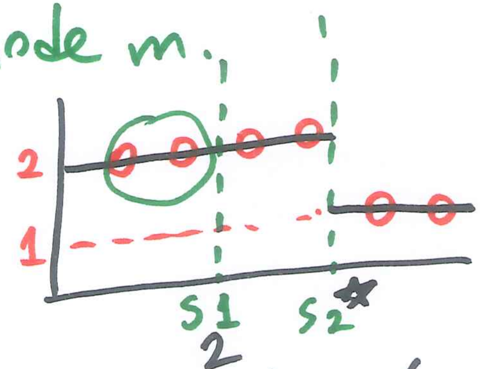
$$E_m = \frac{1}{N_m} \sum_{i=1}^{N} (y_i - g_m)^2 \, b_m(x_i)$$

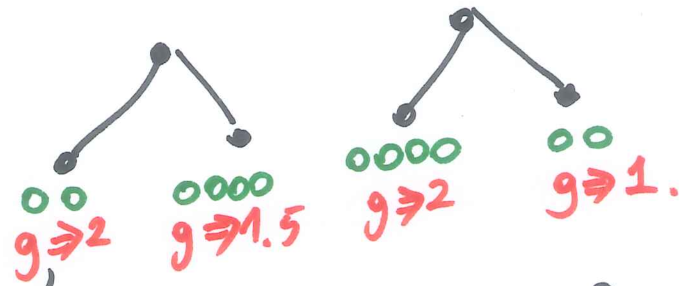
↳ # of data points that reach to node m.

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

g_m: predicted value at node m.

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





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

$0 < E_{s1} = (2-2)^2 + (2-2)^2 + (2-1.5)^2 + (2-1.5)^2 + (1-1.5)^2 + (1-1.5)^2$

$0 = E_{s2} = (2-2)^2 + (2-2)^2 + (2-2)^2 + (2-2)^2 + (1-1)^2 + (1-1)^2$

Multivariate Trees:

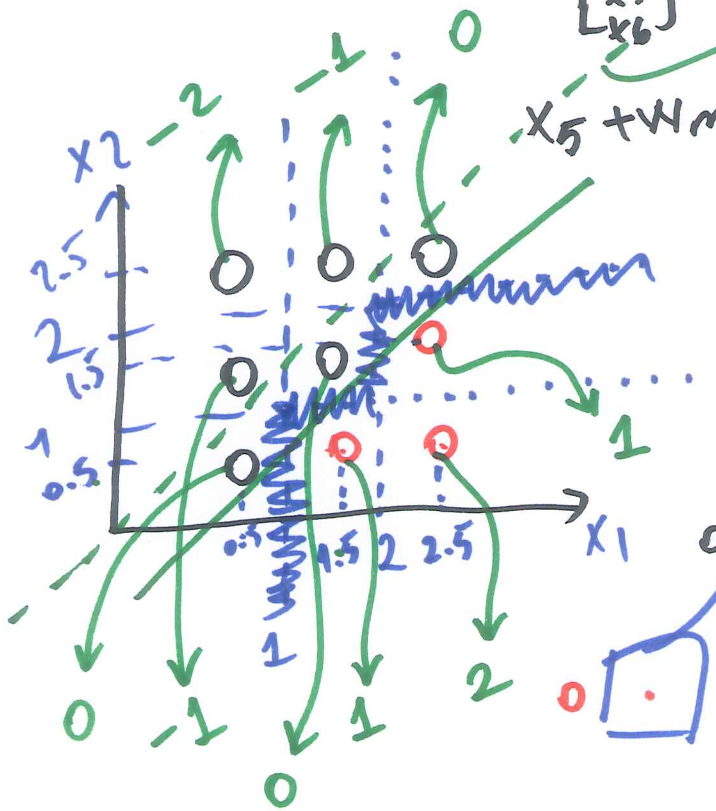
$f_m(x): \; x_{\underline{j}} > \underline{w_{m0}}$ ⇐ univariate split

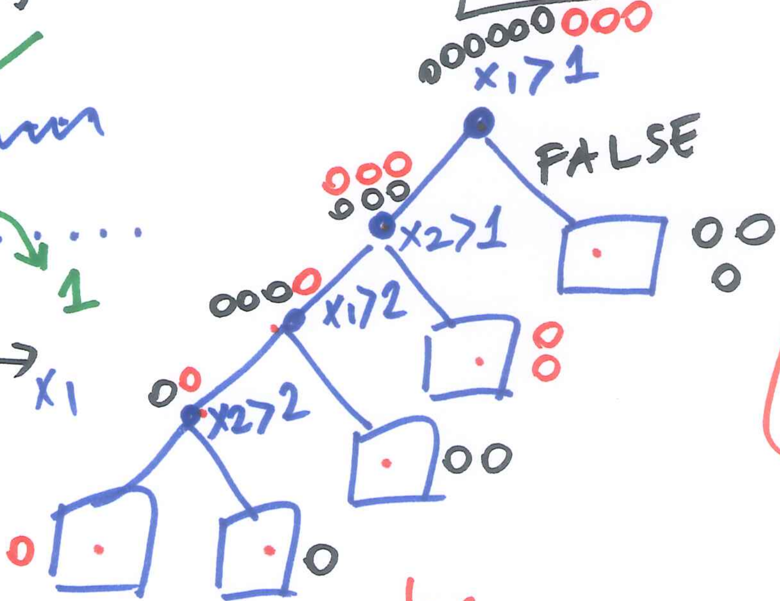
$f_m(x): \; w_m^T \cdot x + w_{m0} > 0$ ⇐ multivariate split.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}$$

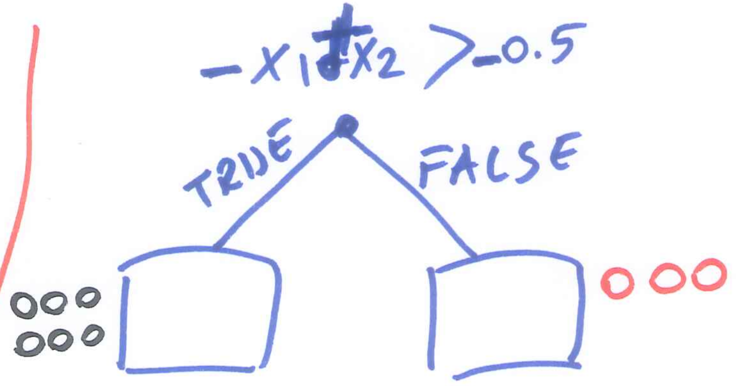
$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & -2 \end{bmatrix} \rightarrow x_1+x_2+x_3+x_4+x_5-2x_6+w_{m0}>0$$

$x_5 + w_{m0} > 0 \Rightarrow \boxed{x_5 > -w_{m0}}$





9 nodes
4 decision / 5 terminal nodes

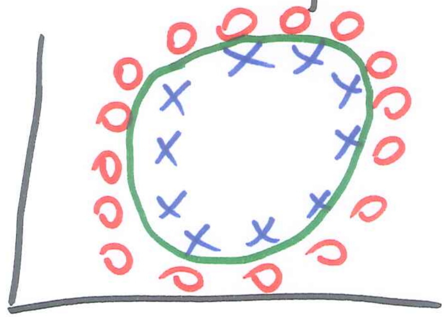


3 nodes
1 decision / 2 terminal nodes.

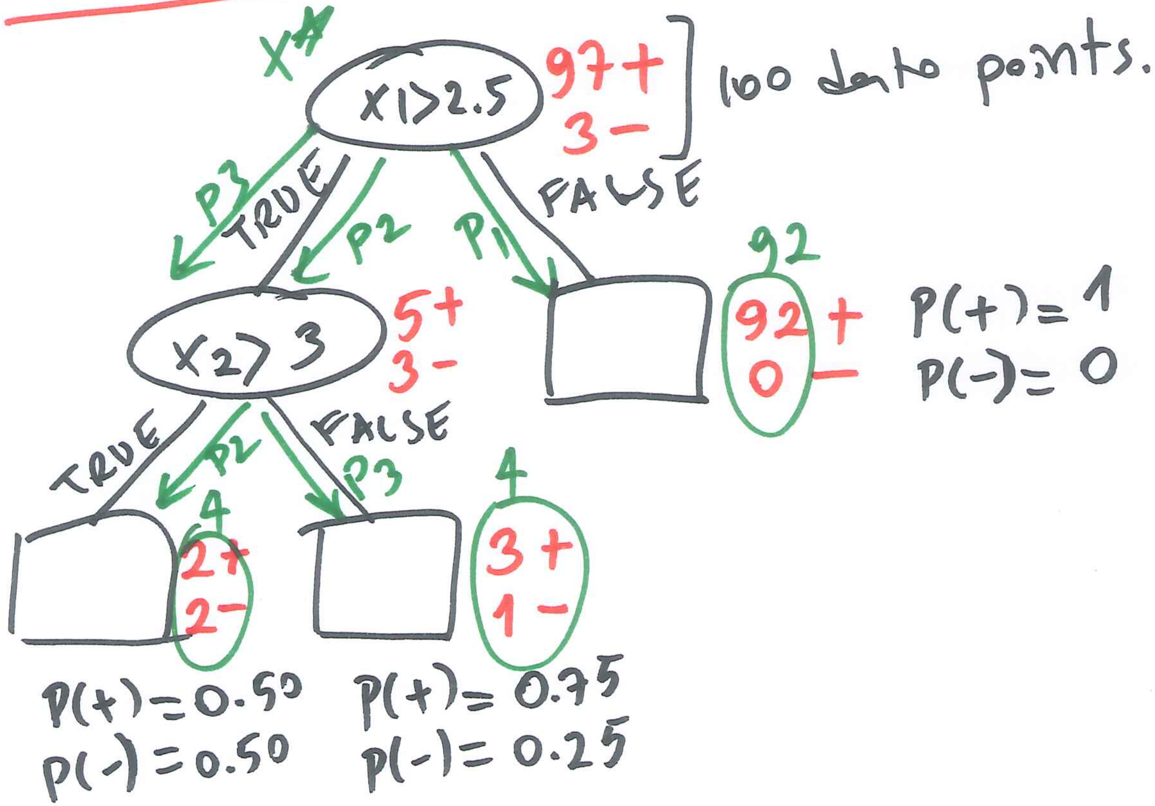
$$f_m(x) \Rightarrow w_m^T \cdot x + w_{m0} > 0$$

$$f_m(x) \Rightarrow x^T \cdot W_m \cdot x + w_m^T \cdot x + w_{m0} > 0 \Rightarrow \text{nonlinear / multivariate split.}$$

$$\underbrace{[\quad]}_{D\times D} \qquad D\times 1 \qquad 1\times 1$$



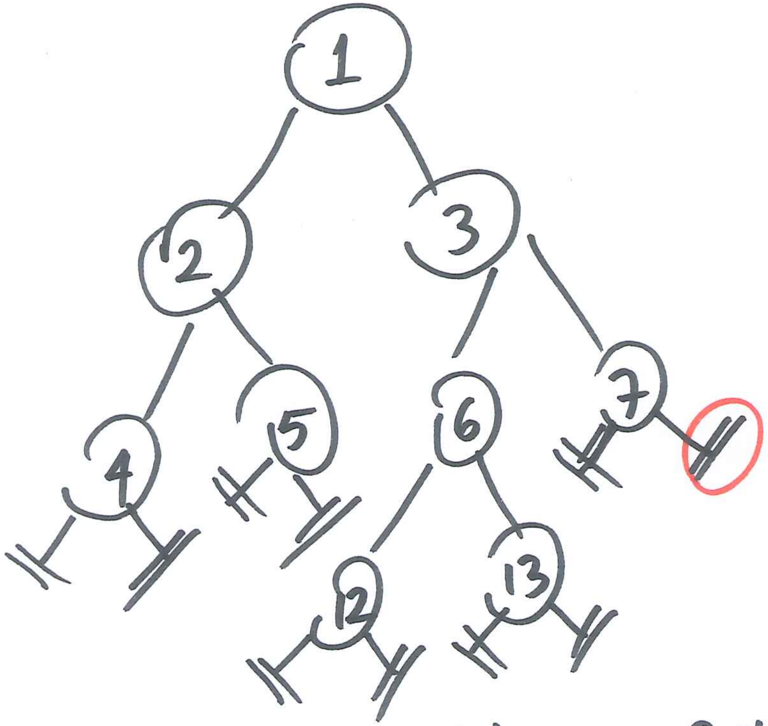
Rule Extraction:



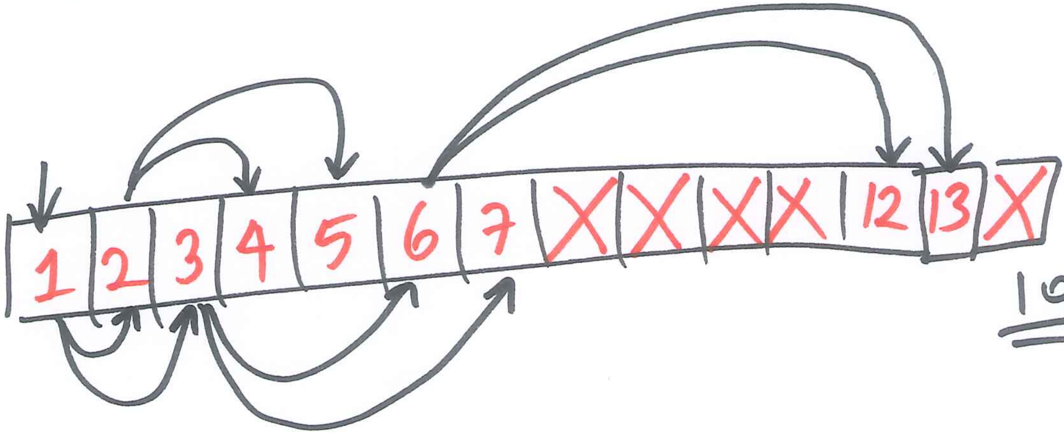
P1: $\boxed{x_1 \le 2.5}$

P2: $\boxed{x_1 > 2.5 \ \& \ x_2 > 3}$

P3: $\boxed{x_1 > 2.5 \ \& \ x_2 \le 3}$



terminel.



100

leftchild → 2 * parent

right child → 2 * parent + 1