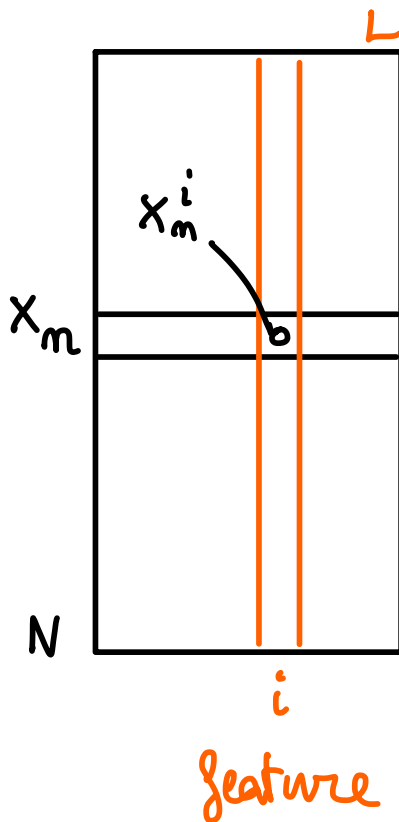


# XGBoost

data  $\{x_m\}$

sample



"labels"



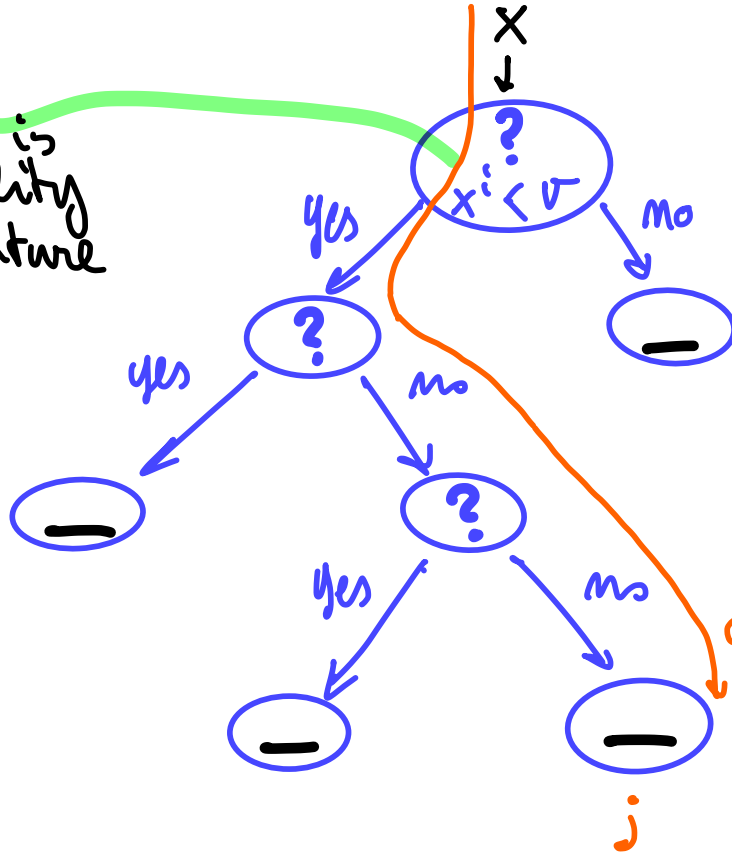
prediction

$$\hat{y}_m = G(x_m) \\ = \sum_{t=1}^T g_t(x_m)$$

linear combination  
of  $t \leq T$  contrib.,  
one per tree

# Prediction of $L$ -th tree

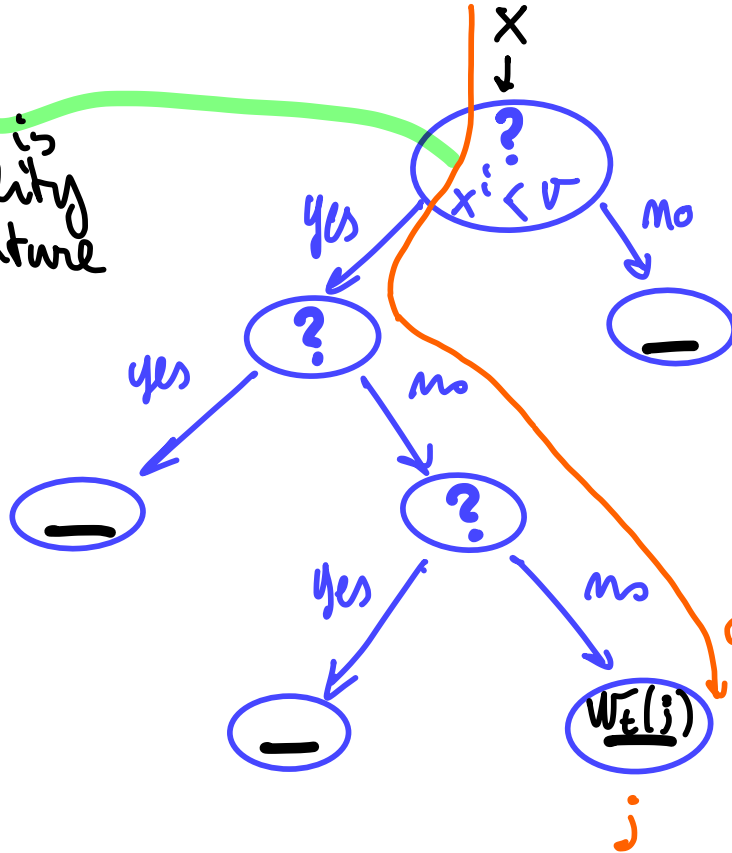
each check is  
an inequality  
for one feature



$g(x)$ : function mapping  
sample  $x$  to  
leaf  $j$

Prediction of  $t$ -th tree  $\Rightarrow g_t(x) = w_t(q(x))$

each check is  
an inequality  
for one feature



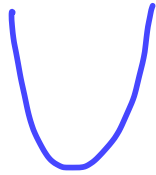
$q(x)$ : function mapping  
sample  $x$  to  
leaf  $j$

# Questions

- $Q_1$  • how to find best split ( $x^i < v$ ) for every node?
- $Q_2$  • when to stop splitting and set a leaf?
- $Q_3$  • why a linear combination of tree's predictions? ( $G = \sum_{t=1}^T g_t$ )

# Loss function

$$C(\{x_m\}, G) = \sum_{m=1}^N \ell(y_m, \hat{y}_m) + \sum_{t=1}^T \Omega(\theta_t)$$



convex

loss function

$N \geq m$  samples

e.g. square deviation  $\frac{1}{2}(y - \hat{y})^2$

• cross entropy for classification  $y = (0, 1)$

# Loss function

$$C(\{x_m\}, G) = \sum_{m=1}^N \ell(y_m, \hat{y}_m) + \sum_{t=1}^T \Omega(g_t)$$

regularization  
for every tree  $t \leq T$

$$\Omega(g_t) = \gamma \cdot J_t + \underbrace{\frac{\lambda}{2} \|w_t\|^2}_{\text{Ridge on } w\text{'s}}$$

"cost"  
of one  
leaf

# leaves

( $J_t = 4$  in the example)

# Loss function

$$C(\{x_m\}, G) = \sum_{m=1}^N \ell(y_m, \hat{y}_m) + \sum_{t=1}^T \Omega(\theta_t)$$

build  $C$  iteratively at every "time"  $t$

$t=1$



$t \gg 1$ : perturbation expansion, Taylor

Loss function, up to time  $\tau$

$$C_{\tau}(\{x_m\}, G) = \sum_{m=1}^N \ell(y_m, \hat{y}_m^{(\tau)}) + \sum_{t=1}^{\tau} \Omega(g_t)$$

$$\hat{y}_m^{(\tau)} = \hat{y}_m^{(\tau-1)} + g_t(x_m)$$

prediction  
up to tree  
 $\tau-1$

addition to  
prediction  
from tree  
 $\tau$

Taylor  
(answers)  
( $R_3$ )

$$= C_{\tau-1} + \Delta C_{\tau}$$



Contribution to loss function from tree  $\tau$

Taylor expansion

$$\Delta C_\tau = \sum_m \left[ \left. \partial_z \ell(y_m, z) \right|_{z=\hat{y}_m^{(\tau-1)}} g_\tau(x_m) + \frac{1}{2} \left. \partial_z^2 \ell(y_m, z) \right|_{z=\hat{y}_m^{(\tau-1)}} g_\tau(x_m)^2 \right] + \mathcal{L}(g_\tau)$$

def.  $a_m, b_m$

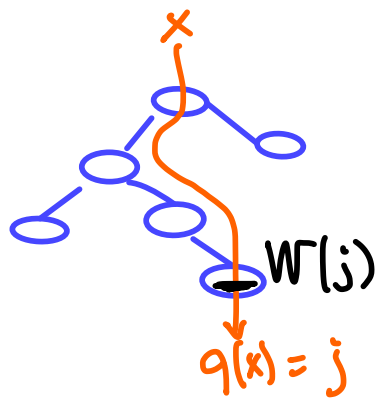
$$\equiv \sum_m \left[ a_m g_\tau(x_m) + \frac{1}{2} b_m g_\tau(x_m)^2 \right] + \mathcal{L}(g_\tau)$$

⚠ neglecting terms  $O(g_t^3)$

Contribution to loss function from tree  $\tau$

Taylor expansion

$$\Delta C_{\tau} = \sum_m \left[ a_m g_{\tau}(x_m) + \frac{1}{2} b_m g_{\tau}(x_m)^2 \right] + \mathcal{O}(g_{\tau})$$



all samples sent by tree  $\tau$  (via  $q_{\tau}(x_m)$ )  
to the same leaf  $j$   
will have the same  $g_{\tau}(x_m) = W_{\tau}(j)$

Contribution to loss function from tree  $\tau$

Taylor expansion

$$\Delta C_\tau = \sum_m \left[ a_m g_\tau(x_m) + \frac{1}{2} b_m g_\tau(x_m)^2 \right] + \Omega(g_\tau)$$

all samples sent by tree  $\tau$  (via  $g_\tau(x_m)$ )  
to the same leaf  $j$

will have the same  $g_\tau(x_m) = W_\tau(j)$

rewrite

$$\sum_{\substack{m \\ \text{samples}}} \dots = \sum_{\substack{j=1 \\ \text{leaves}}}^{J_\tau} \sum_{m \in \mathcal{I}_\tau(j)} \dots \quad \parallel \quad \begin{array}{l} \text{Indicator} \\ \mathcal{I}_\tau(j) = \{m \mid g_\tau(x_m) = j\} \end{array}$$

Contribution to loss function from tree  $\tau$

Taylor expansion

$$\Delta C_{\tau} = \sum_m \left[ a_m g_{\tau}(x_m) + \frac{1}{2} b_m g_{\tau}(x_m)^2 \right] + \underbrace{\Omega(g_{\tau})}_{\text{regularization}}$$

$$= \sum_{j=1}^{J_{\tau}} \left[ \underbrace{\left( \sum_{m \in I_{\tau}(j)} a_m \right)}_{A_{\tau}(j)} w_{\tau}(j) + \frac{1}{2} \left\{ \underbrace{\left( \sum_{m \in I_{\tau}(j)} b_m \right)}_{B_{\tau}(j)} + \underbrace{\lambda}_{\text{regularization}} \right\} w_{\tau}(j)^2 \right] + \underbrace{\gamma J_{\tau}}_{\text{regularization}}$$

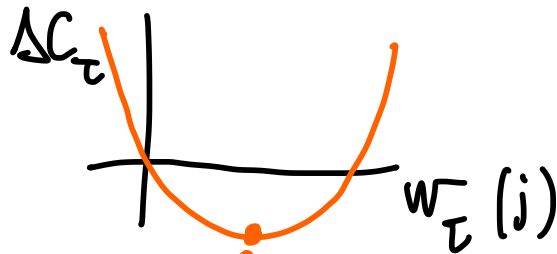
Contribution to loss function from tree  $\tau$

$$\Delta C_{\tau} = \sum_{j=1}^{J_{\tau}} \left[ A_{\tau}(j) w_{\tau}(j) + \frac{1}{2} (B_{\tau}(j) + \lambda) w_{\tau}(j)^2 \right] + \gamma J_{\tau}$$

$\forall$  leaf  $j$ , parabola

$$\frac{B+\lambda}{2} w^2 + Aw + \gamma J$$

$$B+\lambda > 0$$



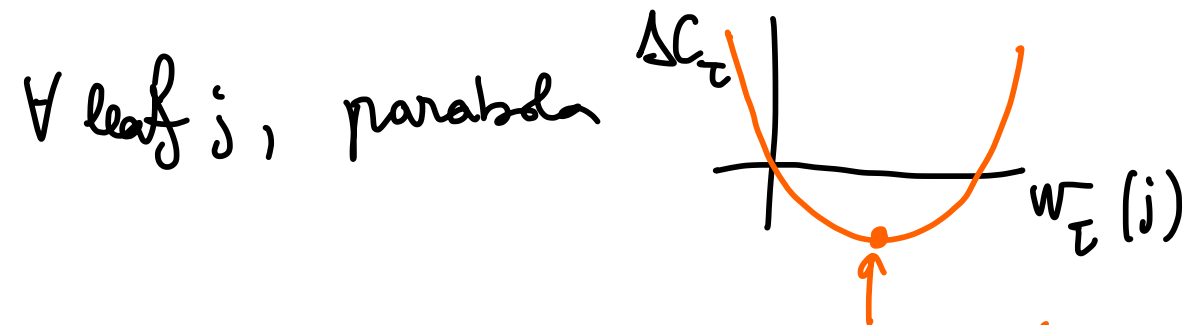
$$\text{best } w_{\tau}^*(j) = - \frac{A_{\tau}(j)}{B(j) + \lambda}$$

Contribution to loss function from tree  $\tau$

$$\Delta C_{\tau} = -\frac{1}{2} \sum_{j=1}^{J_{\tau}} \frac{A_{\tau}(j)^2}{B_{\tau}(j) + \lambda} + \gamma J_{\tau}$$

optimized  
with Newton's  
method ( $w^*$ )

⚠ very quick!  
no gradient descent



$$\text{best } w_{\tau}^*(j) = -\frac{A_{\tau}(j)}{B_{\tau}(j) + \lambda}$$

# Questions

Q<sub>1</sub>

- how to find best split ( $x^i < v$ ) for every node?

Q<sub>2</sub>

- when to stop splitting and set a leaf?



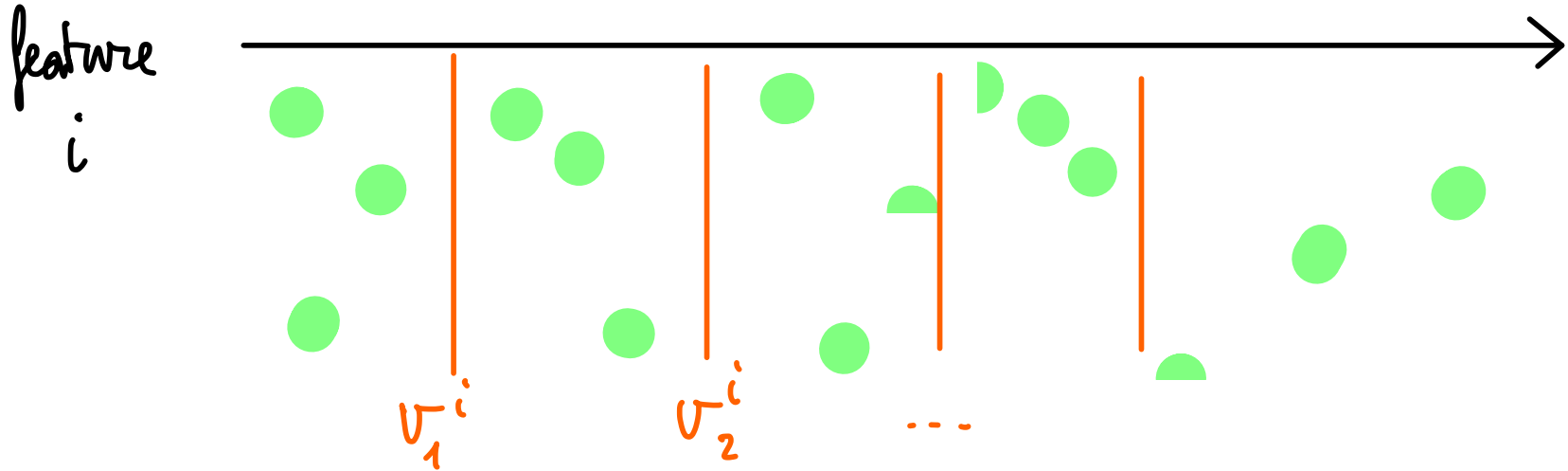
there is a huge amount of possible combinations (9, 9)

weights      tree  
                 structure

# Practical implementation of XGBoost

for many samples ( $N \gg 1$ ) with many features ( $L \gg 1$ )

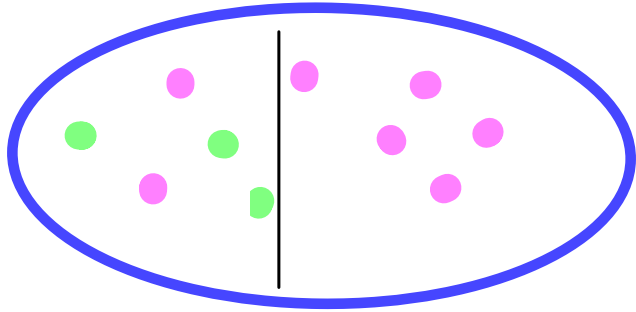
- partition each feature in percentiles,  $v^i$



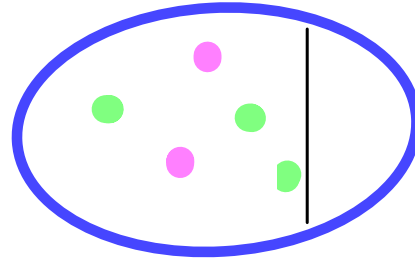


Split

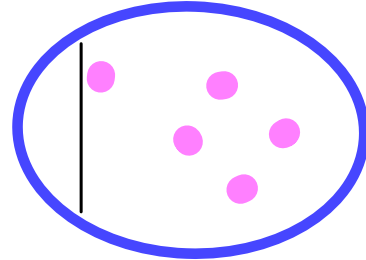
$v^i$



VS



$L$



$R$

Is the original leaf on the left better or worse than the two leaves on the right?

# Practical implementation of XGBoost

Greedy algorithm  
in every current leaf  $\Rightarrow$

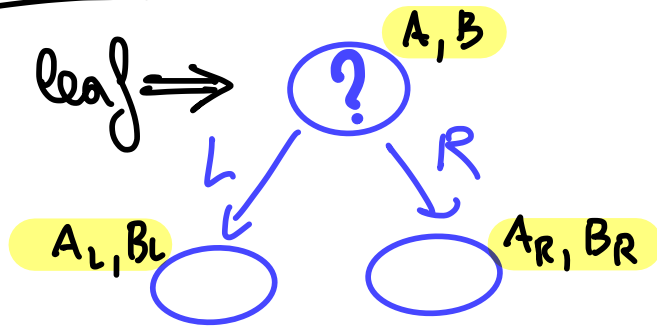
$A, B$   
?

try many splits  
( $\sqrt{A}$  percentile)

# Practical implementation of XGBoost

Greedy algorithm

in every current leaf  $\Rightarrow$



try many splits  
(V percentile)

choose the one providing smallest  $\Delta C_{\text{SPLIT}}$

$$\Delta C_{\text{SPLIT}} = -\frac{1}{2} \left( \frac{A_L^2}{B_L + \lambda} + \frac{A_R^2}{B_R + \lambda} - \frac{A^2}{B + \lambda} \right) + \gamma$$

L/R split

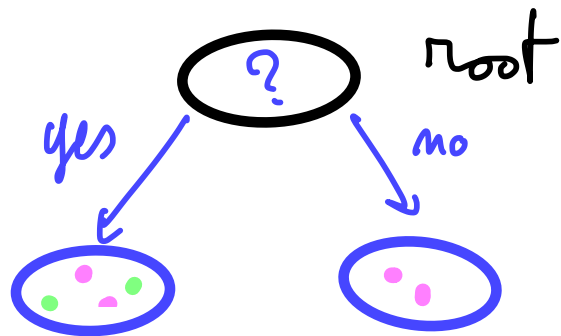
all together  $\Rightarrow$

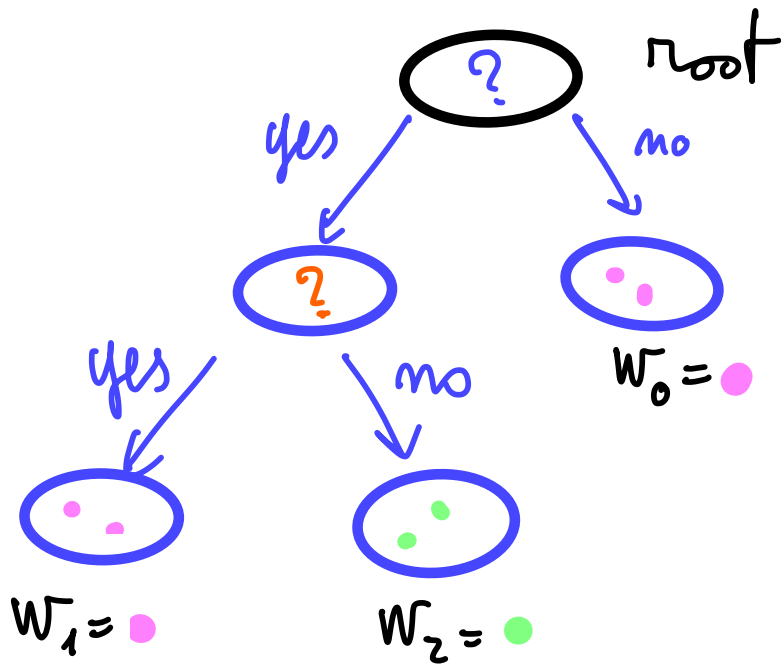
$Q_1 \checkmark$

- start from a root node
- iterate split for every node until
  - a) it is not convenient (all  $\Delta C_{\text{split}} > 0$ )
  - b) conditions max depth, etc., are met

$\Rightarrow Q_2 \checkmark$







this was too easy

normally the 1<sup>st</sup> tree does not solve it all, and the  $\hat{\eta}_m - \eta_m$  difference is the starting point for the next tree

Example of  $A, B$ , for square deviation loss f.

( $\hat{y}_m$  up to previous  
free)

$$\ell(y_m, \hat{y}_m) = \frac{1}{2} (\hat{y}_m - y_m)^2$$

$$a_m = \left. \partial_z \ell(y_m, z) \right|_{z=\hat{y}_m} = \hat{y}_m - y_m \quad \Rightarrow \quad A(j) = \sum_{m \in \mathcal{I}(j)} (\hat{y}_m - y_m)$$

$$b_m = \left. \partial_z^2 \ell(y_m, z) \right|_{z=\hat{y}_m} = 1 \quad \Rightarrow \quad B(j) = \sum_{m \in \mathcal{I}(j)} 1 =$$

= # samples sent to leaf  $j$  by  $g(x)$