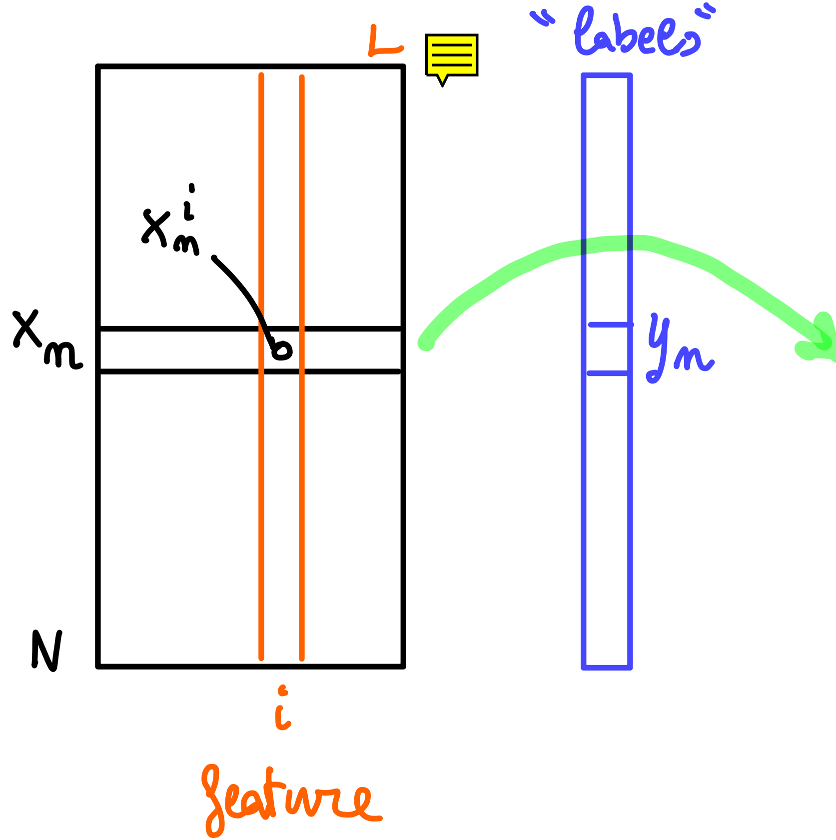
XGBoost

data $\{x_n\}$

sample



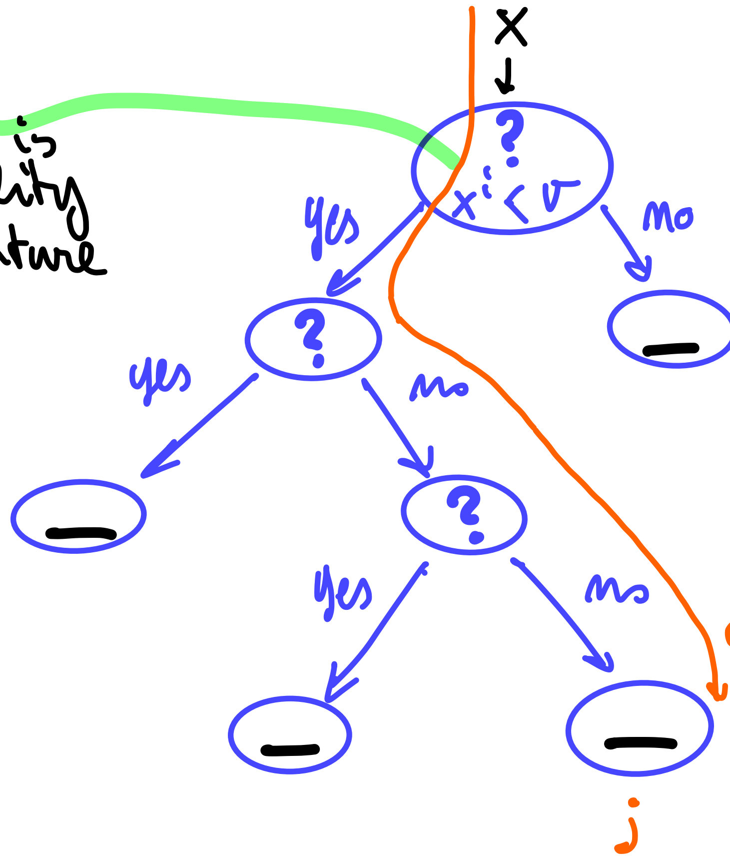
prediction

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

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

Prediction of t-th tree

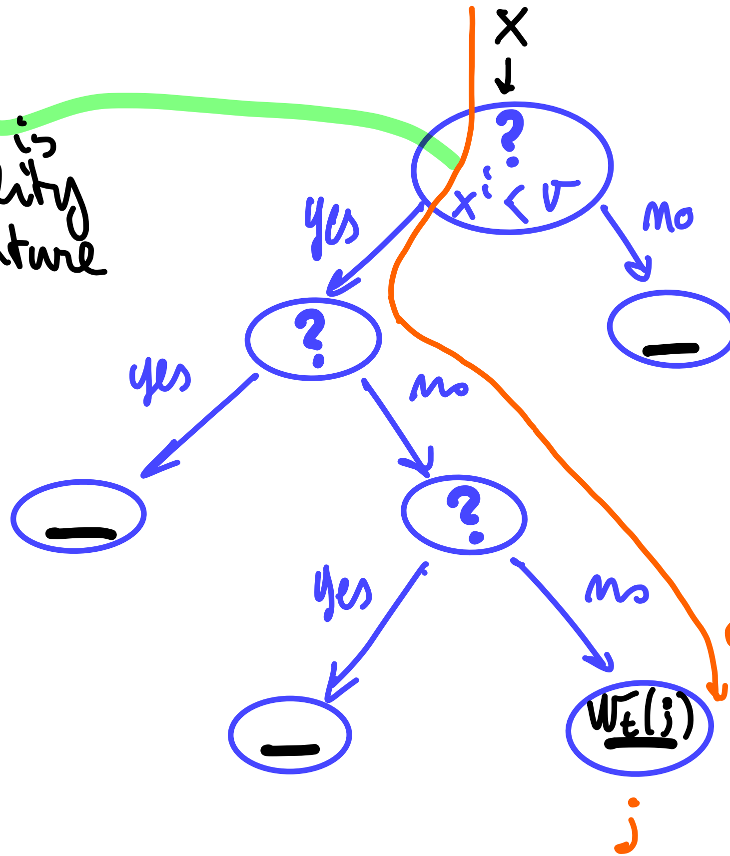
each check is an inequality for one feature



$q(x)$: 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



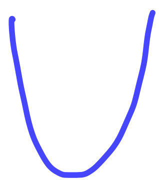
q(x): function mapping sample x to leaf j

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? $\left(G = \sum_{t=1}^{T} g_t \right)$

Loss function

$$C(\{x_n\}, G) = \sum_{n=1}^{N} \underbrace{\ell(y_n, \hat{y}_n)}_{\text{convex loss function}} + \sum_{t=1}^{T} \Omega(g_t)$$



convex
loss function
$\forall$ sample $n \leq N$

e.g.:

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

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

Loss function

$$C(\{x_n\}, G) = \sum_{n=1}^{N} \ell(y_n, \hat{y}_n) + \underbrace{\sum_{t=1}^{T} \Omega(g_t)}_{\text{regularization for every tree } t \leq T}$$

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

"cost" of one leaf

\# leaves

($J_t = 4$ in the example)

Loss function

$$C(\{x_n\}, G) = \sum_{n=1}^{N} \ell(y_n, \hat{y}_n) + \sum_{t=1}^{T} \Omega(g_t)$$

build C iteratively at every "time" t

$t=1$

$$C_1 \rightarrow C_2 \rightarrow \cdots \rightarrow C_{t-1} \rightarrow C_t \rightarrow \cdots \rightarrow C_T \equiv C$$

$t \gg 1$: perturbation expansion, Taylor

Loss function, up to time τ

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

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

prediction up to tree $\tau-1$

addition to prediction from tree τ

Taylor

(answers Q_3)

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

Contribution to loss function from tree τ

Taylor expansion

$$\Delta C_\tau \simeq \sum_n \left[\partial_z \ell(y_n, z)\Big|_{z=\hat{y}_n^{(\tau-1)}} g_\tau(x_n) + \frac{1}{2}\partial_z^2 \ell(y_n, z)\Big|_{z=\hat{y}_n^{(\tau-1)}} g_\tau(x_n)^2 \right] + \Omega(g_\tau)$$

def. a_n, b_n

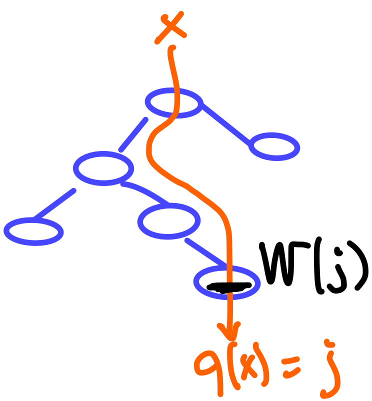
$$\stackrel{\text{def. } a_n, b_n}{\equiv} \sum_n \left[a_n \, g_\tau(x_n) + \frac{1}{2} b_n \, g_\tau(x_n)^2 \right] + \Omega(g_\tau)$$

⚠ neglecting terms $O(g_t^{\,3})$

Contribution to loss function from tree τ

Taylor expansion

$$\Delta C_\tau = \sum_n \left[a_n \, g_\tau(x_n) + \frac{1}{2} b_n \, g_\tau(x_n)^2 \right] + \Omega(g_\tau)$$



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

Contribution to loss function from tree τ

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 τ (via $q_\tau(x_m)$)
to the same leaf j
will have the same $g_\tau(x_m) = W_\tau(j)$

rewrite

$$\sum_{m \atop \text{samples}} \cdots = \sum_{j=1 \atop \text{leaves}}^{J_\tau} \sum_{m \in I_\tau(j)} \cdots$$

Indicator

$$I_\tau(j) = \{ m \,|\, q_\tau(x_m) = j \}$$

Contribution to loss function from tree τ

Taylor expansion

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

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

$$+ \gamma J_\tau$$

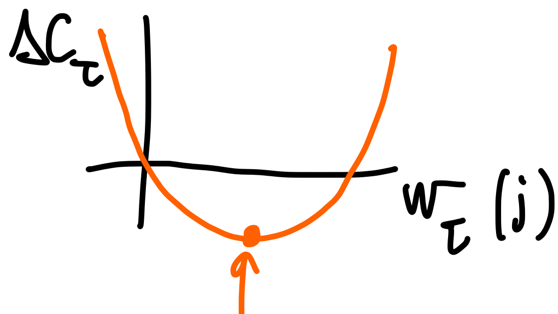
Contribution to loss function from tree τ

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

$\forall$ leaf j, parabola

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

$$B + \lambda > 0$$



best $w^*_\tau(j) = -\dfrac{A_\tau(j)}{B(j) + \lambda}$

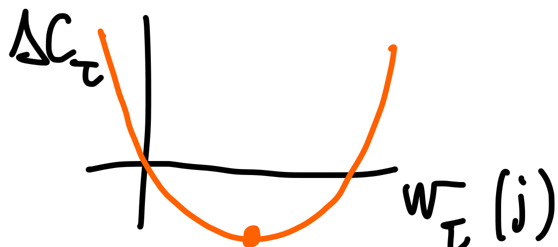
Contribution to loss function from tree τ

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

optimized with Newton's method (w^*)

⚠ very quick! no gradient descent

$\forall$ leaf j, parabola



best $w^*_\tau(j) = -\dfrac{A_\tau(j)}{B(j)+\lambda}$

Questions

$\boxed{Q_1}$ • how to find best split $(x^i < v)$ for every node?

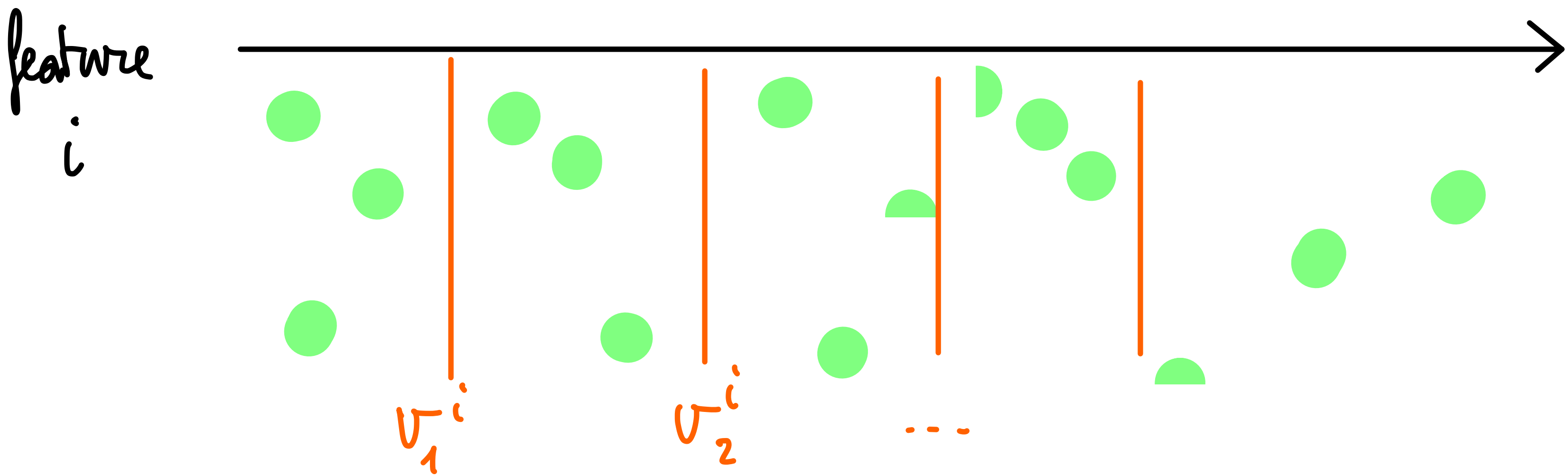
$\boxed{Q_2}$ • when to stop splitting and set a leaf?

⚠ There is a huge amount of possible combinations $(\underbrace{g}_{\text{weights}}, \underbrace{q}_{\text{tree structure}})$

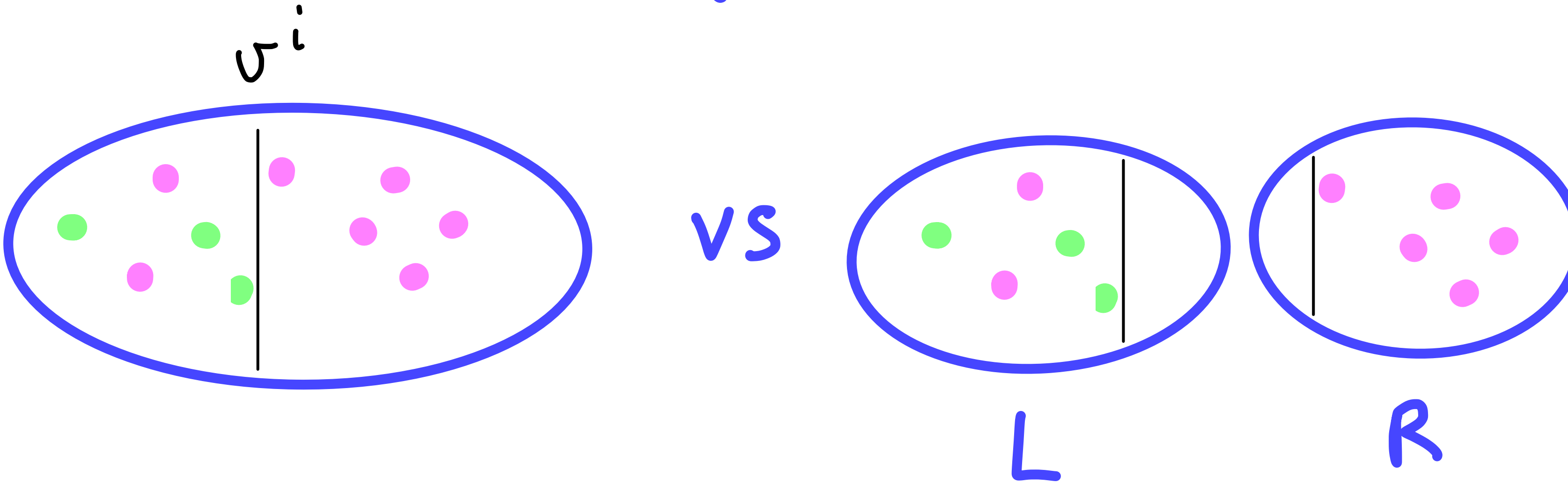
Practical implementation of XGBoost

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

- partition each feature in percentiles v^i



Split



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

Practical implementation of XGBoost

<u>Greedy algorithm</u>

in every current leaf $\Longrightarrow$? A, B

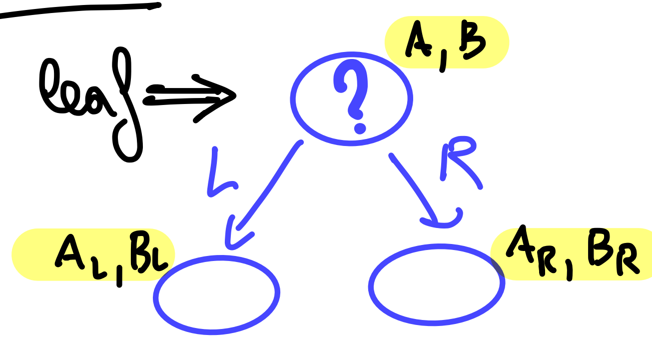
try many splits

($\forall$ percentile)

Practical implementation of XGBoost

Greedy algorithm

in every current leaf $\Longrightarrow$



try many splits
(∀ percentile)

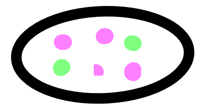
choose the one providing smallest ΔC_{SPLIT}

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

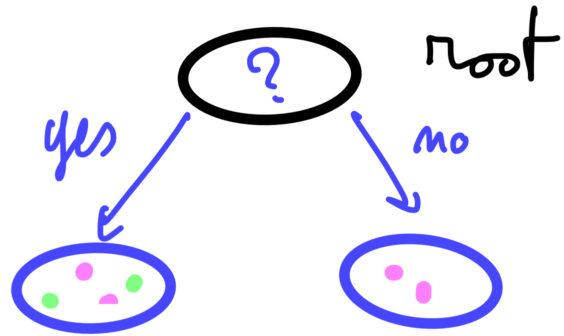
all together $\Longrightarrow Q_1$ ✓

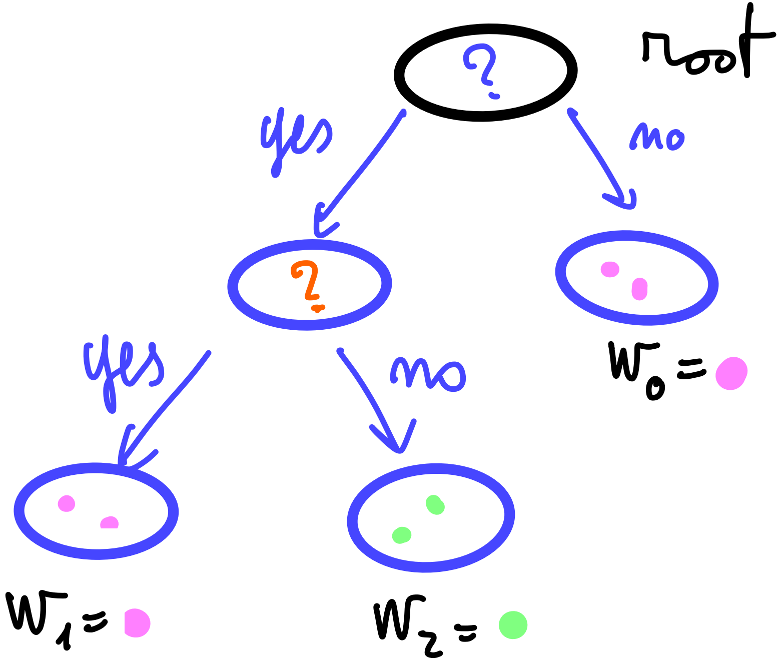
- start from a root node
- iterate split for every node until
  a) it is not convenient (all $\Delta C_{split} > 0$)
  b) conditions max depth, etc., are met

$\Rightarrow Q_2 \checkmark$



root


?
root
yes
no



this was too easy

normally the 1st tree does not solve it all, and the $\hat{y}_n - y_n$ difference is the starting point for the next tree

Example of A, B, for square deviation loss ℓ.

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

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

$$a_n = \partial_z \ell(y_n, z)\Big|_{z=\hat{y}_n} = \hat{y}_n - y_n \implies A(j) = \sum_{n \in I(j)} (\hat{y}_n - y_n)$$

$$b_n = \partial_z^2 \ell(y_n, z)\Big|_{z=\hat{y}_n} = 1 \implies B(j) = \sum_{n \in I(j)} 1 =$$

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