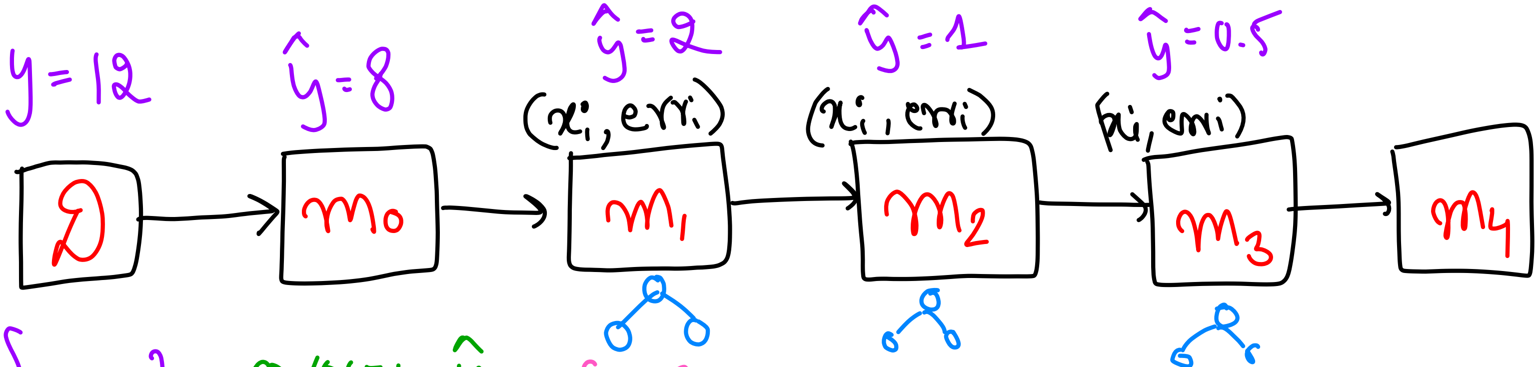
BOOSTING-2



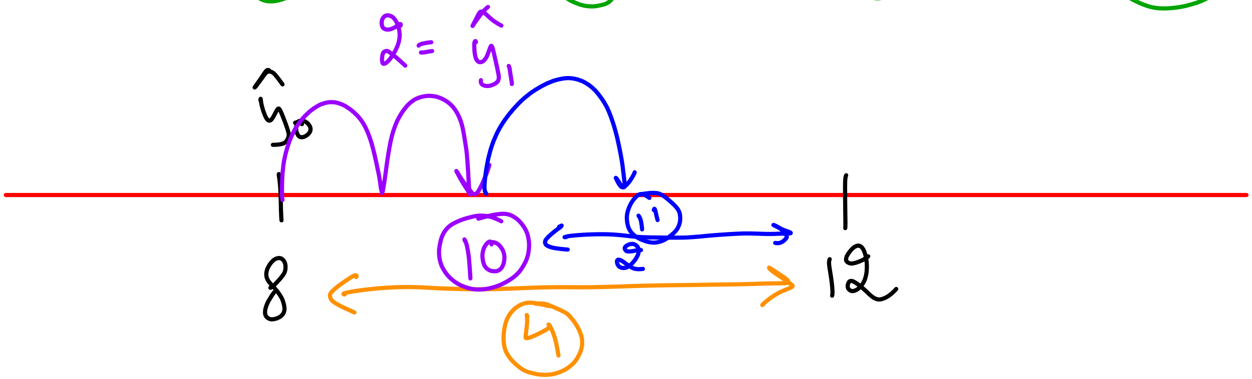
$\{x_i, y_i\}$

error $= y - \hat{y}$
$= 12 - 8$
$= 4$

$F_0 = 8$
$F_1 = 8 + 2$
error $= 2$

$F_2 = F_1 + 1$
error $= 1$

$F_3 = F_2 + 0.5$
error $= 0.5$



$F_0(x) = 8$
$F_1(x) = 8 + 2$
$F_2(x) = 8 + 2 + 1$
$= 11$

f_1	f_2	f_3	y	m_0 $\hat{y}_0$	err_0	m_1 $\hat{y}_1$ (f_1, f_2, f_3, err_0)	err_1	m_2 $\hat{y}_2$ (f_1, f_2, f_3, err_1)	err_2	m_3 $\hat{y}_3$		
			3	5.5	-2.5	-1.5	-1	-0.5	-0.5			
			4	5.5	-1.5	-0.5	-1	-0.3	-0.7			
			6	5.5	0.5	0.5	0	0	0			
			9	5.5	3.5	2	1.5	1	0.5			

mean of y's

$y - \hat{y}_0 - \hat{y}_1$

$y - \hat{y}_0 - \hat{y}_1 - \hat{y}_2$

$12 - 8 - 2 = 2$

$$\boxed{err_m = y - \left[\hat{y}_0 + \sum_{i=1}^{k} \hat{y}_i\right]}$$

$$err_0 = y - \hat{y}_0$$

$$y = err_m + \left[\hat{y}_0 + \sum_{i=1}^{k} \hat{y}_i\right]$$

$$err_1 = y - \hat{y}_0 - \hat{y}_1$$

$$err_2 = y - \hat{y}_0 - \hat{y}_1 - \hat{y}_2$$

Boosting

$$D_{Train} = x_i$$

Step 1

Create m_0

$$\underbrace{f_0(x)}_{\text{function value after stage 0}} = \underbrace{\hat{y}_0}_{\text{predicted value at stage 0}} = \underline{Avg(x)} = \underbrace{h_0(x)}_{\text{output of } m_0}$$

$$err_0 = y - \hat{y}_0$$

$$y = \underbrace{\hat{y}_0}_{h_0(x)} + err_0$$

$$\underline{y} = \underbrace{\hat{y}_i}_{h_i(x)} + err_i$$

For each datapoint $\{x_i, y_i, err_i\}_{i=1}^{k}$

Step 2 Create $\underline{\underline{m1}}$ → Train on $\{x_i, err_0\}$

input label/Target

$$F_1(x) = h_0(x) + \gamma_1 h_1(x)$$

~~Step 3~~ Create m_2 $\{x_i, \underbrace{err_1^{(i)}}\}$

$\hookrightarrow y - F_1(x)$

$$F_2(x) = h_0(x) + \underbrace{\gamma_1 h_1(x) + \gamma_2}_{\text{weights}} h_2(x)$$

$\hookrightarrow$ weights

~~Step G~~

$$F_G(x) = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x) + \ldots \gamma_G h_G(x)$$

$$\boxed{F_G(x) = h_0(x) + \sum_{i=1}^{G} \gamma_i h_i(x)}$$

Linear Regression / Gradient Boosting

$$F_3(x) = \underbrace{h_0(x)}_{\text{Const}} + \gamma_1 h_1(x) + \gamma_2 h_2(x) + \gamma_3 h_3(x)$$

learning γ's is a sequential process.

$$\hat{y} = \underbrace{w_0}_{\text{Const}} + w_1 x_1 + w_2 x_2 + w_3 x_3$$

Learning w's is a parallel process

errors | Residuals

$$F_k(x) = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x) \cdots \gamma_k h_k(x)$$

Residual $= y_i - F_k^i(x)$ $\qquad k = 2$

12 $\qquad F_k(x) = 10$

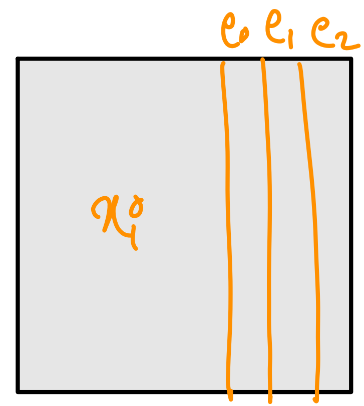
Training for $(k+1)^{th}$

↳ m

Squared loss

$$F_k(x) \Rightarrow \hat{y}_k \Rightarrow (y - \hat{y}_1)^2$$

Loss function



$$D = \{x_i, \text{error}_k^i\}$$

$$L(y, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

$$L(y, F_k(x)) = (y_i - F_k(x))^2 = 2^2 = 4$$

12 10

Let's $f_k(x) = z_i$

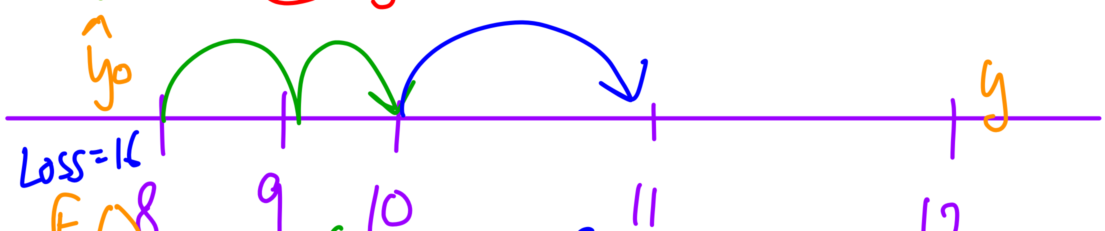
error as function

$err_0 = 4$ m_1 $\hat{y}_1 = 2$ m_2 $\hat{y}_2 = 1$

$L(y, F_k(x))$

$k = 2$

$F_2(x) = 11$



$F_0(x) = 8$ $f_1(x) = 10$ $F_2(x) = 11$

$L(y, f_2(x)) = (12 -$

$L(y, f_2(x)) = (y - f_2(x))^2$

$L(y, f_k(x)) = (y - f_k(x))^2$

$f_k(x) = z_i$

$L(y, z_i) = (y - z_i)^2$

$$\frac{\partial L}{\partial z_i} \Rightarrow \frac{\partial}{\partial z_i}(y_i - z_i)^2$$

$$= -2(y_i - z_i)$$

$$\frac{\partial L}{\partial z_i} = -2(y_1 - z_i)$$

$$-\frac{\partial L}{\partial z} = 2(y_i - z_i)$$

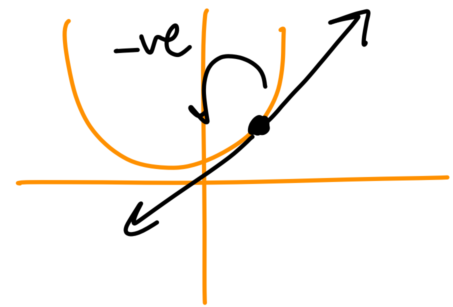
Const

neg grad.

$$-\frac{\partial L}{\partial z_i} \approx (y_i - z_i)$$

$$-\frac{\partial L}{\partial z_i}$$

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w}$$



$$-\frac{\partial L}{\partial F_k(x_i)} \approx \underbrace{\left(y^{(i)} - f_k(x_i)\right)}_{\text{error / Residual}}$$

pseudo error

L = Loss $\rightarrow$ Boosted trees

$$h_k(x) = \left\{ x^{(i)}, \text{err}^{i}_{k-1} \right\}$$

pseudo error

Loss function. Sq. Loss.

$(y-\hat{y})^2$

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

$y = 2, 4, 6$

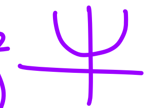
Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

$y = 2, 4, 6$

$L = (2-\gamma)^2 + (4-\gamma)^2 + (6-\gamma)^2$

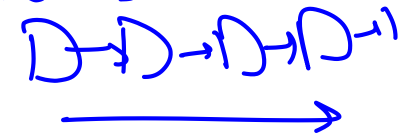


$\frac{\partial L}{\partial \gamma} = \cancel{-2}(2-\gamma) \cancel{-2}(4-\gamma) \cancel{-2}(6-\gamma) = 0$

$\cancel{-2}(2-\gamma + 4-\gamma + 6-\gamma) = 0$

$12 - 3\gamma = 0 \Rightarrow \gamma = 4$

$m_0 \quad m_1 \; - - - \quad M$



2. For m = 1 to M:

   1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

   2. Fit a base learner (or weak learner, e.g. tree) closed under scaling $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.

   3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

m_1

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

$m = 1$

$L(y_i, F_{m-1}(x) + \gamma h_m(x_i)) =$

   4. Update the model:

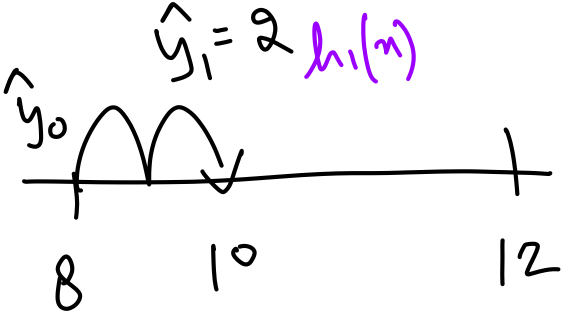
$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

$F_1(x) = F_{1-1}(x) + 2\, h_1(x)$

$[y_i - F_{m-1}(x) - \gamma h_m(x_i)]^2$

$12 = F_1(x) = 8 + 2 \times 2 = 12$



$F_0(x)$ $\quad F_1(x) = 10$

$$m = 1 \qquad [y_i - F_{1-1}(x) - \gamma h_1(x)]^2$$

$$f(\gamma) = [y_i - F_0(x) - \gamma\, h_1(x)]^2$$

$$\frac{\partial}{\partial \gamma} = -2[y_i - F_0(x) - \gamma\, h_1(x)]\, h_1(x)$$

$$\cancel{-2}[12 - 8 - \gamma 2]\cancel{2} = 0$$

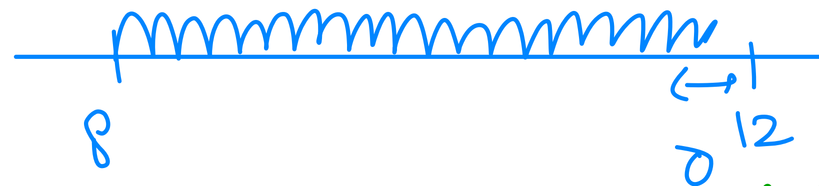
$$4 = \gamma 2 \Rightarrow \gamma_1 = 2$$

$$y = h_0(x) + \gamma_1 h_1(x)$$

$$y = 8 + (2 \times 2) \Rightarrow 8 + 4$$

$$\Rightarrow 12$$

~~Hyper parameter~~



$$m_0 \quad m_1 \quad m_2 \quad \cdots\cdots \quad \boxed{\underline{m_{1000}}} \rightarrow 0$$

# of model

M ↑ → error tend to 0 → Overfitting

M ↓ → Values will be close to mean value

Underfitting

Value of d → depth of each tree

base learner

d ↑ → Overfitting

d ↓ → Underfitting

Regularisation by Shrinkage

$$F_M(x) = h_0(x) + \nu \sum_{i=1}^{M} \gamma_m h_m(x)$$

M → hyperparameter

underfitting

Overfitting

ν → Learning Rate $[0 < \nu \leq \infty]$

recommend

Problem

* Slow
* Sequential
* It overfits very Quickly.

Randomisation → RS + CS

GBDT = Pseudo Residual + Additive Combining

RS + CS + PR + AC

Stochastic GBDT

Categorical data

Catboost

Adaboost

LIGHTGBM

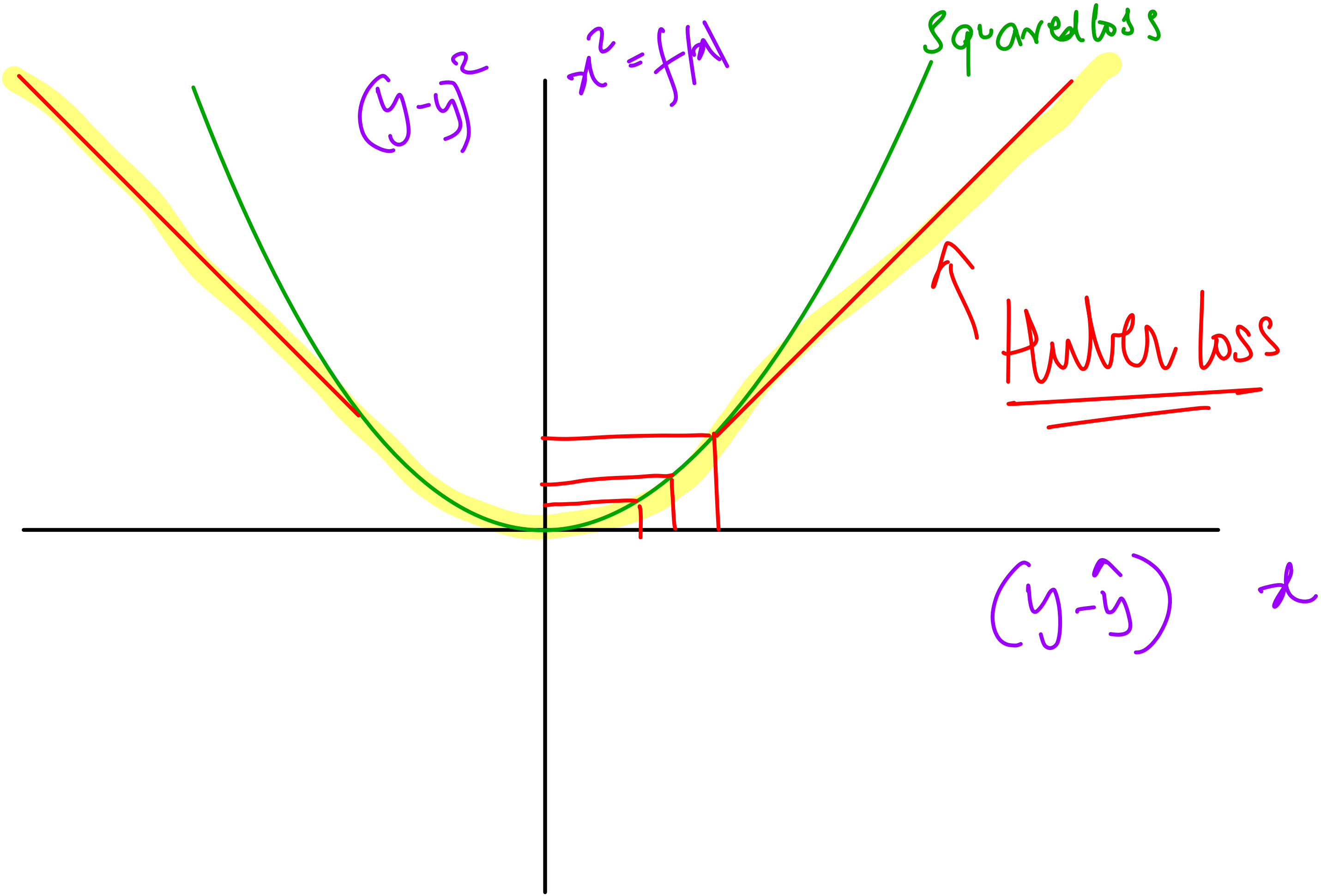
Impact of Outlier → GBDT

Cremes

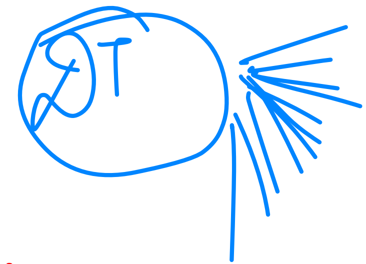
y	$\hat{y}$	error
20	48	-28
4	48	-44
6	48	-42
10	48	-38
200	48	152

Outlier Impact ↑↑↑↑↑

model will focus on reducing errors associated with outliers.

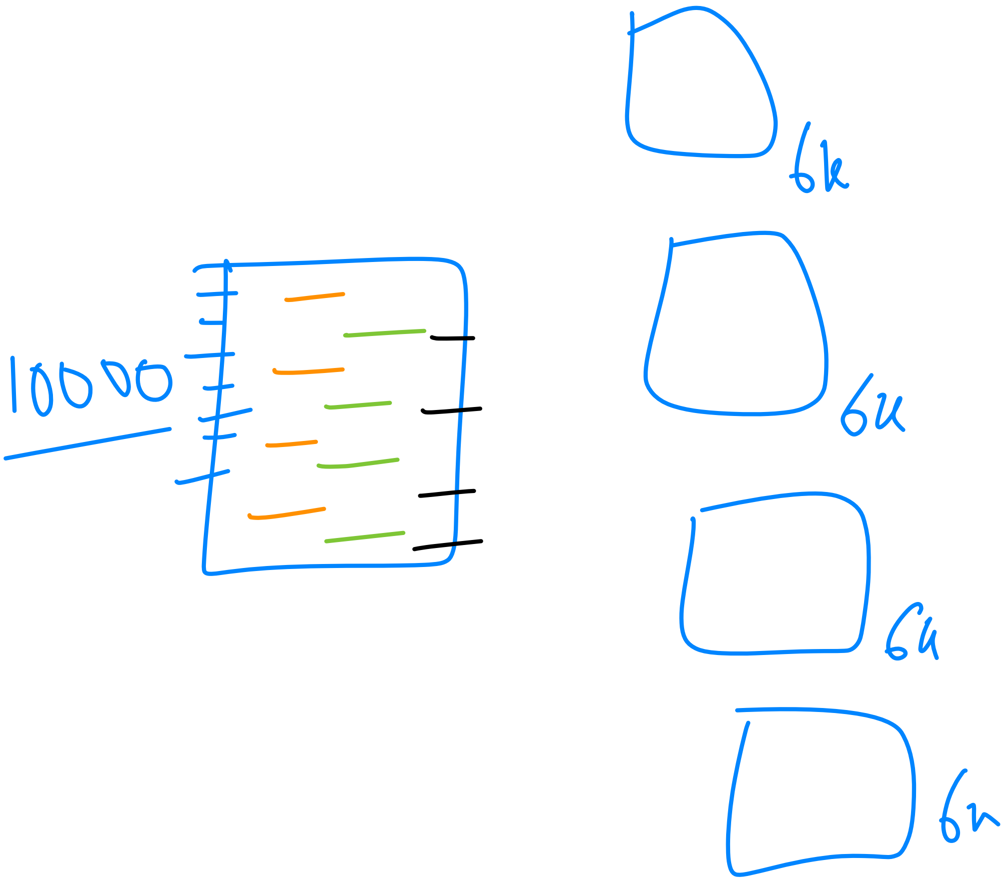

(y-y)²
x² = f(x)
Squared loss
Huber loss
(y-ŷ)
x

GBDT → slow



XGBoost → Not slow

LIGHT GBM → Not slow


10000
6k