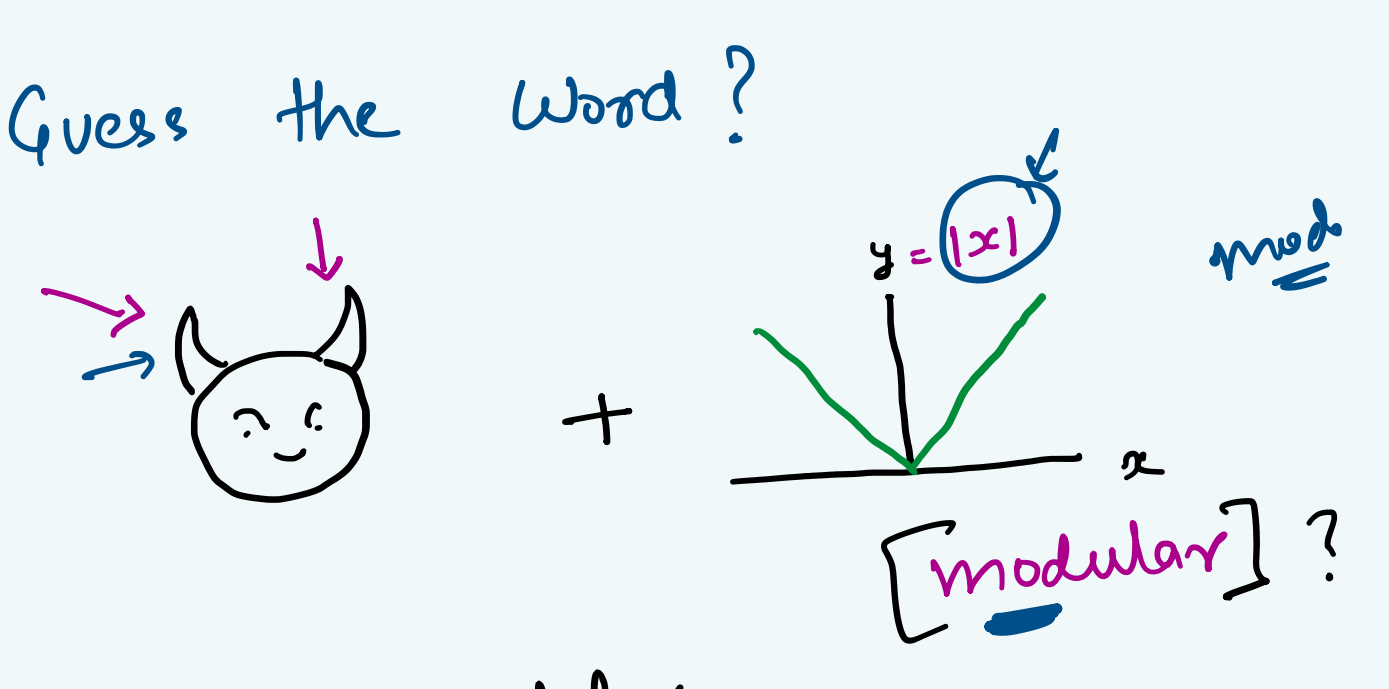
GBDT-XGBoost

Friday, 9 September 2022 8:11 PM

Q→ Guess the Word ?



de modular

⇒ Sigma ?

"Sigmoid"

Agenda :

1. GBDT
2. Bias Variance
3. Sklearn
4. XG Boost
5. Use-Case (EMG) "Robotics"

Residuals, Pseudo Residual, Loss function

trained model at end of K stage: $F_K(x) \Rightarrow h_0(x) + \alpha_1 \cdot h_1(x) + \dots + \alpha_K h_K(x)$

residuals ← $errors^{(i)} = y^{(i)} - F_K(x^i)$

for training of Model (K+1) using $\{x^{(i)}, errors^{(i)}\}$ → form Previous model.

Loss function.

$x^{(i)}$ → $y^{(i)}$, → $F_K(x^i) \simeq \hat{y}^{(i)}$ ⇒ $(y^{(i)} - F_K(x^i))^2$ ← squared loss

$$L(y^{(i)}, F_K(x^i)) = (y^{(i)} - F_K(x^i))^2$$

Let $z^{(i)} = F_K(x^i)$

$$\frac{\partial L}{\partial z^{(i)}} \Rightarrow \frac{d(y^{(i)} - z^{(i)})^2}{dz^{(i)}}$$

$$\frac{\partial L}{\partial z^{(i)}} \Rightarrow -2(y^{(i)} - z^{(i)})$$

$\frac{d(5-x)^2}{dx}$ ⇒ $\frac{d(5-x)^2}{d(5-x)} \cdot \frac{d(5-x)}{dx}$

$2(5-x) \cdot (-1)$

⇒ $-2(5-x)$

$$-\frac{\partial L}{\partial F_K(x^i)} = +2(y^{(i)} - F_K(x^i))$$ residual.

neg. derivative or neg. gradient

neg. derivative / neg. gradients ≈ (residuals) approx.

Pseudo Residual

"Huber loss" → $y^{(i)} - f_K(x^i)$ ← MSE; 0|1, 0.9

At j^{th} Model / iteration

$h_j(x)$ ← $\{x^{(i)}, err^{(i)}\}$ → $err^{(i)} = y^{(i)} - F_{j-1}(x)$

dy get-gradient( )

$\left[\lim_{h \to 0} \frac{f(x+h) - f(x)}{h}\right]$

automated { $error^{(i)}$ → residuals. $error^{(i)} \approx$ pseudo residual$^{(i)}$ $\approx -\frac{\partial L}{\partial f_{j-1}(x)}$

Rf → minize "hingloss" ✗

GBDT → minize "hinge loss" ✓

Stage p :

$F_p(x) \Rightarrow F_{p-1}(x) + \alpha_p \cdot h_p(x)$

$h_p(x)$ ← trained on $\{x^{(i)}, err^{(i)}\}$ → Not ~~$y^{(i)} - f_{p-1}(x^i)$~~ → $-\frac{\partial L}{\partial F_{p-1}(x^i)}$ Pseudo Residual.

The Algorithm

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

Algorithm:

1. Initialize model with a constant value: (→ avg.)

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$ → avg := $h_0(x) = F_0(x)$; → log. loss; $P(y=1)$; 80%

2. For m = 1 to M: (m=2)
   1. Compute so-called *pseudo-residuals*:
   $$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \dots, n.$$
   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:
   $$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)).$$ ($F_0(x) + \gamma_1 h_1(x)$) = 0
   4. Update the model:
   $$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$ (new tree) $f_1(x) = F_0(x) + \gamma_1 h_1(x)$
3. Output $F_M(x)$.

$F_M(x) \Rightarrow h_0(x) + \alpha_1 \cdot h_1(x) + \dots + \alpha_M h_M(x)$ (γ_1, γ_2, γ_M)

α's → γ's

$F_1(x) =$

$F_2(x) =$

$F_3(x) =$

$F_M(x) =$

$f(x) = (x-5)^2$

At $x = ?$

$\arg\min_x (x-5)^2$

As Base learners → DT shallow depth $d = 2, 3, 4, 5$

Bias Variance.

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

How to know Best M ?

→ Grid Search → Hyper param tuning on 'M'

V.Imp. [as M↑ → overfit ↑ ; M↓ → overfit ↓]

Regularization by shrinkage.

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_m \cdot h_m(x)$$ $0 < \nu \leq 1$

ν ↓ learning rate

1. if $\nu = 0.0001$ (small)
   as ν ↓ → underfit. overfit ↓
2. if $\nu = 1$ (V. large)
   as ν ↑ → overfit ↑ (high variance)

good practice : $\nu = 0.1$

GBDT → M, → ν } → "grid Search"

Proper "Hyper-Param tuning" is required.

⇒ easy to overfit.

Time & Space Complexity.

Train time Complexity.

1 DT → $O(n \log n \cdot d)$

M DT → $O(n \log n \cdot d \cdot M)$

Rf was trivally parallizable algo.

GBDT is not easily parallizable.

"Sequentially"

Runtime Complexity. :

1 DT → O(depth)

M DT → O(M * depth) — 300; 2-4-5-6

Space Complexity :

$O(M \cdot DT's + M \cdot \gamma's)$ → M gammas.

Very Powerful

Tomorrow — RF, DT : Doubts, Outliers / Missing Values. PSP ↑

[40 mins]

[9pm]?

x	y
	2
	5
	6
	4

$x^{(i)}$

$\bar{y}$ avg = 8.4

$x^{(i)} \rightarrow y^{(i)} \approx 12$

12

x

$F_M(x) \Rightarrow h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x)$

$h_0(x)$ → 8.4 ; 1.4 ; 2.2

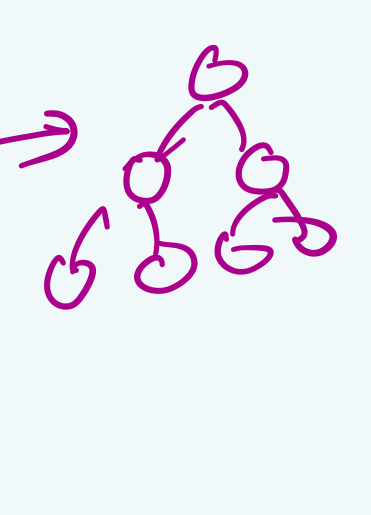
3.6 − 9.8 ; 1.08 → res = 0.12

γ → 0.1 | 0.2

8.4 → 11 ; → 8.81 ; → 8.62 → 8.75 → 9.01 ... 12 ; 30 ...

3.8 → 0.38 ; 0.74

$F_{m-1}(x)$ ← $F_M(x)$; $h_m(x)$ →



$\{x^{(i)}, err^{(i)}\}$ → $-\frac{\partial L}{\partial f_{m-1}(x)}$

γ_M

$$F_M(x) = F_{m-1}(x) + \gamma_M \cdot h_m(x)$$