

Session - 7

BOSTING - 2

Sep 08, 2025



AGENDA

- 1 Questions
- 2 GBDT- Pseudo Code
- 3 EXPLANATION

When they put you on a project where the previous developer has not commented a single line



GBDT - ALGORITHM

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

Algorithm:

1. Initialize model with a constant value:

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

2. For $m = 1$ to M :

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

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

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:

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

4. Update the model:

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

3. Output $F_M(x)$.

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

STEP-1

1. Initialize model with a constant value:

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

$$L = \sum_{i=1}^n (y_i - \gamma)^2$$

$$\underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \gamma)^2$$

derivative = 0

$$\sum_{i=1}^n 2(y_i - \bar{y})(-1) = 0$$

$$\sum_{i=1}^n (y_i - \bar{y}) = 0$$

$$(y_1 - \bar{y}) + (y_2 - \bar{y}) + (y_3 - \bar{y}) + \dots + (y_n - \bar{y}) = 0$$

$$\therefore \sum_{i=1}^n y_i - n \times \bar{y} = 0$$

$$\Rightarrow \bar{y} = \sum_{i=1}^n y_i \times \frac{1}{n}$$

$$\text{Asgm} \underset{n}{\left(5 - n \right)^2} ; \quad n=5 \\ = 0$$

2. For $m = 1$ to M :

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

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

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 $\{(x_i, r_{im})\}_{i=1}^n$.

Fit a base Learner on
 r_{im} of $F_{m-1}(x)$

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

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

For each base learner, we will have different level of contribution, so model will assign diff weight parameter for each base-learner.

| n | y | $\pi_0(x)$ | ϵ_0 | (π_1, ϵ_1) | $\hat{y}_i(x_i)$ |
|-----|-----|------------|--------------|-----------------------|-------------------------|
| 1 | 3 | 0 | 3 | 2.5 | $0 + 2.5 \times 2.5$ |
| 2 | 0 | 0 | 0 | 0.1 | $0 + 2.5 \times 0.1$ |
| 3 | -3 | 6 | -3 | -2.5 | $0 + 2.5 \times (-2.5)$ |

$\hat{y} = F_1(m)$

$$L = \frac{1}{3} \sum_{i=1}^3 (y_i - \hat{y})^2$$

$$= \frac{1}{3} \left[(3 - 2.5 \times 2.5)^2 + (0 - 2.5 \times 0.1)^2 + (-3 - 2.5 \times -2.5)^2 \right]$$

↓ ↓

$$= \frac{1}{3} \left[(3 - 2.5 \times 2.5)^2 + (0 - 2.5 \times 0.1)^2 + (-3 + 2.5 \times -2.5)^2 \right]$$

$$(a-b)^2 = a^2 + b^2 - 2ab$$

$$A = 9 + 6.25 \gamma^2 - 15\gamma$$

$$C = 9 + 6.25 \gamma^2 - 15\gamma$$

$$B = 0.01 \gamma^2$$

$$A + B + C = 12.51 \gamma^2 - 30\gamma + 18$$

$$\therefore 2 \times 12.51 \gamma - 30 = 0$$

$$\therefore \gamma = \frac{30}{2 \times 12.51} = 1.199$$

4. Update the model:

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

$$F_3(n) = h_0(n) + h_1(n) + h_2(n) + h_3(n)$$

$$= F_2(n) + h_3(n) \times \gamma_3$$

After learning $h_1(x)$ - We need to find γ

STEP 2 . 3

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

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



In simple terms, this equation is saying ;

- To find γ which minimizes given loss

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

\downarrow \downarrow
Actual Value \hat{y} (Predicted Value)

$$\hat{y} = F_{m-1}(x^i) + \gamma h_m(x^i)$$

To find

$$\text{For } m = 1, \quad \hat{y} = F_{1-1}(x^i) + \gamma_1 h_1(x^i)$$

$$= F_0(x^i) + \gamma_1 h_1(x^i)$$

Stage 0 model (mean model) Calculated is step 2.2



Only variable here is γ

(Rest everything is constant -
already calculated)

To find γ which minimizes given loss

- Take derivative of loss w.r.t γ & equate it to 0

STEP 2 . 4

After finding $h_1(x)$ and γ_1

- We need to make final prediction

(i.e combine previous model with current model predictions)

4. Update the model:
 $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$

For $m = 1$

$$= F_0(x) + \gamma_1 h_1(x)$$

calculated in Step - 1 calculated in Step - 2.3 calculated in Step - 2.2

We keep doing sub-steps for M iteration i.e finding

$$\begin{bmatrix} h_1(x), h_2(x), h_3(x) \dots h_m(x) \\ \gamma_1, \gamma_2, \gamma_3 \dots \gamma_m \end{bmatrix}$$

Finally, we use these to make final model $F_m(x)$

3. Output $F_M(x).$

$$F_m(x) = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x) + \dots + \gamma_m h_m(x)$$

Which of the following statement(s) are True?

0 users have participated

- A F_0(x) model will be a mean model 0%
- B For additive combining, we do weighted addition of consecutive models instead of simple addition 0%
- C We perform weighted addition in order to control the influence of the base learners. 0%
- D All of the above 0%

[End Quiz Now](#)



As we increase the value of M (number of base learners), the model will:

0 users have participated

- A Underfit 0%
- B Overfit 0%
- C No change 0%
- D None of the above 0%

[End Quiz Now](#)

Leaderboard
Based on all quizzes from this session



Hyperparameter : # of base learners (M)

$M \uparrow \longrightarrow$ Overfit

As M increases , model will overfit .

WHY ?

- Because as base learners increase more likely training error will tend to 0

$M \downarrow \longrightarrow$ Underfit

As M decreases , model will underfit.

WHY ?

- Say M = 1 , Stage 0 & Stage 1 model.
- Prediction will be close to mean model - Underfit

Hyperparameter

Depth

- As depth increases
- Model will overfit

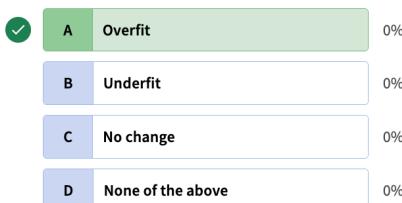
Why ?

- Increase in depth → Variance Increase
- Model will overfit quickly



What will happen if we increase max_depth of base learner in boosting?

0 users have participated



[End Quiz Now](#)

| Leaderboard | | | |
|---------------------------------------|-------------------------------|--------|--------|
| Based on all quizzes from the session | | | |
| SA | Souvik Adhikary 3/3 | 276.27 | 276.27 |
| I | Shreya Gupta 3/3 | 276.95 | 276.95 |
| S | Sumanth Andhav... 3/3 | 274.01 | 274.01 |
| 4 | Kiran Hebasur 3/3 | 270.75 | 270.75 |
| 5 | tejas sinha 3/3 | 267.62 | 267.62 |
| 6 | Perisetta Pavan Kalyan 3/3 | 266.37 | 266.37 |
| 7 | Purushottam Kumar 3/3 | 265.51 | 265.51 |
| 8 | Nachiket Pawar 3/3 | 264.54 | 264.54 |
| 9 | Mohanankrishna 3/3 | 260.84 | 260.84 |
| 10 | SHASHANK JHA 3/3 | 258.68 | 258.68 |

Regularization by Shrinkage

$$F_n(x) = h_0(x) + \eta \times \gamma_1 \times h_1(x) + \eta \times \gamma_2 \times h_2(x) + \eta \times \gamma_3 \times h_3(x) + \eta \times \gamma_4 \times h_4(x) + \dots + \eta \times \gamma_n \times h_n(x)$$

weights for each base learner

$$\eta = 10^{-2}$$

$$m_0, y = 12$$

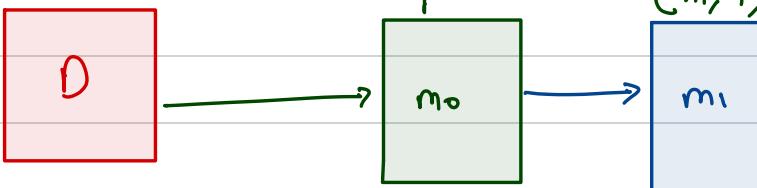
$$\epsilon_0 = 4$$

$$\hat{y}_0 = 8$$

$$\epsilon_1 = 3.98$$

$$\hat{y}_1 = 2 \times 0.01$$

$$(m_1, y)$$



$$\epsilon_2 = 0.2$$

$$\hat{y}_2 = 0.8$$



Regularization by Shrinkage

Final Model equation is: $F_m(x) = h_0(x) + \sum_{m=1}^M \gamma_m \cdot h_m(x)$

To regularize, we add an regularization term i.e learning rate

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

Learning Rate

$$Range : 0 \leq \nu \leq 1$$

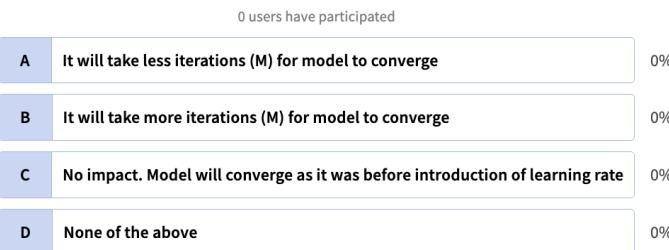
Notes :

- Adding learning rate is reducing the impact of Mth .

Hence, reducing overfit.

This learning-rate was applied post-hoc. Need to check in sklearn.

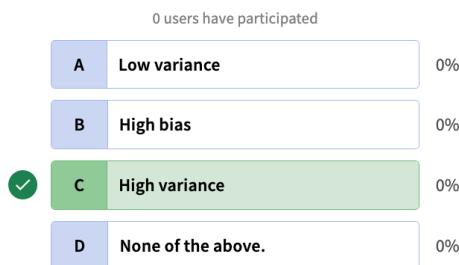
What happens if we have learning rate as a very small value ?



Based on all quizzes from the session

| | | | | |
|--|---------------------|-----|--------|--------|
| | Souvik Adhikary | 4/4 | 370.31 | 360.84 |
| | Shoreya gupta | 4/4 | 372.55 | 357.09 |
| | Sumanth Andhavar... | 4/4 | 361.70 | 354.59 |
| | Kiran Hebasur | 4/4 | 354.24 | 353.24 |
| | tejas sinha | 4/4 | 352.88 | 352.05 |
| | Nachiket Pawar | 4/4 | 352.05 | 352.05 |
| | Purushottam Kumar | 4/4 | 345.24 | 345.24 |
| | Mohanan krishna | 4/4 | 345.09 | 345.09 |
| | SHASHANK JHA | 4/4 | 345.09 | 345.09 |
| | MADEEHA REHMAN | 4/4 | 345.09 | 345.09 |

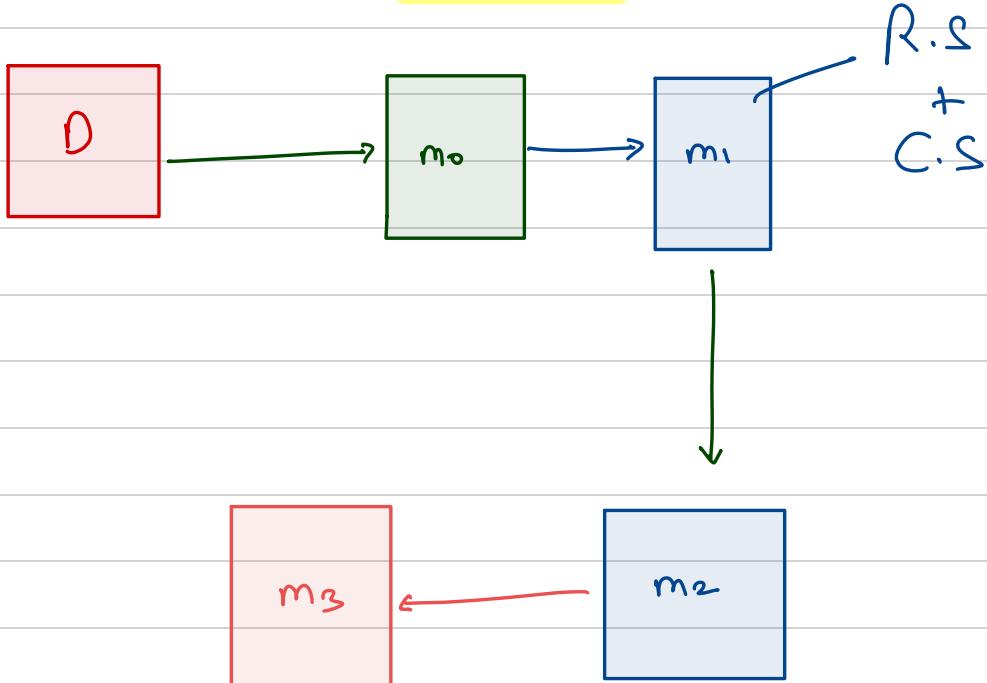
What does it mean when model is overfit ? Model will have ___



Based on all quizzes from the session

| | | | | |
|--|---------------------|-----|--------|--------|
| | Souvik Adhikary | 5/5 | 459.84 | 443.65 |
| | Shoreya gupta | 5/5 | 466.88 | 443.45 |
| | tejas sinha | 5/5 | 447.08 | 440.51 |
| | Nachiket Pawar | 5/5 | 443.45 | 443.45 |
| | SHASHANK JHA | 5/5 | 443.45 | 440.51 |
| | Kiran Hebasur | 5/5 | 440.51 | 440.51 |
| | Sumanth Andhavar... | 5/5 | 435.90 | 440.51 |
| | MADEEHA REHMAN | 5/5 | 421.16 | 421.16 |
| | Rakesh Karade | 5/5 | 408.76 | 408.76 |
| | Sri Harsha Nanduri | 5/5 | 397.11 | 397.11 |

Stochastic GBDT



| | f_1 | f_2 | f_3 | f_4 | f_5 | f_6 | y |
|-------|-------|-------|-------|-------|-------|-------|-----|
| n_1 | | | | | | | |
| n_2 | | | | | | | |
| n_3 | | | | | | | |
| n_4 | | | | | | | |
| n_5 | | | | | | | |
| n_6 | | | | | | | |
| n_7 | | | | | | | |
| n_8 | | | | | | | |

$m_2 \rightarrow n_2, n_3 ; f_2, f_5$

$m_3 \rightarrow n_7, n_8 ; f_1, f_6$

This is HW.

Stochastic Gradient Boosting



PROBLEM : GBDT overfit a lot

|
(High Variance)

How to reduce variance ? - Randomization

Row sampling +
column sampling

- Can use the same concept of randomization (as used in RF) to reduce variance

- This variation of GBDT is called ' Stochastic Gradient Boosting'

GBDT → Pseudo Residual + Additive Combing

Stochastic GDBT → GBDT + Randomization

|
Row Sampling + Column Sampling

- skLearn provides ability of row sampling
 - Using **subsample** hyperparameter
 - And column sampling using **max_features** hyperparameter

Time Left: 26s

Which of the following techniques can help in regularizing GBDT

0 users have participated

A Using Learning rate

0%

B Using Sampling (Stochastic GBDT)

0%

C All of the above

0%

End Quiz Now



Souvik Adhikary
6/6 ⚡ 551.77



Shreyas Gupta
6/6 ⚡ 557.28



tejas sinha
6/6 ⚡ 537.55

| | | |
|----|---------------------|--------------|
| 4 | Nachiket Pawar | 6/6 ⚡ 535.05 |
| 5 | Kiran Hebasur | 6/6 ⚡ 529.71 |
| 6 | Sumanth Andhavarapu | 6/6 ⚡ 526.63 |
| 7 | MADEEHA REHMAN | 6/6 ⚡ 497.02 |
| 8 | Rakesh Karade | 6/6 ⚡ 480.46 |
| 9 | Sri Harsha Nanduri | 6/6 ⚡ 477.51 |
| 10 | SHASHANK JHA | 5/6 ⚡ 443.45 |

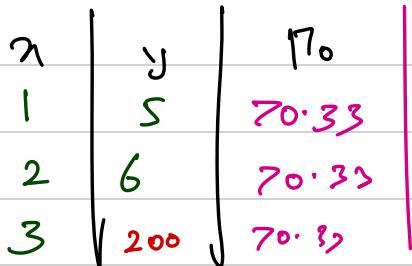
What do you think: Do outliers impact GBDT?

11 users have participated



[End Quiz Now](#)

| | | |
|---------|---------------------------------|------------------------------|
| S 2 | Sumanth Andhav... 7/7 610.16 | Shoreya gupta 7/7 557.28 |
| SA 1 | Souvik Adhikary 7/7 640.54 | Rakesh Karade 7/7 555.27 |
| MR 3 | MADEEHA REHMAN 7/7 579.32 | SHASHANK JHA 6/7 540.12 |
| | | tejas sinha 6/7 537.55 |
| | | Mohankrishna 6/7 537.28 |
| | | Nachiket Pawar 6/7 535.05 |
| | | Kiran Hebasur 6/7 529.71 |



 ATTENTION PLEASE!



Does outlier impact GBDT ?

- As each model is fit on residual of previous model
- Outliers will have **high residual**

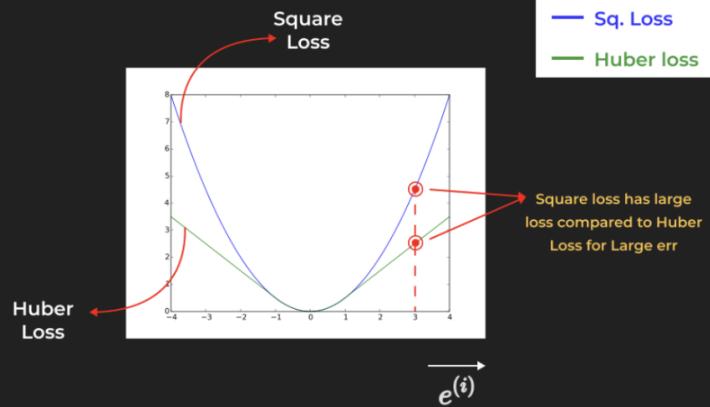
This causes GBDT to focus its attention on reducing these residual for outlier points.

Instead of using squared loss ,

- We can use **huber loss**

Notice :

- For smaller residual values.
- Both squared loss & hyper loss have same values
- As value of residual increases Huber loss doesn't explode like squared loss



How can we deal with outliers in GBDT? Statement 1: Remove the outliers

Statement 2: Use loss function which isn't impacted by outliers

0 users have participated

- A Statement 1 0%
- B Statement 2 0%
- C Statement 1 & 2 0%
- D None of the above 0%

[End Quiz Now](#)

| User Leaderboard | | |
|------------------|---------------------------|------------|
| MR | MADEEHA REHMAN | 8/8 653.14 |
| SA | Souvik Adhikary | 8/8 716.32 |
| M | Mohanakrishna | 7/8 625.46 |
| 4 | PP Perisetta Pavan Kalyan | 7/8 612.54 |
| 5 | S Sumanth Andhavarapu | 7/8 610.16 |
| 6 | N Nachiket Pawar | 7/8 609.01 |
| 7 | Shreya gupta | 6/8 557.28 |
| 8 | RK Rakesh Karade | 7/8 555.27 |
| 9 | T Tanvi Singh | 7/8 542.02 |
| 10 | SJ SHASHANK JHA | 6/8 540.12 |

MSE vs MAE

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

$$|y_i - \hat{y}|$$

$$2 \rightarrow 4$$

$$4 \rightarrow 16$$

$$(0.01 - 0.03)$$

$\frac{1}{2}$ with fraction \rightarrow or MAE