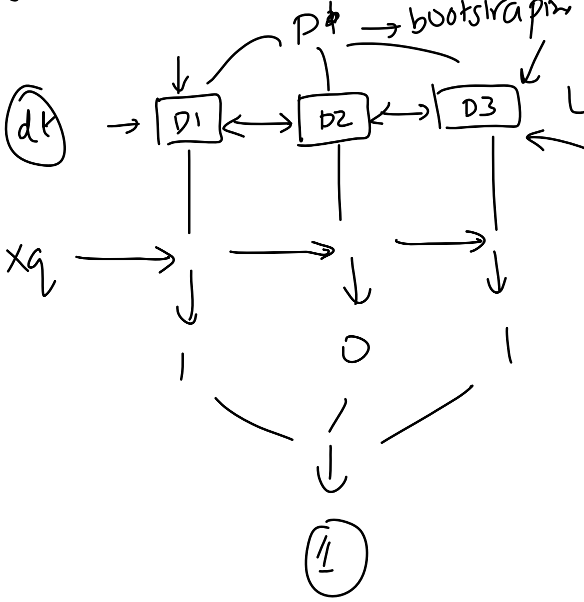
Boosting

02 August 2023 15:55

high bias

Boosting is a general ensemble method in machine learning that aims to create a strong classifier or regressor by combining the predictions of several weaker models. The idea is to build the strong model incrementally, by sequentially adding weak models that are trained to correct the mistakes made by the existing ensemble.

- Bagging
  - Random forest
- Boosting
  - Gradient Boost
  - Xgboost



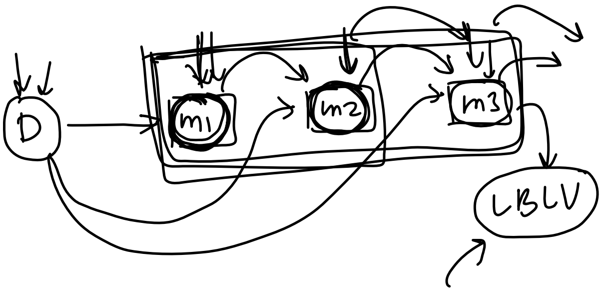
LBHV (deep decision tree)

variance reduce

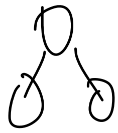
↓

wisdom of the crowd

LBHV



1) sequentially train
2) all the model ↓ same data
3) HBLV model
   - shallow dt
   - linear model



powerful → Boosting
- Adaboost
- Gradient Boost
- → xgboost

{ lightgbm catboost }

industry ⟷ competitive Kaggle → Xgboost

What is Gradient Boosting

02 August 2023 07:38

Gradient Boosting is a machine learning ensemble technique that aims to build a strong predictive model by combining the predictions of several weaker models using the concept of Additive Modelling, typically decision trees. The method works by iteratively adding models to the ensemble, with each new model trained to correct the mistakes made by the combined ensemble of existing models.

Gradient Boosting is a powerful and flexible method that can be used for both regression and classification tasks. → ranking systems

It is particularly effective when the data has complex, non-linear relationships, and it often performs well even with little hyperparameter tuning.

Its popularity in real-world applications and machine learning competitions is testament to its effectiveness, and it has implementations in most major machine learning libraries, such as scikit-learn, XGBoost, LightGBM, and CatBoost.

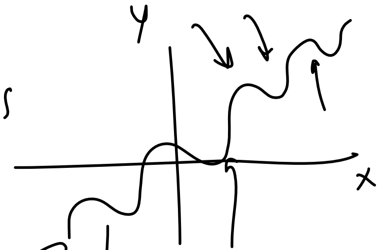
m1 → m2 → m3

$f_1(x)$ $f_2(x)$ $f_3(x)$

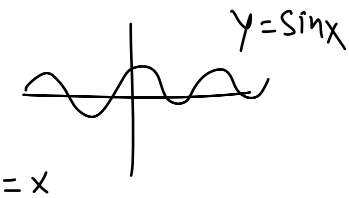
Additive modelling

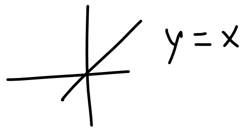
$f_1(x) + f_2(x) + f_3(x)$

Additive modelling



eq





$y = \underbrace{x + \sin x}$

$y = x \quad y = \sin x$

Additive modelling

X | Y → $y = \boxed{f(x)}$

$f_1(x)$ $f_2(x)$ $f_3(x)$

$f(x) = \underbrace{f_1(x) + f_2(x) + \ldots + f_n(x)}$

The How?

02 August 2023 16:20

Wikipedia m dt

$L(y,\hat{y}) = \frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ $\gamma = f_0(x)$

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

→ 1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.

2. For $m = 1$ to M: → 10 times

(a) For $i = 1, 2, \ldots, N$ compute → rows

$i = 1, 2, 3$

Pseudo residual → $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$.

(b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$. $j = 1, 2, 3$ $m = 1$

(c) For $j = 1, 2, \ldots, J_m$ compute

$\gamma_{11}\ \gamma_{12}\ \gamma_{13}$

$$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

3. Output $\hat{f}(x) = f_M(x)$.

$m = 1$

regression

$\boxed{\gamma = f_0(x) = Y_{mean}}$

γ_{i1} → γ_{11}, γ_{21}, γ_{31}

dt

mean → m1

mo deviation

mo m=1 m=2

$\{(x_i, y_i)\}_{i=1}^{n}$ →

$\{x_i, y_i\}_{i=1}^{3}$

$\boxed{L(y, f(x))}$

x | y

$y = f(x)$

$\boxed{L(y, \hat{y})}$ ←

$\hat{y} = f(x)$

actual output   prediction

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

squared loss $L = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$x > y$

$\boxed{\frac{x}{2} > \frac{y}{2}}$

$\frac{x}{2}$ $\frac{y}{2}$

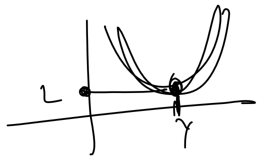
$$\boxed{L(y, \hat{y}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$L(y, \hat{y}) = \boxed{\frac{1}{2}} \sum_{i=1}^{n}(y_i - \hat{y}_i)$$

min

$$f_0(x) = \arg\min_\gamma \sum_{i=1}^{n} L(y, \gamma)$$

$$\boxed{\frac{\partial L}{\partial \gamma} = 0}$$



$$\frac{\partial f_0(x)}{\partial \gamma} = \frac{\partial}{\partial \gamma}\sum_{i=1}^{n}(y_i - \gamma)^2$$

$\gamma = Y_{mean}$

$$= -2\sum_{i=1}^{3}(y_i - \gamma) = 0$$

$$\sum_{i=1}^{3}(y_i - \gamma) = 0$$

Y_{mean}

$$y_1 - \gamma + y_2 - \gamma + y_3 - \gamma = 0$$

$$192 - \gamma + 144 - \gamma + 91 - \gamma = 0 \Rightarrow 3\gamma = 192 + 144 + 91$$

$$\gamma = \boxed{\frac{192 + 144 + 91}{3}}$$

$$f(x) = f_0(x) + f_1(x) + f_2(x) + \ldots + f_m(x)$$

$f_0(x)$ → Y_{mean}

10 dt

m1 m2 m3

m1 → m2 → m3 → □ □ □

mean of y col — decision tree — dt

m=1

$$\gamma_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

$$\gamma_{i1} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_0} = \left[\frac{\partial L(y_i, f_0(x_i))}{\partial f_0(x_i)}\right]$$

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

$$\gamma_{i1} = \frac{1}{2}\frac{\partial}{\partial f_0(x)} \sum_{i=1}^{n} (y_i - f_0(x_i))^2$$

$$= -\frac{2}{2}(y_i - f_0(x_i)) \Rightarrow \boxed{\gamma_{i1} = f_0(x_i) - y_i}$$

□ → □ → □

avg

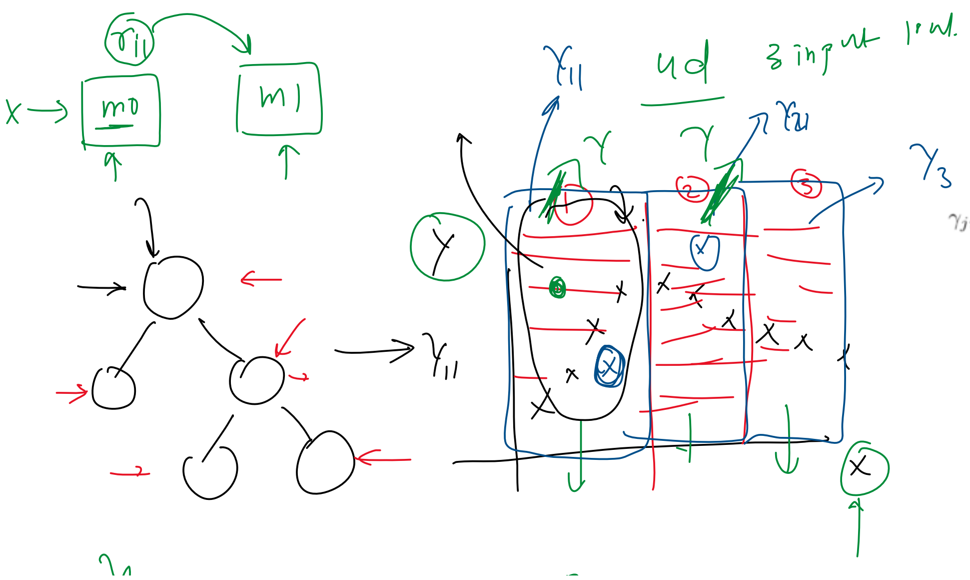
$$\gamma_{11} = f_0(x_1) - y_1 = 142 - 192$$

$$\gamma_{21} = f_0(x_2) - y_2 = 142 - 144$$

$$\gamma_{31} = f_0(x_3) - y_3 = 144 - 91$$

m0 → m1

avg — dt (regression)



$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

$m=1$ $\quad \gamma \quad \gamma_{21}$

$x_i \in R_{j1}$ → terminal region

x

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$

$$\boxed{\gamma_{j1}} = \arg\min_{\gamma} \sum_{\boxed{x_i \in R_{j1}}} L\left(y_i, f_0(x_i) + \gamma\right)$$

γ_{11} $\quad j=1$

$$\boxed{\gamma_{11}} = \arg\min_{\gamma} \underbrace{L\left(y_3, f_0(x_3) + \gamma\right)}$$

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

$$L = (y_3 - f_0(x_3) + \gamma)^2$$

$$\gamma_{11} = \arg\min_{\gamma} \underbrace{(y_3 - f_0(x_3) - \gamma)^2}$$

$$= \frac{\partial}{\partial \gamma} (y_3 - f_0(x_3) - \gamma)^2$$

-1

$$\Rightarrow -2(y_3 - f_0(x_3) - \gamma) = 0$$

$$y_3 - f_0(x_3) - \gamma = 0$$

$$\gamma = y_3 - f_0(x_3)$$

$$= 91 - 142$$

$$\boxed{\gamma = -51}$$

$$\gamma_{11} = -51$$

γ_{21}

γ_{31}

$m=1$

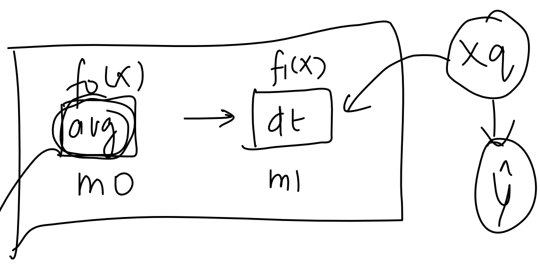
m1 output

$$f_m(x) = f_{m-1}(x) + \left[\sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})\right].$$

$$f_1(x) = \boxed{f_0(x)} + \underbrace{\text{output of decision tree}}$$

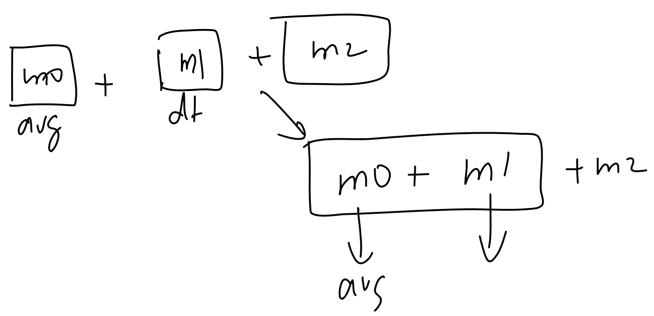
$$142 + 49 \qquad 142 + 1.66 = 143.6$$

191



avg + output

$$\hat{y} = f_1(x)$$



$f_m(x) = f_{m-1}(x) + \underbrace{\sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})}.$

$f_5(x) = f_4(x) + \underbrace{dt5}$

$\swarrow$

$f_4(x) = f_3(x) + dt4$

$\swarrow$

$f_2(x) + dt3$

$\swarrow$

$f_1(x) + dt2$

$\swarrow$

$\boxed{f_0(x) + dt1}$

$f_5(x) = f_0(x) + f_1(x) + f_2(x) + \ldots$

$f_5(x) +$

$= f_0(x) + dt1 + dt2 + dt3 \ldots + dt4$

The What?

02 August 2023 16:20

	R&D	Ops	Marketing	Profit
0	165	137	472	192
1	101	92	250	144
2	29	127	201	91

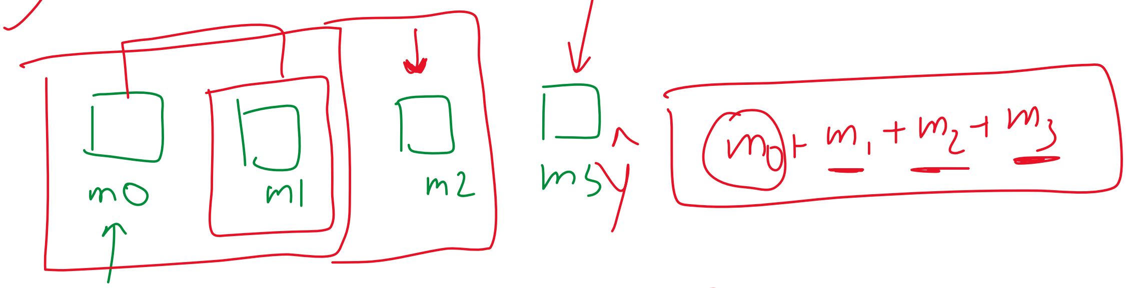
$f_0(x)$ | r_1 | $f_1(x)$ | $y - f_1(x)$ | $f_2(x)$ | $y - f_2(x)$ | $f_3(x)$

142

142

142

gradient boosting regression



$m_0 + m_1$

$m_0 + m_1 + m_2$

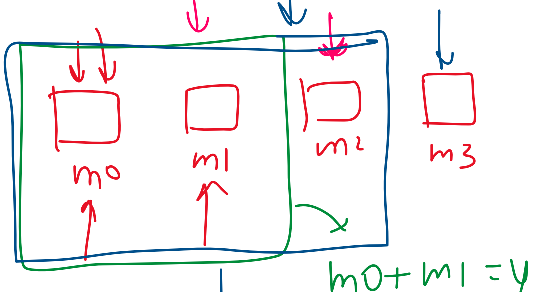
X y

	R&D	Ops	Marketing	Profit
0	165	137	472	192
1	101	92	250	144
2	29	127	201	91

$f_0(x)$ | $y - f_0(x)$ | $f_1(x)$ | $y - f_1(x)$ | $f_2(x)$ | $y - f_2(x)$ | $f_3(x)$

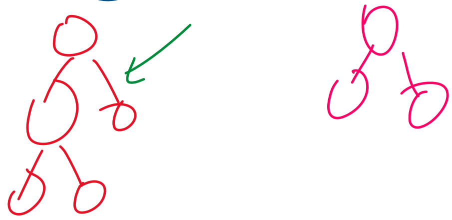
X r_2

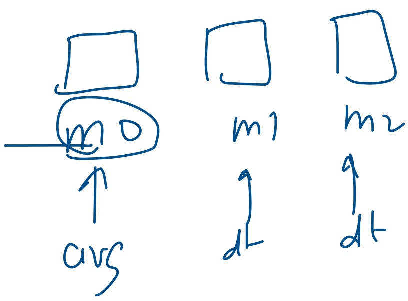
$f_3(x) = m_0 + m_1 + m_2 + m_3$



$m_0 + m_1 = y$

$m_0 + m_1 + m_2$





The Why?

02 August 2023 16:20

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute

   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$

   (b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm},\ j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute

   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$

   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.