Градиентный бустинг в XGBoost

$\mathcal{F}$ – мн-во базовых моделей

Строим композицию вида:

$$\hat{y}_t(x) = \sum_{r=1}^{t} \gamma_r b_r(x) = \hat{y}_{t-1}(x) + \gamma_t b_t(x).$$

Пусть $L(y, z)$ – ф-ция потерь

Функционал качества: $Q(y, \hat{y}) = \sum_{i=1}^{n} L(y_i, \hat{y}(x_i))$

где $(x_1, y_1), \dots, (x_n, y_n)$ – обуч. выборка.

Работа не волк.
Волк - это ходить,
работа это ворк :)

(не стирать!!!)

Обид нет.
Просто сделаны
выводы.

Перейдем в пр-во $\mathbb{R}^n$.

Таблица соотв. обозначений:

Объекты в $\mathbb{R}^n$	Их значения в бустинге.
$S \in \mathbb{R}^n$	$\left(\hat{y}_{t-1}(x_i)\right)_{i=1}^{n}$
$\Delta S \in \mathbb{R}^n$	$\left(b_t(x_i)\right)_{i=1}^{n}$
$Q(y,S) = \sum_{i=1}^{n} L(y_i, s_i)$	$Q(y, \hat{y})$
$g = \nabla_s Q(y, s)$	$\tilde{g} = \nabla_s Q(y, \hat{y})$
$h_i = \frac{\partial^2 L(y_i, s_i)}{\partial s_i^2}$	$\tilde{h}_i = \frac{\partial^2 L(y_i, \hat{y}(x_i))}{\partial s_i^2}$

Рассм. разложение по Тейлору $Q(y, s+\Delta s)$ до 2-й произв.:

$$Q(y, s+\Delta s) = \sum_{i=1}^{n} L(y_i, s_i + \Delta s_i) =$$

$$= \sum_{i=1}^{n} \Big[\underbrace{L(y_i, s_i)}_{\text{не зависит от } \Delta s} + \Delta s_i \cdot g_i + \frac{1}{2} \Delta s_i^2 \underbrace{\frac{\partial^2 L(y_i, s_i)}{\partial s_i^2}}_{= h_i} \Big] \to \min_{\Delta s \in \mathbb{R}^n}$$

$$\sum_{i=1}^{n} \Big[\Delta s_i \cdot g_i + \frac{1}{2} \Delta s_i^2 h_i \Big] \to \min_{\Delta s \in \mathbb{R}^n}$$

Вспомним,
при постр

В терми

Вывод: в об

$\to \min\limits_{\Delta S \in \mathbb{R}^n}$

Вспомним, что мы оптимизировали в простран. град. буст. при построении новой базовой модели.

$$\sum_{i=1}^{n} \left(b(x_i) + \tilde{g}_i \right)^2 \to \min_{b \in \mathfrak{B}}.$$

в терминах пр-ва $\mathbb{R}^n$:

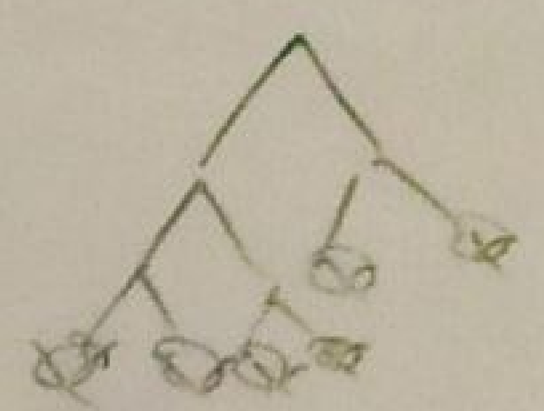
$$\sum_{i=1}^{n} \left(\Delta S_i + g_i \right)^2 \to \min_{\Delta S \in \mathbb{R}^n}.$$

$$= \sum_{i=1}^{n} \left(\Delta S_i^2 + 2 g_i \Delta S_i + \cancel{g_i^2} \right) \sim \sum_{i=1}^{n} \left(\Delta S_i g_i + \frac{1}{2} \Delta S_i^2 \right) \to \min_{\Delta S \in \mathbb{R}^n}$$

т.к. не зав. от ΔS

<u>Вывод</u>: в обычном бустинге мы отбрасываем информацию о 2-й произв., по сути считая функцию квадратичной.

Пусть структура дерева построена.
Выбор значение в листе.



Дерево представимо в виде:

$$b(x) = \sum_{k=1}^{\ell} \gamma_k I\{x \in R_k\},$$

где ℓ - кол-во листьев
γ_k - значение в листе
$R_k \subset \mathbb{R}^d$ - область, объекты которой попадают в один лист с номером k.

Работа не волк
Волк - это ходить
работа это ворк :)

не спорим.

Обид нет.
Просто сделали выводы.

XGBoost использует доп. регуляризацию:

1). на кол-во листьев

2). норму значений в листах

Получаем функционал:

$$\sum_{i=1}^{\ell}\left[\tilde{g}_i b(x_i) + \frac{1}{2}\tilde{h}_i b^2(x_i)\right] + \mu \ell + \frac{\lambda}{2}\sum_{k=1}^{\ell} r_k^2 =$$

$$= \sum_{k=1}^{\ell} \left[\gamma_k \underbrace{\left(\sum_{i=1}^{n} \tilde{g}_i \, I\{x_i \in R_k\}\right)}_{G_k} + \frac{1}{2} \gamma_k^2 \left(\underbrace{\sum_{i=1}^{n} \tilde{h}_i \, I\{x_i \in R_k\}}_{H_k} + \lambda\right) + \mu \right] =$$

$$= \sum_{k=1}^{\ell} \left[\gamma_k G_k + \frac{1}{2} \gamma_k^2 (H_k + \lambda) + \mu \right] \to \min_{\gamma_k}$$

Задача разбивается на подзадачи по отдельным листьям.

$$\gamma_k = \frac{-G_k}{H_k + \lambda}$$ — оптимальное значение в листе.

$\ldots) + \mu \Big] =$

чи

Критерий информативности.

Подставим оптим. γ_k в исходной ф-л ;

$$Q^* = \sum_{k=1}^{l} \left(\frac{-G_k^2}{H_k+\lambda} + \frac{1}{2} \frac{G_k^2}{(H_k+\lambda)^{2}} \cdot \cancel{(H_k+\lambda)} + \mu \right) =$$

$$= \sum_{k=1}^{l} \left[-\frac{1}{2} \frac{G_k^2}{H_k+\lambda} + \mu \right] = \sum_{k=1}^{l} H(R_k)$$

Критерий информ.

X_m

$X_l \quad X_r$

$$H(X_m) - H(X_l) - H(X_r) \to \min$$