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

модель обучения

$$g(\omega, x_i)$$

$$\hat{y}_i^0 = g(\omega_0, x_i)$$

1) $$s_i^t = L'_y(y, y_i)\Big|_{y = \hat{y}_i^{t-1}}$$

2) Обучение $(x_i, -s_i^t)$

$$\omega_t = \operatorname*{argmin}_{\omega \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \Big(g(\omega_t, x_i) \underbrace{-\,(-s_i^t)}_{+} \Big)^2$$

3) Обучение коэффициент

$$\hat{\gamma}_t = \operatorname*{argmin}_{\gamma \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} L\big(\hat{y}_i^{t-1} + \gamma g(\omega_t, x_i), y_i\big)$$

4) $$\hat{y}_i^t = \hat{y}_i^{t-1} + \eta \hat{\gamma}_t g(\omega_t, x_i)$$

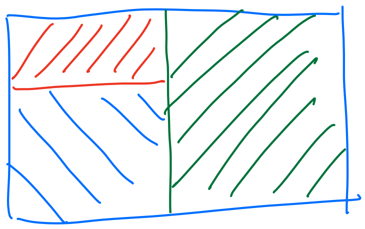
$$g^t_{\text{итоговая}}(\cdot) = g^{t-1}_{\text{итоговая}}(\cdot) + \eta \gamma g(\omega_t, \cdot)$$

Дерево:

$$\underset{\text{дерево}}{g}(w, x_i) = \sum_{k=1}^{L} q_k \, \mathbb{I}\{x_i \in R_k\}$$

w: $\{R_k\}_{k=1}^{L}$, $\{q_k\}_{k=1}^{L}$ — разбиение пр-ва

$\mathbb{I}\{x_i \in R_k\} = \{1 \text{ да}, 0 \text{ нет}\}$



Бустинг на деревьях:

$$g(w_t, x_i) = \sum_{k=1}^{L} q_k \, \mathbb{I}\{x_i \in R_k\}$$

$$\hat{y}_i^{t} = \hat{y}_i^{t-1} + \eta \gamma_t \, g(w_t, x_i)$$

$$\hat{y}_i^{t} = \hat{y}_i^{t-1} + \sum_{k=1}^{L} \underbrace{q_k \eta \gamma_t}_{\gamma_{tk}} \, \mathbb{I}\{x_i \in R_k\}$$

$$\gamma_{tk} = \underset{\gamma_k \text{ по всем } k}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L\left(\hat{y}_i^{t-1} + \sum_{k=1}^{L} \gamma_k \, \mathbb{I}\{x_i \in R_k\}; y_i\right)$$

$\Downarrow$ сепарабельно по γ_k

$$\gamma_{tk} = \underset{\gamma \in \mathbb{R}}{\arg\min} \sum_{i \in R_k} L(\hat{y}_i^{t-1} + \gamma; y_i)$$

Почему не применять исп. град. спуска?

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{L(\hat{y}_i; y_i)}$$

$$\approx L(\hat{y}_i; y_i) + \overbrace{L'_y(y; y_i)\Big|_{y=\hat{y}_i}}^{g_i} \cdot \Delta\hat{y}_i + \frac{1}{2}\underbrace{L''_y(y, y_i)\Big|_{y=\hat{y}_i}}_{h_i} \cdot (\Delta\hat{y}_i)^2$$

по Тейлору в окр $\hat{y}_i$

квадр — приближение

$$\min_{(\Delta\hat{y}_i)\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\Big[L(\hat{y}_i; y_i) + g_i \Delta\hat{y}_i + \frac{1}{2} h_i (\Delta y_i)^2\Big]$$

$$\Delta\hat{y}_i = \operatorname{argmin}_{\Delta y_i\in\mathbb{R}} \Big[\underbrace{g_i \Delta y_i + \frac{1}{2} h_i (\Delta y_i)^2}_{\text{произв.} = 0}\Big]$$

$$g_i + h_i \Delta y_i = 0$$

$$\Delta\hat{y}_i = -\frac{g_i}{h_i}$$

$$\boxed{\begin{array}{ll} \text{Раньше} & s_i^t = -g_i^t \\ \text{Теперь} & s_i^t = -\dfrac{g_i^t}{h_i} \end{array}}$$

учитываем кривизну

XG Boost:

две идеи дополнительно к h_i:

- штрафовать за кол-во листьев
- γ тк нужно регуляризовать

Итого: в базовом варианте

$$g(w_t, x_i) \to s_i = -\frac{g_i^t}{h_i}$$

с точки зрения оптимизации

$$w_t = \underset{w\in\mathbb{R}^d}{\arg\min}\left[\sum_{i=1}^{n}\left(g(w, x_i) - s_i\right)^2\right]$$

$$= \underset{w\in\mathbb{R}^d}{\arg\min}\left[\sum g^2(w, x_i) - 2s_i\, g(w, x_i) + \cancel{s_i^2}\right]$$

$$= \underset{w\in\mathbb{R}^d}{\arg\min}\left[\sum_{i=1}^{n} g^2(w, x_i) + \frac{2g_i^t}{h_i^t}\, g(w, x_i)\right]$$

$\frac{h_i^t}{2}$

$$= \underset{w\in\mathbb{R}^d}{\arg\min}\left[\sum_{i=1}^{n} g_i^t\, g(w, x_i) + \frac{1}{2} h_i^t\, g^2(w, x_i)\right]$$

с регуляризаторами:

$$w_t = \underset{w\in\mathbb{R}^d}{\arg\min}\left[\sum_{i=1}^{n} g_i^t\, g(w, x_i) + \frac{1}{2} h_i^t\, g^2(w, x_i) + \mu L + \frac{\lambda}{2}\sum_{k=1}^{L}\gamma_k^2\right]$$

здесь $g(\omega, x_i)$ задает дерево

$$g(\omega, x_i) = \sum_{k=1}^{L} \gamma_k \, \mathbb{I}\{x_i \in R_k\}$$

<u>Как искать γ_k?</u> Пусть дерево R_k построено

$$\min_{\gamma_k \in \mathbb{R}} \Bigg[\sum_{i=1}^{n} \Bigg(g_i^t \sum_{k=1}^{L} \gamma_k \, \mathbb{I}\{x_i \in R_k\} + \frac{1}{2} h_i^t \Big(\sum_{k=1}^{L} \gamma_k \, \mathbb{I}\{x_i \in R_k\} \Big)^2 \Bigg) + \mu L + \frac{\lambda}{2} \sum_{k=1}^{L} \gamma_k^2 \Bigg]$$

Зафиксируем R_k и найдем γ_k для нее

$$\gamma_k = \operatorname*{argmin}_{\gamma \in \mathbb{R}} \underbrace{\sum_{i \in R_k} g_i^t \gamma + \sum_{i \in R_k} \gamma^2 \frac{h_i^t}{2} + \frac{\lambda}{2} \gamma^2}$$

Берем производную

$$\sum_{i \in R_k} g_i^t + \gamma \cdot \Big(\lambda + \sum_{i \in R_k} h_i^t \Big) = 0$$

$$\boxed{\gamma_k^* = - \frac{\sum_{i \in R_k} g_i^t}{\lambda + \sum_{i \in R_k} h_i^t}}$$

Теперь ищем R_k, исходя из того, что всегда найдём оптим. γ_k^*

$$\min_{\{R_k\}} \left[\sum_{i=1}^{n} \left(g_i^t \sum_{k=1}^{L} \gamma_k \, \mathbb{I}\{x_i \in R_k\} + \frac{1}{2} h_i^t \left(\sum_{k=1}^{L} \gamma_k \, \mathbb{I}\{x_i \in R_k\} \right)^2 \right) + \mu L + \frac{\lambda}{2} \sum_{k=1}^{L} \gamma_k^2 \right]$$

$$\boxed{\min_{\{R_k\}} \sum_{k=1}^{L} \left[\mu - \frac{1}{2} \frac{G_k}{H_k + \lambda} \right] = \sum_{k=1}^{L} Err(\text{всех } x_i \in R_k)}$$

$$G_k = \sum_{i \in R_k} g_i^t \qquad H_k = \sum_{i \in R_k} h_i^t$$

общая задача поиска дерева для XGBoost

Ищем $\mathbb{I}\{j \text{ признак объекта} < t\}$

всех $x_i \in R_k$

$x_i \in R_{k,l}$ $\qquad$ $x_i \in R_{k,r}$

Поиск дерева:

$$\min_{j,t} \; Err(\text{всех } x_i \in R_{k,l}) + Err(\text{всех } x_i \in R_{k,r}) - Err(\text{всех } x_i \in R_k)$$

итеративная процедура

LightGBM

что нового:

- давайте при построении R_k не все объекты, а согласно g_i^t:
  1) брать 1 объект с большим g_i^t
  2) много объектов с меньшим g_i^t

CatBoost

- категориальные признаки
- усовершенствовали выбор объектов случай