Ансамбль:

$$\underbrace{y_T}_{\text{ответ ансамбля}} = \overbrace{\sum_{t=1}^{T} g_t(x)}^{\text{суммирование}} \qquad g_t \text{ из семейства}$$

$g(w_t, x)$

Идея бустинга:

1) строим модель
2) оцениваем её ошибки
3) строим новую модель, которая наилучшим образом компенсирует ошибки первой
4) ...

Пример:

квадратичная ф. потерь в регрессии

- $w_1 = \arg\min\limits_{w \in \mathbb{R}^d} \left[\frac{1}{2} \sum\limits_{i=1}^{n} \left(g(w, x_i) - y_i \right)^2 \right]$

- ошибки модели w_1: $e_{1,i} = y_i - g(w_1, x_i)$

- $w_2 = \arg\min\limits_{w} \left[\frac{1}{2} \sum\limits_{i=1}^{n} \left(g(w, x_i) - e_{1,i} \right)^2 \right]$

зацикливание $e_{t,i} = y_i - \sum\limits_{\tau=1}^{t} g(w_\tau, x_i)$

$w_{t+1} = \ldots$

NB g может разные на каждом шаге: линейные и другие

NB На самом деле мы решили другую задачу

Обучение модели | Бустинг:

$$\sum_{i=1}^{n} L\Big(\underbrace{\sum_{t=1}^{T-1} g_t(x_i)}_{\text{фиксир.}} + \underset{\substack{\uparrow \\ \text{учим} \\ \text{argmin}}}{g_T};\, y_i\Big) \qquad \Bigg| \qquad \sum_{i=1}^{n} L\Big(g_T;\, y_i - \sum_{t=1}^{T-1} g_t(x_i)\Big)$$

Забудем про $g(\omega, \cdot)$ и сформулируем

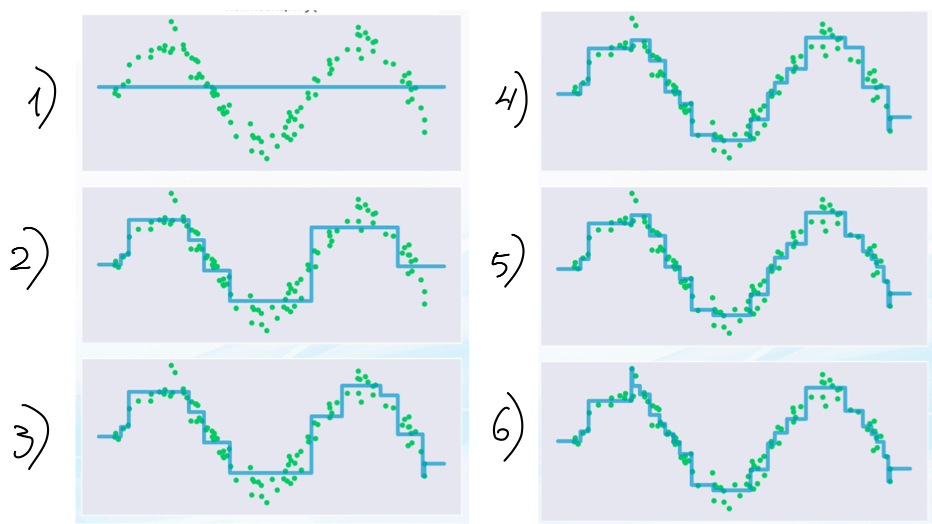
$$\min_{g_i \in \mathbb{R}} \underbrace{\frac{1}{2} \sum_{i=1}^{n} (g_i - y_i)^2}_{L(g_1 \ldots g_n)}$$

$$\nabla_{g_i} L = g_i - y_i = -e_i$$

Вспомним, что модель, построена на всех пред. шагах

Получаем, что учимся на $-\nabla L$

- мы шли против градиента => идём к min потерь
- минимизируем потери новой суммарной



много шагов бустинга = переобучение

Ансамбль взвешенный:

$$y_T(x) = \sum_{t=0}^{T} \gamma_t g_t(x)$$

$g(w_t, x)$

$t=0$ — создаем базовую модель

- $\gamma_0 = 1$
- g_0 — очень простую
  - $g_0(x) \equiv 0$
  - классиф.:
    $g_0(x) \equiv$ самый популярный
  - регрессия:
    $g_0(x) \equiv \frac{1}{n} \sum_{i=1}^{n} y_i$

от 0 к пред. итерации

Забудем про модели:

$$L(g) = \sum_{i=1}^{n} L(g_i, y_i)$$

$$\min_g L(g)$$

давайте возьмём град спуск

$$g^t = g^{t-1} - \eta \underbrace{\nabla_g L(g^{t-1})}_{\begin{pmatrix} \cdots \\ \nabla_g L(g_i^{t-1}) \\ \cdots \end{pmatrix} \Big\}n}$$

Вспомним про модели

$$w^t = \arg\min_w \sum_{i=1}^n L\left(g(w, x_i); -\nabla_g L(g_i^{t-1})\right)$$

Осталось найти γ_t:

$$\hat{\gamma}_t = \arg\min_{\gamma \in \mathbb{R}} \sum_{i=1}^n L\Big(\underbrace{g_i^{t-1}}_{y_{t-1}(x_i)} + \gamma g(w_t, x_i); y_i\Big)$$

$\hat{\gamma}_t = 1$ более практично

$$\underbrace{y_t(x_i)}_{g_i^t} = y_{t-1}(x_i) + \underbrace{\eta \hat{\gamma}_t}_{\gamma_t} g(w_t, x_i)$$

чем меньше, тем сходимость медленнее и итоговая модель тяжелее,

но модель точнее

Полный алгоритм:

инициализация: $y_0(x_i) = g_0(x_i)$

for $t = 1 \dots T$

1) $$s_i = -\nabla_g L(g)\Big|_{g = y_{t-1}(x_i)}$$

2) $$w_t = \arg\min_{w \in \mathbb{R}^d} \underbrace{\sum_{i=1}^{n} L\big(g(w, x_i),\ s_i\big)}_{\text{можно использовать случ. подвыборку}}$$

3) $$\hat{\gamma}_t = \arg\min_{\gamma \in \mathbb{R}} \overbrace{\sum_{i=1}^{n} L\big(y_{t-1}(x_i) + \gamma\, g(w_t, x_i);\ y_i\big)}$$

4) $$y_t(x_i) = y_{t-1}(x_i) + \hat{\gamma}_t \eta\, g(w_t, x_i)$$

Частные случаи:

1) MSE регрессия

$$L(z, y) = \frac{1}{2}(y - z)^2 \qquad L'_z(z, y) = z - y$$

$$s_i = -\big(y_{t-1}(x_i) - y_i\big)$$

2) MAE регрессия

$$L(z, y) = |z - y| \qquad L'_z(z, y) = \operatorname{sign}(z - y)$$

$$s_i = -\operatorname{sign}\big(y_{t-1}(x_i) - y_i\big)$$

3) Классификация $y_i \in \{-1; 1\}$

$$\sum_{i=1}^{n} \mathbb{I}[y_i z_i < 0]$$

сглаженная её аппроксимация

- $\log(1+e^{-a})$ Logit Boost
- $1-a^2$ Gentle Boost
- $\exp(-a)$ Ada Boost

Ada Boost

$$\sum_{i=1}^{n} \exp(-y_i z_i)$$

$$\sum_{i=1}^{n} \exp(-y_i \cdot y_t(x_i)) =$$

$$= \sum_{i=1}^{n} \exp(-y_i \cdot (y_{t-1}(x_i) + \gamma_t g(w_t, x_i)))$$

$$= \sum_{i=1}^{n} \underbrace{\exp(-y_i \cdot y_{t-1}(x_i))}_{\text{фиксировано},\ v_i} \cdot \exp(-y_i \underbrace{\gamma_t}_{} \underbrace{g(w_t, x_i)}_{\{-1;1\}})$$

веса объектов

$\{-1; 1\}$

модель выдает ответ $\{-1; 1\}$

предположение

$$= \sum v_i \exp(-\gamma_t) \cdot \mathbb{I}\underbrace{[y_i = g(w_t, x_i)]}_{\{1;0\}}$$

$$+ \sum v_i \exp(+\gamma_t) \cdot \mathbb{I}[y_i \neq g(w_t, x_i)]$$

$$= \exp(-\gamma_t) \cdot \underbrace{\left(\sum v_i \mathbb{I}[y_i = g(w_t, x_i)]\right)}_{Positive}$$

$$+\exp(\gamma_t) \cdot \underbrace{\left(\sum v_i \mathbb{I}[y_i \neq g(w_t, x_i)]\right)}_{Negative}$$

дифференцирую γ_t:

$$-\exp(-\gamma_t) \cdot Positive + \exp(\gamma_t) Negative = 0$$

$$\exp(2\gamma_t) = \frac{Positive}{Negative} \qquad \gamma_t = \boxed{\frac{1}{2} \ln \frac{Positive}{Negative}}$$

$$\frac{1}{2} \ln \frac{P+\varepsilon}{N+\varepsilon} \qquad \varepsilon \sim \frac{1}{n}$$

подставляем γ_t:

$$\exp(-\gamma_t) \cdot P + \exp(+\gamma_t) \cdot N =$$

$$= \sqrt{\frac{N}{P}} \cdot P + \sqrt{\frac{P}{N}} \cdot N = 2\sqrt{PN}$$

$$w_t = \arg\min_{w \in \mathbb{R}^d} \left[\sum v_i \mathbb{I}[y_i = g(w_t, x_i)] + \sum v_i \mathbb{I}[y_i \neq g(w_t, x_i)] \right]$$

Скорость сходимости:

$$\sum_{i=1}^{n} \mathbb{I}[y_i \, y_T(x_i) < 0] \leq \sum_{i=1}^{n} \underbrace{\exp(-y_i \cdot y_T(x_i))}_{V_i^T}$$

$$= \sum_{i=1}^{n} V_i^{T-1} \cdot \exp(-y_i \gamma_T \, g(w_T, x_i))$$

$$\leq \sum V_i^{T-1} (1-\gamma^2) \qquad \gamma = \min \gamma_t$$