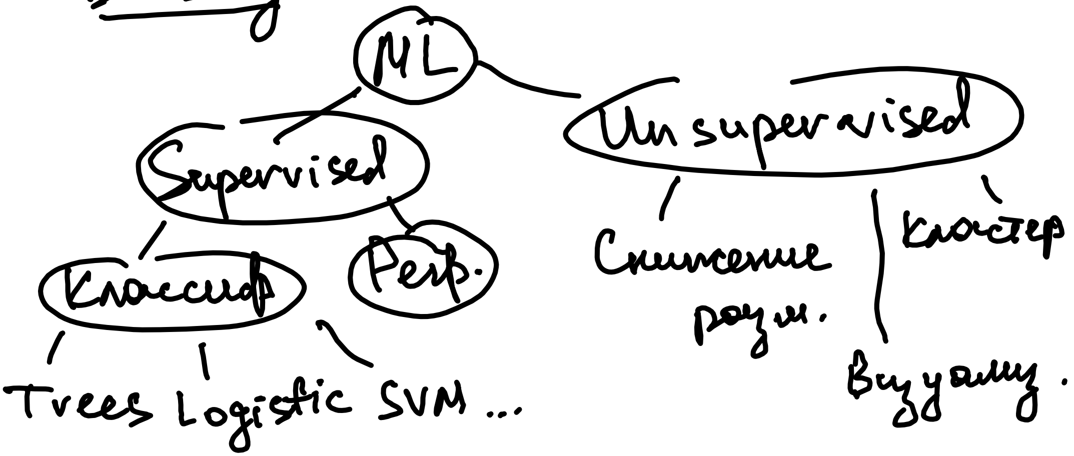
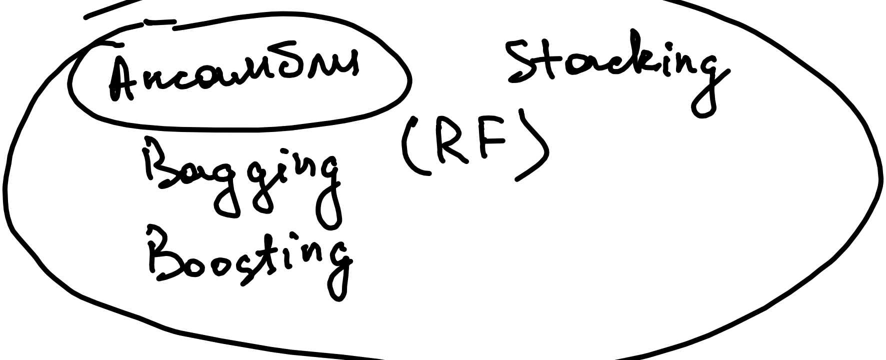
Boosting





$y \in \mathbb{R}$ $\qquad a_N(x) = \sum_{j=0}^{N} b_j(x)$ ← weak learners.

$$b_0(x) = \bar{y}$$

$$s_i^{(1)} = y_i - b_0(x_i) \Rightarrow y_i = b_0(x_i) + \underbrace{s_i^{(1)}}_{\sim b_1(x_i)}$$

$$b_1(x) = \underset{b \in \mathcal{A}}{\arg\min} \sum_{i=1}^{n} \left(b(x_i) - s_i^{(1)}\right)^2$$

$$s_i^{(2)} = y_i - b_0(x_i) - b_1(x_i) \Rightarrow$$

$$\Rightarrow y_i = \underbrace{b_0(x_i) + b_1(x_i)}_{a_1(x_i)} + \underbrace{s_i^{(2)}}_{\sim b_2(x_i)}$$

$$\ldots$$

$$\begin{cases} s_i^{(N)} = y_i - a_{N-1}(x_i) \ * \\ b_N(x) = \underset{b \in \mathcal{A}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(b(x_i) - s_i^{(N)} \right)^2 \end{cases}$$

$$a_N(x_i) = \sum_{j=0}^{N} \gamma_j b_j(x)$$

$$\sum_{i=1}^{n} L(y_i, a_N(x_i)) \to \min$$

$$\sum_{i=1}^{n} L\left(y_i, a_{N-1}(x_i) + \gamma_N b_N(x_i)\right) \to \min_{\gamma_N, b_N}$$

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

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y, z) \Big|_{z = a_{N-1}(x_i)} =$$

$$= -\left(\frac{1}{\cancel{2}} \cdot \cancel{2}(y - z) \cdot (-1) \right) \Big|_{z = a_{N-1}(x_i)} =$$

$$= y_i - a_{N-1}(x_i)$$

$$\sum_{i=1}^{n} L(y_i, a_{N-1}(x_i) + s_i) \to \min_{s_1, \dots, s_n}$$

$$s_i = y_i - a_{N-1}(x_i)$$

$$s_i = -\frac{\partial}{\partial z} L(y_i, z_i) \Big|_{z_i = a_{N-1}(x_i)}$$

$$f(x) = f(x_0) + f'(x_0) \cdot (\underbrace{x - x_0}_{\varepsilon}) + o(x - x_0)$$

$$x = x_0 + \varepsilon$$

$$f(x_0 + \varepsilon) = f(x_0) + f'(x_0) \cdot \varepsilon + o(\varepsilon)$$

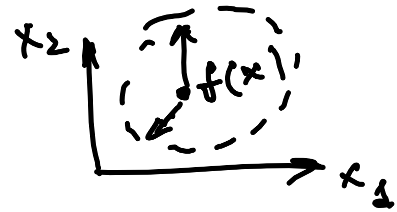
$$f(x + s) = f(x) + \nabla f(x)^T \cdot s + o(\|s\|)$$

$$\begin{cases} f(x+s) \to \min\limits_{s} \\ \|s\| = 1 \end{cases} \quad \begin{aligned} L &= f(x+s) + \lambda(s^T s - 1) \\ L &= f(x) + \nabla f(x)^T s + \lambda(s^T s - 1) \end{aligned}$$

$$\nabla_s L = \nabla f(x) + 2\lambda s = 0 \Rightarrow s = -\frac{1}{2\lambda} \nabla f(x)$$

$$s^T s = \frac{1}{4\lambda^2}\|\nabla f(x)\|^2 = 1 \Rightarrow \lambda = \pm\frac{1}{2}\|\nabla f(x)\|$$

$$\boxed{s = \pm \frac{\nabla f(x)}{\|\nabla f(x)\|}}$$



---

1) $$s_i^{(N)} = -\frac{\partial L(y_i, z_i)}{\partial z_i}\Big|_{z_i = a_{N-1}(x_i)}$$

2) $$b_N(x) = \underset{b \in A}{\operatorname{argmin}} \sum_{i=1}^{n} \left(b(x_i) - s_i^{(N)}\right)^2$$

3) $$\gamma_N = \underset{\gamma \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(y_i, a_{N-1}(x_i) + \gamma \cdot b_N(x_i)\right)$$

gradient boosting

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

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z_i)\Big|_{z_i = a_{N-1}(x_i)} =$$

(сдвиг)

$$= y_i - a_{N-1}(x_i)$$

$L(y, z) = |y - z|$

$s_i^{(N)} = -\left(\mathrm{sign}(y - z) \cdot (-1)\right) =$

$= \mathrm{sign}(y_i - a_{N-1}(x_i))$

<u>классификация</u>:

$y \in \{-1; +1\}$ $\quad p_+(x) = p(y = 1 | x)$

$a_N(x) \in \mathbb{R}$ $\quad \overset{\shortparallel}{\sigma}(a_N(x)) = \dfrac{1}{1 + \exp(-a_N(x))}$

$\mathrm{Likelyhood} = \prod_{i=1}^{n} p_+(x_i)^{[y_i = 1]} \cdot (1 - p_+(x_i))^{[y_i = -1]}$

$\log Lik = \sum_i -[y_i = 1] \log(1 + \exp(-a_N(x_i))) -$

$- [y_i = -1] \log(1 + \exp(a_N(x_i)))$

$\mathrm{Loss} = \sum_{i=1}^{n} [y_i = 1] \log(1 + \exp(-a_N(x_i))) +$

$+ [y_i = -1] \cdot \log(1 + \exp(a_N(x_i))) =$

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

$$L(y,z) = \log\left(1+\exp(-\underbrace{yz}_{\text{Margin}})\right)$$

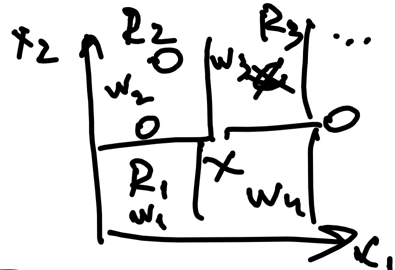
1) $$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i,z)\Big|_{z=a_{N-1}(x_i)} =$$

$$= + \frac{+y_i \exp(-y_i \cdot a_{N-1}(x_i))}{1+\exp(-y_i a_{N-1}(x_i))} = \frac{y_i}{1+\exp(y_i a_{N-1}(x_i))}$$

2) $$b_N(x) = \arg\min_{b\in A} \sum_{i=1}^{n} \left(b(x_i) - \frac{y_i}{1+\exp(y_i a_{N-1}(x_i))}\right)^2$$

3) $$\gamma_N = \arg\min_{\gamma\in\mathbb{R}} \dots$$

XGBoost, LightGBM, Catboost



Tree GB:

$$b(x) = \sum_{t=1}^{T} w_t \cdot [x\in R_t]$$

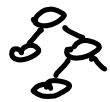
$$\sum_{i=1}^{n} L\left(y_i, a_{N-1}(x_i) + \underbrace{\gamma_N \sum_{t=1}^{T} w_{Nt}}_{w_{Nt}} [x_i\in R_t]\right)$$

$$\sum_{i=1}^{n} L\left(y_i, a_{N-1}(x_i) + \sum_{t=1}^{T} w_{Nt}\,[x_i\in R_t]\right) \to \min_{w_{N1},\dots,w_{NT}}$$

$$\sim \sum_{i:(x_i,y_i)\in R_t} L(y_i,\ a_{N-1}(x_i) + w_{Nt})$$

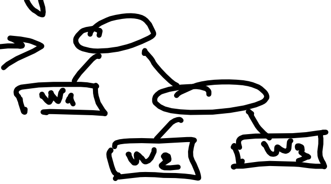
1) $s_i^{(N)} = -\dfrac{\partial}{\partial z_i} L(y_i, z_i) \Big|_{z_i = a_{N-1}(x_i)}$

2) $b_N(x) = \underset{b \in}{\operatorname{argmin}} \sum_{i=1}^{n} \left(b(x_i) - s_i^{(N)}\right)^2$



b_N = DecisionTree Regressor( )

b_N.fit(X, s_N)



3) $w_{Nt}^{*} = \underset{w\in R}{\operatorname{argmin}} \sum_{i:(x_i,y_i)\in R_t} L(y_i, a_{N-1}(x_i) + w)$

$L(y, z) = |y - z|$

3) $w_{Nt} = \underset{w\in \mathbb{R}}{\operatorname{argmin}} \sum_{R_t} |y_i - a_{N-1}(x_i) - w|$

$w_{Nt}^{*} = \underset{i:(x_i,y_i)\in R_t}{\operatorname{median}} \{ (y_i - a_{N-1}(x_i)) \}$

Семинар

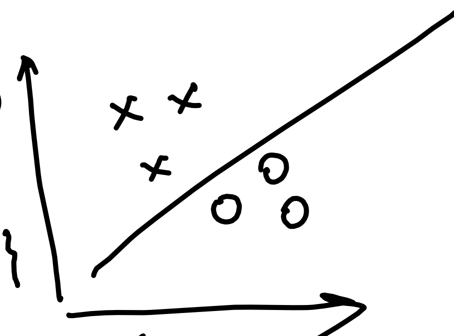
④ $f(x+\Delta x) = f(x) + Df(x)[\Delta x] + o(\|\Delta x\|)$

1) $Df(x)[\Delta x] = \nabla f(x)^T \cdot \Delta x$

2) $Df(x)[\Delta x] = tr(\nabla f(x)^T \Delta x)$

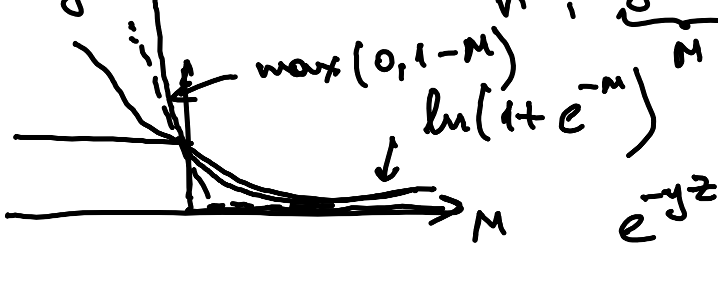
$$Df(x)[\Delta x] = \lim_{t \to 0} \frac{f(x+\Delta x \cdot t) - f(x)}{t}$$

⑤ $a(x) = sign(\langle w, x\rangle + b)$

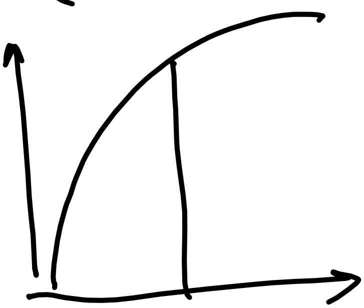


$Margin_i = y_i \langle w, x_i \rangle$, $y_i \in \{+1, -1\}$

$$Q = \frac{1}{n}\sum_{i=1}^{n}[y_i \neq a(x_i)] = \ldots = \frac{1}{n}\sum_i [\underbrace{y_i a_i}_{M} < 0]$$



(11)
$$\begin{cases} \sum_{k=1}^{d} a_k^T S a_k \to \max \\ \|a_k\|^2 = 1 \end{cases} \qquad S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$$



(16) Hard Margin

$$\rho(x, \ldots) = \frac{|w^T x + b|}{\|w\|}$$

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,b} \\ y_i(\langle w, x_i \rangle + b) \geq 1 \end{cases}$$

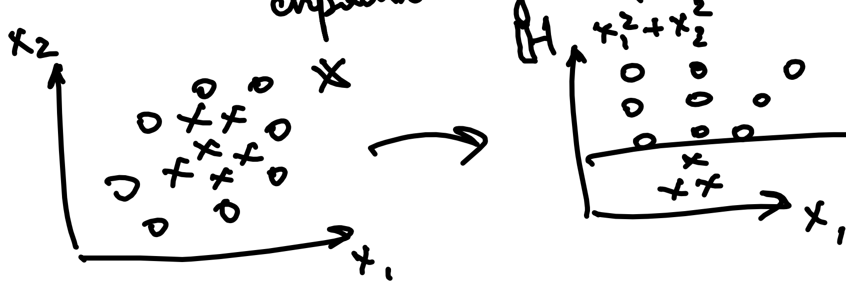
Soft Margin

$$\begin{cases} \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{n} \xi_i \to \min\limits_{w,b,\xi_i} \\ y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

(17) $K(x,z) = \langle \varphi(x), \varphi(z) \rangle$

$\varphi: \mathbb{X} \to \mathbb{H}$

спрямляющее пр-во.



$$w = \sum_{i=1}^{n} \lambda_i y_i x_i$$

$$a(x) = \operatorname{sign}(\langle w, x \rangle + b) = \operatorname{sign}\Big(\sum_i \lambda_i y_i \langle x_i, x \rangle + b\Big)$$

$$\cancel{\mathbb{E}_{X^n}\left(\mathbb{E}_{x,y}\left(y - \mu(X^n)\right)^2\right)} =$$

$$\mathbb{E}_{X^n}\left(\mathbb{E}_{x,y}\left(y - \mu(X^n)(x)\right)^2\right) =$$

$$= \mathbb{E}_{x,y}\left(y - \mathbb{E}(y|x)\right)^2 \text{ — noise}$$

$$+ \mathbb{E}_{x,y}\left(\mathbb{E}(y|x) - \mathbb{E}_{X^n}\left(\mu(X^n)(x)\right)\right)^2 \text{ — bias}$$

$$+ \mathbb{E}_{x}\left(\mathbb{E}_{X^n}\left[\mu(X^n)(x) - \mathbb{E}_{X^n}\left(\mu(X^n)(x)\right)\right]^2\right)$$

variance

$$\tilde{\mu}(x^n)(x) = \frac{1}{N}\sum_{j=1}^{N} \mu(x^j)(x)$$

bias:

$$\mathbb{E}_{x^n}\left(\frac{1}{N}\sum_{j=1}^{N} \mu(x^j)(x)\right) =$$

$$= \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{x^n}\left(\mu(x^j)(x)\right) = \mathbb{E}_{x^n}\left(\mu(x^n)(x)\right)$$

variance:

$$\underbrace{\mathbb{E}_x\big(\mathbb{E}_{x^n}}_{\tilde{\mathbb{E}}}\left[\frac{1}{N}\sum_{j=1}^{N} \mu(x^j)(x) - \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{x^n}(\mu)\right]^2\big)$$

$$= \frac{1}{N^2}\tilde{\mathbb{E}}\left[\sum_j \left(\mu(x^j)(x) - \mathbb{E}_{x^n}(\mu(x^n)(x))\right)\right]^2$$

$$\left(\Big(\sum_j a_j\Big)^2 = \sum_j a_j^2 + \sum_{i \neq j} a_i a_j\right)$$

$$= \frac{1}{N^2}\cdot\tilde{\mathbb{E}}\left(\sum_j (\mu - \mathbb{E}\mu)^2 + \sum_{j \neq k} (\mu - \mathbb{E}\mu)\cdot(\mu - \mathbb{E}\mu)\right)$$

$$= \frac{N}{N^2} \underbrace{\mathbb{E}_x \left[\mathbb{E}_{x^n} \left(\mu(x^n)(x) - \mathbb{E}_{x^n}(x^n)(x) \right)^2 \right]}_{\text{variance } (\mu)} +$$

$+ \dots$

bagging: bias —

variance ↓

boosting: bias ↓ (Trees

variance ↑ ↓ depth)