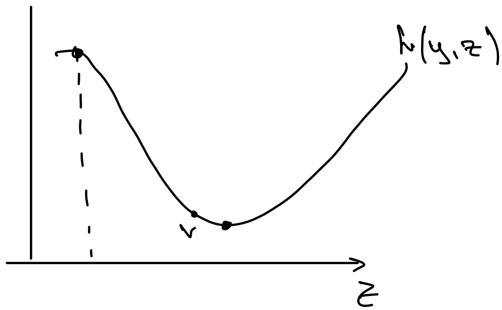
Формат:

Задача 2 листок 0



КВО w2.3.

шум: $\mathbb{E}_{x,y}[(y-E[y|x])^2]$ – шум просто как данность

смещение увеличится ↗ на баз. дереве => для композ. тоже увеличится, но это не очень страшно

разброс возрастает для одной модели, ↗ => для композ.

т.к. они более случайные модели

$$Var\left[\frac{1}{N}\sum b_i(x)\right] = \frac{Var(b_i(x))}{N} + \frac{N(N-1)}{N^2} Cov(b_i, b_j) \Rightarrow$$

=> разброс для комп. ↓

т.к. ковар. уходит

КРО 1.5

Logreg: $\frac{1}{\ell} \sum \log\left(1+e^{-y_i\langle x_i, w\rangle}\right) \longrightarrow \min\limits_{w}$

SVM: $\underbrace{\frac{1}{2}\|w\|^2}_{\text{регуляризация}} + C \sum \max\left(0, 1 - y_i\langle x_i, w\rangle\right) \longrightarrow \min\limits_{w}$

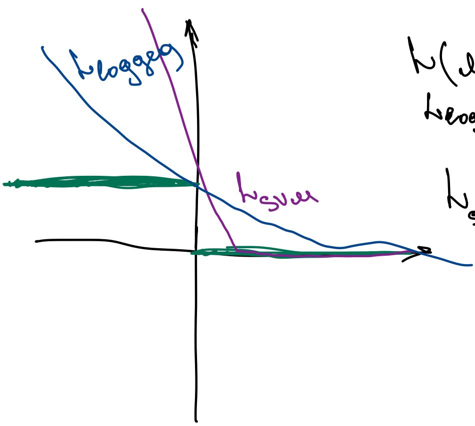
у logreg меньше смещение больше разброс

у SVM больше смещение $\nearrow$

разброс ниже $\swarrow$

Пороговая ф-я потерь:

$L(M) = [M < 0]$, $M = y_i\langle x_i, w\rangle$



$L(M) = \log\left(1+e^{-M}\right)$

L_{logreg}

$L_{svm}(M) = \max(0, 1 - M)$

У SVM больше смещение, но меньше разброс

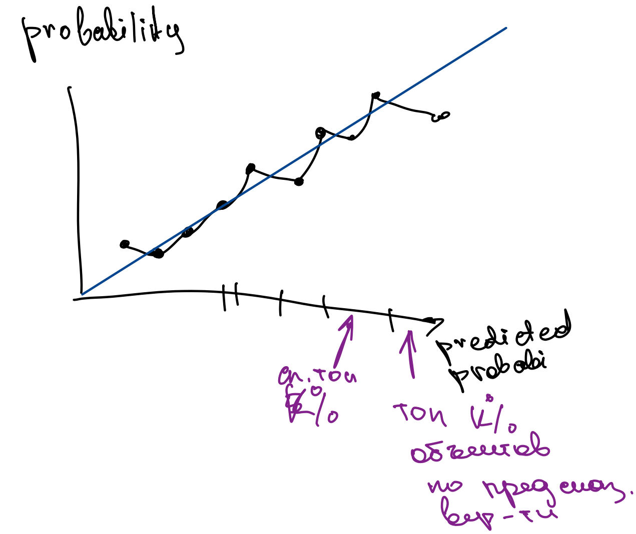
$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y, z)\Big|_{\substack{y = y_i \\ z = a_{N-1}(x_i)}}$$

— таргет для обучения N-ной модели в ГБМ

↓

MSE

Качество калибровки



SVM: предсказание вер-ти

↓

$b(x): w \Rightarrow \langle w, x\rangle$: $\left.\begin{matrix} -100 \\ 10 \\ 20 \\ \vdots \end{matrix}\right\}$ хотим в вер-ти

$\tilde{x}_i = \langle w, x\rangle \in \mathbb{R}$

⇓

учим новую модель на одной переменной

(logreg)

$a(x) = \sigma(\tilde{w}_0 + \tilde{w}_1 \tilde{x})$, $\tilde{x} = \langle w, x\rangle$

Прошлый дз

$$L_a(y,z) = \log(1+e^{-yz})$$

$$L_b(y,z) = \max(0, 1-yz)$$

$$L_c(y,z) = (yz-1)^2$$

1.1 One-vs-all

1.2 $y=-1$

$$\underset{z \in \mathbb{R}}{\arg\min}\ L(-1,z)$$

L_a: $\log(1+e^{z}) \to \min \iff z=-\infty$

L_b: $\max(0, 1+z) \Rightarrow \min \iff z=-1$

L_c: $(-z-1)^2 = (z+1)^2 \to \min \iff z=-1$

1.3

$$a(x) = \text{sign}(\langle w,x\rangle)$$

$$\tilde{a}(x) = \text{sign}(\langle w,x\rangle - 10)$$

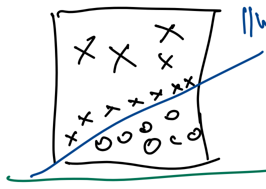
L_a: $\log\left(1+e^{-y(\langle w,x\rangle-10)}\right)$ ф-я поменялась

L_b: $\max(0, 1-y(\langle w,x\rangle-10))$ может,

$\langle w,x\rangle = 100 \Rightarrow$

$\max(0, 1-100)=0 \quad \max(0,1-90)=0$

1.4



$\|w\|$ - большое, а $\sum \xi_i = 0$

SVM:

$\|w\|$ маленькое, но $\sum \xi_i$ чуть больше

ош-ка

$$\|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \to \min$$

чем меньше $\|w\|$, тем больше разд. полоса

Если $C \searrow$, то модель может выбрать модель

2

2.1 out-of-bag

2.2

$$(b_N(x_i) - s_i)^2$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \cancel{L(s_i, b_N(x_i))} \to \min_{b_N}$$

grad-boost.

$$s_i = -\frac{\partial L(y_i, z)}{\partial z}\Bigg|_{z = \boxed{b_{N-1}(x_i)}}$$

должен быть, т.к. антиград

$a_{N-1}(x_i)$

Где ошибки?

2.3 $y_i > 0$

$a(x) < 0$ для реш. дерева на MSE

Нет, т.к. реш. деревья на MSE усред. прогнозы $\Rightarrow$ среднее > 0

Случ. лес тоже нет, т.к. ср. прогноз

На град.бустинге могут, т.к. там сдвиги

(2.е) $\sum_{j=1}^{J_N} b_{N_j}^2$ - нельзя такой рег. в отдельном дереве xgboost.

т.к. он запрещает большие веса

(3.1) error rate → min

$\frac{1}{\ell} \sum [a(x_i) \neq y_i] \to \min$

нельзя дифф., поэтому не используем

(3.2) η_t: не перескок минимума

шаг t: доля ош ↑
лог. ош ↑

Не может, т.к. мы уменьшаем лог. ош.

доля ош ↑
лог. ош ↓ - может

(3.3)

$z \geq y$ $10 \log\cosh(z-y)$

$z < y$ $\log\cosh(z-y)$

3.4) • $\sum_{j=1}^{d} |w_j|^3$ – сильно нак. за большие веса

• $\max_{j=1,..,d} |w_j|$ – не дифф

• $\sum_{i=1}^{d} [w_j \neq 0]$ – самый прореж., но мин. хз как т.к. не дифф.

4) $x_i \sim \mathcal{N}(0, \sigma^2)$

$y_i = \varepsilon_i x_i$

$\varepsilon_i \sim Exp(\lambda)$

$$\mu(X) = \frac{1}{\lambda \ell}\left(\sum_{i=1}^{\ell} x_i\right) sign(x)$$

$$\mathbb{E}[y|x] = \mathbb{E}[\varepsilon x | x] = x\,\mathbb{E}[\varepsilon|x] = x \cdot \frac{1}{\lambda} = \frac{x}{\lambda}$$

$$noise := \mathbb{E}_{x,y}[y - \mathbb{E}[y|x]]^2 =$$

$$= \mathbb{E}_{x,y}\left[y - \frac{x}{\lambda}\right]^2 = \mathbb{E}_{x,y}\left[\varepsilon x - \frac{x}{\lambda}\right]^2 = \underbrace{\mathbb{E}[x]^2}_{=\sigma^2} \cdot \underbrace{\mathbb{E}\left[\varepsilon - \frac{1}{\lambda}\right]^2}_{=(\mathbb{E}[\varepsilon - \mathbb{E}[\varepsilon]])^2 = Var[\varepsilon] = \frac{1}{\lambda^2}} =$$

$$= \frac{\sigma^2}{\lambda^2}$$

bias:

$$\underbrace{\mathbb{E}_X[\mu(X)]}_{\text{ср. модель}} = \mathbb{E}_X \frac{1}{\lambda \ell}\left(\sum x_i\right) sign(x) = \frac{1}{\lambda \ell} sign(x) \sum \mathbb{E}[x_i] = 0$$

$$\text{bias} = \mathbb{E}_x\Big[\underbrace{\mathbb{E}_X[\mu(X)]}_{\text{ср. модель}} - \underbrace{\mathbb{E}[y|x]}_{\text{лучш. модель}}\Big]^2 =$$

$$= \mathbb{E}_x\Big[0 - \frac{x}{\lambda}\Big]^2 = \frac{1}{\lambda^2}\underbrace{\mathbb{E}_x[x]^2}_{= \operatorname{Var} - (\mathbb{E}[x])^2 = \sigma^2 - 0} = \frac{\sigma^2}{\lambda^2}$$

variance:

$\mathbb{E}[y(x)] + \mathbb{E}[\ ]$

$$\mathbb{E}_{x,X}\big(\mu(X)(x)\big)^2 = \mathbb{E}_{x,X}\Big[\frac{1}{\lambda \ell}\Big(\sum x_i\Big)\operatorname{sign}(x)\Big]^2 =$$

$$= \frac{1}{\lambda^2\ell^2}\,\mathbb{E}_x\underbrace{\operatorname{sign}^2(x)}_{=1}\;\mathbb{E}_X\Big[\sum x_i\Big]^2 =$$

$$= \frac{1}{\lambda^2\ell^2}\Big(\underbrace{\sum_{i\neq j}\mathbb{E}x_i x_j}_{\text{нез} = \mathbb{E}x_i \cdot \mathbb{E}x_j = 0,\ \mathbb{E}x_i = 0,\ \mathbb{E}x_j = 0} + \underbrace{\sum_{i=1}^{\ell}\mathbb{E}x_i^2}_{= \ell\sigma^2}\Big) = \frac{\sigma^2}{\lambda^2\ell}$$

$\mathbb{E}x_i^2 = \sigma^2$