① BVD для Регуляризации

$$\mathbb{E}_{x,\varepsilon}[y(x,\varepsilon)-a(x,X)]^2 = \quad \text{bias}_X^2[a(x,X)]$$

$$= \sigma^2 + \mathrm{Var}_x(a(x,X)) + (f(x)-\mathbb{E}_X[a(x,X)])^2$$

Упражнение

$f(x)=w\cdot x$

Вероятностная модель:

$y(x,\varepsilon) = w\cdot x + \varepsilon$

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

$(y(x,\varepsilon)-a(x,X))^2$

$x \sim \mathcal{N}(0,\alpha^2)$

Так надо будет страдать с интегралами

$\hat{w} = \dfrac{\sum x_i y_i}{\sum x_i^2}$ из обучающей выборки

x – детерминированы

новое наблюдение

$$a(\tilde{x},X) = \gamma\cdot\underbrace{\frac{\sum x_i y_i}{\sum x_i^2}}_{\hat{w}}\cdot\tilde{x}$$

Подобрать γ: чтобы MSE ↓

$$\mathrm{Var}_{x,\varepsilon}(a(\tilde{x},X)) = \gamma^2\cdot\tilde{x}^2\cdot\frac{1}{(\sum x_i^2)^{\cancel{2}}}\cdot\cancel{\sum x_i^2}\cdot\overbrace{\mathrm{Var}_{x,\varepsilon}(y_i)}^{\sigma^2} =$$

y_i = $\underbrace{wx_i+\varepsilon_i}$

$$= \gamma^2\tilde{x}^2\cdot\frac{\sigma^2}{\sum x_i^2}$$

$\mathbb{E}[\varepsilon_i]=0$

$$\mathbb{E}_X[a(\tilde{x},X)] = \mathbb{E}_X\left[\gamma\cdot\frac{\sum x_i\overbrace{[wx_i+\varepsilon_i]}^{y_i}}{\sum x_i^2}\cdot\tilde{x}\right] =$$

$$= \gamma\cdot\tilde{x}\cdot w$$

$$\text{bias}_X[a(\tilde{x}, X)] = \gamma \tilde{x} w - \tilde{x} w$$

$\gamma = 1 \Rightarrow$ смещение нулевое

$$\mathbb{E}_{x,\varepsilon}[y - \underbrace{\mathbb{E}(y|x)}_{wx}]^2 = Var_{x,\varepsilon}(\varepsilon) = \sigma^2$$

$$MSE = \sigma^2 + (\gamma \tilde{x} w - \tilde{x} w)^2 + \gamma^2 \tilde{x}^2 \cdot \frac{\sigma^2}{\sum x_i^2} \to \min_\gamma$$

$$\cancel{2}\cancel{\tilde{x}} w(\gamma \tilde{x} w - \tilde{x} w) + \cancel{2}\gamma \tilde{x}^{\cancel{2}} \cdot \frac{\sigma^2}{\sum x_i^2} = 0$$

$$w^2 \cancel{\tilde{x}} \gamma \underbrace{- w^2 \cancel{\tilde{x}}} + \gamma \cancel{\tilde{x}} \cdot \frac{\sigma^2}{\sum x_i^2} = 0$$

$$\gamma = \frac{w^2}{w^2 + \frac{\sigma^2}{\sum x_i^2}} = \frac{1}{1 + \frac{\sigma^2 / \sum x_i^2}{w^2}} \neq 1$$

$$\frac{Var(\hat{w})}{\mathbb{E}(\hat{w})^2} \qquad \frac{\sigma^2}{\alpha^2}$$

$\Rightarrow \exists \gamma \neq 1$

$$\text{bias}_X(a(\tilde{x}, X)) > 0$$

$$Var_X(a(\tilde{x}, X)) \downarrow$$

Задание

$$y(x, \varepsilon) = w \cdot x + \varepsilon$$

$$\hat{w}^{Ridge} = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \lambda} \qquad \sum_i \left(y(x,\varepsilon) - a(x, X)\right)^2 + \lambda w^2 \to \min_w$$

λ – ?

λ и γ – одно и то же

Вывести формулу как связаны λ и γ

$$\mathbb{E}_{X,\varepsilon}\left[y(x,\varepsilon) - a(x, X)\right]^2 = \dots \to \min_\lambda$$

$$\Rightarrow \lambda^* = \dots$$

Это не точно, но вроде $\lambda^* = \frac{1}{\gamma^*}$

прогнозирование	интерпретир.
смещение $\uparrow$ разброс $\downarrow$ $\Rightarrow$ MSE $\downarrow$ Регуляризация	• смещ $= 0$ Эконометрика • Попытка расшифровать ML-алгоритм LIME SHAPE

② Бэггинг

$b_n(x)$ – базовый алгоритм (независимы друг от друга разный X_{train})

$$a_N(x) = \frac{1}{N} \sum_{j=1}^{N} b_j(x)$$

Пусть у меня есть много несмещённых алг.
Но они все слабые $\Rightarrow$ MSE большое
$Var_X(b(x, X)]$ большая

$$Var_X(a_N(x, X)) = \frac{1}{N^2} \cdot \sum_{j=1}^{N} Var_X[b_j(x, X^j)] =$$

$$= \frac{1}{N^2} \cdot N \cdot Var_X\, b_j(x, X) = \frac{Var_X(b_j(x, X))}{N}$$

Дисперсия падает в N раз

$$bias_X[a_N(x, X)] = f(x) - \mathbb{E}_X[a_N(x, X)] =$$

$$= f(x) - \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_X[b_j(x, X^j)] = f(x) - \mathbb{E}[b_j(x, X)] =$$

$$= bias_X[b_j(x, X)]$$

Смещение не изменилось

$\Rightarrow$ MSE на тесте $\Downarrow$

Случ. лес — развитие идеи бэггинга для деревьев.

$\boxed{\text{Глубокие деревья} \quad \begin{matrix} \downarrow \text{смещ.} \\ \uparrow \text{разброс} \end{matrix}}$

3 Бустинг в задаче регрессии

$$a_N(x) = \sum_{j=1}^{N} b_j(x)$$

Каждая новая модель исправляет ошибки предыдущей

$\boxed{\text{1999 Jerome Friedman}}$

A — семейство алгоритмов (деревья)

Оптимизируем MSE

$$\frac{1}{2} \sum_{i=1}^{n} [a_N(x_i) - y_i]^2 \longrightarrow \min_a$$

$$a_N(x) = \sum_{j=1}^{N} b_j(x)$$

$$b_1(x) = \arg\min_{b \in A} \frac{1}{2} \sum_{i=1}^{n} (b(x_i) - y_i)^2$$

$$s_i^{(1)} = y_i - a_1(x_i) \quad \text{„остатки"}$$

$$b_2(x) = \underset{b \in A}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} \left(b(x_i) - s_i^{(1)}\right)^2$$

$$a_2(x) = \underbrace{b_1(x)}_{\substack{\text{совершал} \\ \text{ошибку} \\ s_i^{(1)}}} + \underbrace{b_2(x)}_{\substack{\text{предсказывает} \\ \text{ошибку } s_i^{(1)} \text{ и} \\ \text{фиксит её}}}$$

$$s_i^{(2)} = y_i - a_2(x_i)$$

$$b_3(x) = \underset{b \in A}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} \left(b(x_i) - s_i^{(2)}\right)^2$$

$$a_3(x) = a_2(x) + b_3(x)$$

. . . . .

$$s_i^{(N)} = y_i - a_{N-1}(x_i) = y_i - \sum_{j=1}^{N-1} b_j(x_i)$$

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

„Остатки" s_i в общем случае мы будем искать через производную ошибки по прогнозу.

$$s_i^{(N)} = -\frac{\partial}{\partial z}\left[\frac{1}{2}(y_i - z)^2\right]\Bigg|_{z = a_{N-1}(x_i)} = y_i - a_{N-1}(x_i)$$

④ Градиентный бустинг

$L(y, z)$ — дифф. ф. потерь

$$a_N(x) = \sum_{j=0}^{N} \gamma_j \cdot b_j(x)$$ — взвеш. сумма базовых алгоритмов

$b_0(x)$ — наивный прогноз

$\gamma_0 = 1$

$b_0 = 0$ $\quad b_0 = \bar{y}$ $\quad b_0 =$ доля самого частого класса $\quad \dots$

] у нас уже есть $a_{N-1}(x)$ и я хочу $b_N(x)$

$$\sum_{i=1}^{n} L(y_i, a_{N-1}(x_i) + \gamma_N \cdot b_N(x)) \to \min_{\substack{b \in A \\ \gamma_N}}$$

b_N — могу взять что угодно!

$$\sum_{i=1}^{n} L(y_i, a_{N-1}(x_i) + \gamma_N \cdot s_i) \to \min_{s_1, s_2, \dots, s_n}$$

Для MSE я хотел $s_i = y_i - a_{N-1}(x_i)$

У меня новая ф. потерь, $L(y, z)$

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

$$b_N(x) = \arg\min_{b \in A} \sum_{i=1}^{n} (b(x_i) - s_i)^2$$

информация об L уже учтена в S

$\Rightarrow$ Всегда для поиска s используем MSE

$$\gamma_N = \arg\min_{\gamma \in \mathbb{R}} \sum_{i=1}^{n} L(y_i, a_{N-1}(x_i) + \gamma \cdot b_N(x_i))$$

Классический GBM

$$a_N(x) = \sum_{j=1}^{N} b_j(x) \cdot \gamma_j$$

Чаще всего на них забивают т.к. используют деревья, а в них можно просто корректировать прогнозы в листьях

b_0 - наивная, $\gamma_0 = 1$

<u>Обучение N-ой модели:</u>

$$s_i = -\left.\frac{\partial L(y_i, z)}{\partial z}\right|_{z = a_{N-1}(x)}$$

Направления куда хочется сдвинуть композицию на x_i, чтобы ошибка упала

① $\frac{1}{n} \sum_{i=1}^{n} (b_N(x_i) - s_i)^2 \to \min\limits_{b_N}$

② $\frac{1}{n} \sum_{i=1}^{n} L(y_i, a_{N-1}(x_i) + \gamma_N \cdot b_N(x_i)) \to \min\limits_{\gamma_N}$

⑤ Регуляризация

Бустинг жёстко переобучается

- Смещение — бустинг ↓ его даже для слабых базовых алг.
- Разброс — может ↑ из-за переобучения

<u>Логично:</u> брать смещенные модели с маленьким разбросом

⇒ обычно берут неглубокие деревья

→ глубина базового дерева
→ число деревьев
→ структура дерева

} гиперпарам.

xgboost catboost lightGBM

они предлагают разную регул. и разные способы учить деревья

Общее:

базовые модели простые => надо не очень им доверять

$$a_N(x_i) = a_{N-1}(x_i) + \underbrace{\eta}_{\text{learning rate}} \cdot \gamma_N \cdot b_N(x_i)$$

$$\eta \in [0; 1]$$

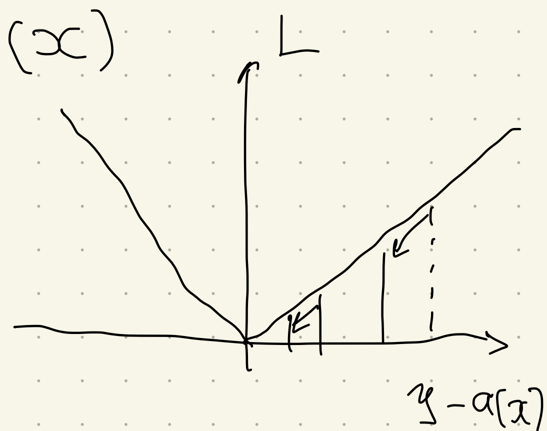
При $\eta \downarrow$ надо $\uparrow N$

Упражнение

$$L(y, z) = |y - z| = \begin{cases} y - z, & z < y \quad -1 \\ z - y, & z \geq y \quad +1 \end{cases}$$

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

$$= \begin{cases} 1, & z < y \\ -1, & z \geq y \end{cases}$$



$y_1 = 0 \quad a_{N-1}(x_1) = 5 \quad L = 5$

$y_2 = 0 \quad a_{N-1}(x_2) = 50 \quad L = 50$

Мы хотим ↓ ошибку линейно

$s_1 = \text{sign}(0 - 5) = -1$

$s_2 = \text{sign}(0 - 50) = -1$

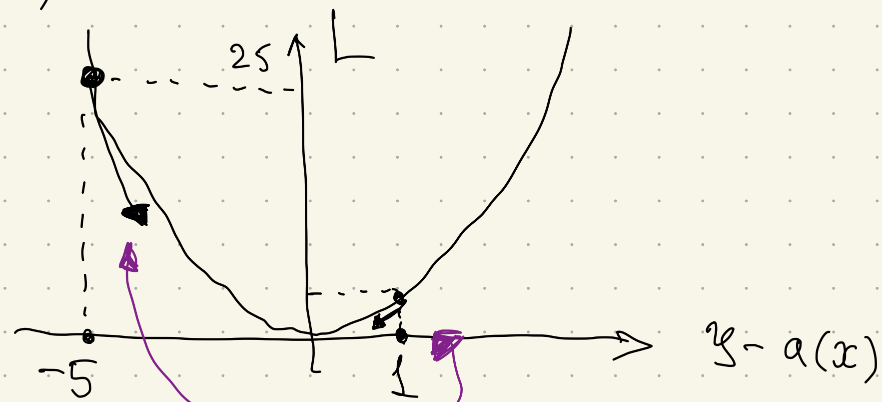
Просто двигаемся к правильному ответу не обращая внимание на размер ошибки

Упражнение

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

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

y_i	5	3	-2
$a_{N-1}(x_i)$	4	8	-1
S_i	1	-5	-1



небольшое ↓ серьезной ошибки лучше, чем доведение маленькой до нуля

$$\sum(b_N(x_i)-S_i)^2\to\min_b$$

$$a_N(x_i)=a_{N-1}(x_i)+\gamma_N b_N(x_i)$$

Упражнение

$$L(y, z) = \frac{1}{2}\left(10 \cdot [z \geq y] + [z < y]\right) \cdot (y - z)^2$$

$y_1 = 0 \quad a_{N-1}(x) = 5 \qquad L = 125$

$y_2 = 0 \quad a_{N-1}(x) = -5 \qquad L = 12.5$

Различие в 10 раз

$$s_i = -\left.\frac{\partial L(y,z)}{\partial z}\right|_{z = a_{N-1}(x_i)} = \left(10\,[z \geq y] + [z < y]\right) \cdot (y - z)$$

$s_1 = -50$

$s_2 = 5$

$\Rightarrow$ модель больше старается на 1ом наблюдении

Дальше — $\sum_i \left(b_N(x_i) - s_i\right)^2 \to \min\limits_{b \in A}$

$s_i = y_i - a_{N-1}(x_i)$ — тогда бы оба объекта имели один приоритет