① Бустинг в бустинге глубже

$L(y, z)$ – дифф. ф. потерь

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

можно забить

Когда бустинг над деревьями подбор γ_j меняют на изменение листьев

Обучение N-ой модели:

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

"псевдоостатки"

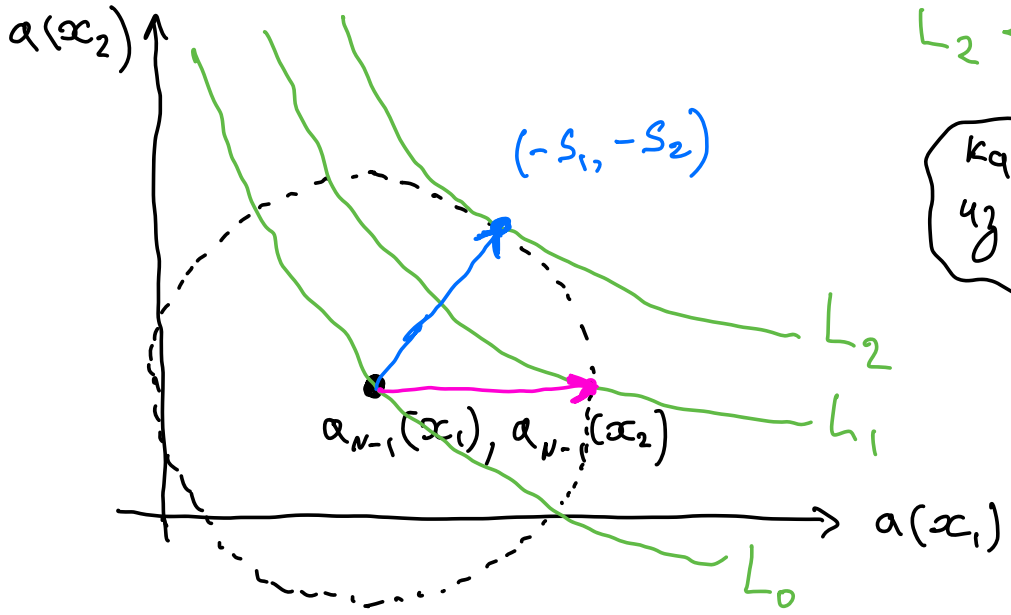
① $$\frac{1}{n}\sum_{i=1}^{n}(b_N(x) - s_i)^2 \to \min_{b_N}$$

② $$\frac{1}{n}\sum_{i=1}^{n} L(y_i;\ a_{N-1}(x) + \gamma \cdot b_N(x)) \to \min_{\gamma}$$

① Почему $b_N(x)$ обучается на s_i, а не $y_i - a_{N-1}(x_i)$?

→ Высокий риск переобучиться на $y_i \sim a_{N-1}(x_i)$

→ Игнорирование исходной ф. потерь



$(b(x_i) - \varepsilon_i)^2$ $\quad$ $(b(x_i) - s_i)^2$

$\varepsilon_i = a_N(x_i) - y_i$

② Почему базовая модель учится на MSE?

а) Разложение ∀ разумной ф. в ряд Тейлора до $2^{\text{ого}}$ слагаемого

б) Сдвиги s_i уже включают инфу про L

Разумно аппроксимировать s_i с одинаковой важностью

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

$y_1 = 0 \quad a_{N-1}(x_1) = 5 \quad L = 125 \quad s_1 = -50$

$y_2 = 0 \quad a_{N-1}(x_2) = -5 \quad L = 12.5 \quad s_2 = 5$

Если мы будем учить b_N на L, тогда при $z \geq y$ будет странное усл. штрафа

$(y-z)^2 \qquad |y-z|$

$$\sum_{i=1}^{n} (b_N(x_i) - s_i)^2 = \sum_{i=1}^{n} (b_N^2(x_i) - 2 \cdot s_i \cdot b_N(x_i) + s_i^2)$$

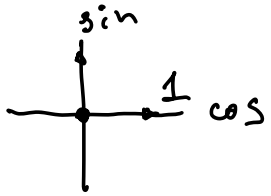
$$\sum_{i=1}^{n} b_N^2(x_i) - 2 \cdot \|s\| \cdot \|b_N(x_i)\| \cdot \cos(s, b_N(x)) \longrightarrow \min_{b_N(x)}$$

$$\langle x, y \rangle = \sum x_i y_i$$

$$\cos(x, y) = \frac{\langle x, y \rangle}{\|s\| \cdot \|b_N(x_i)\|}$$

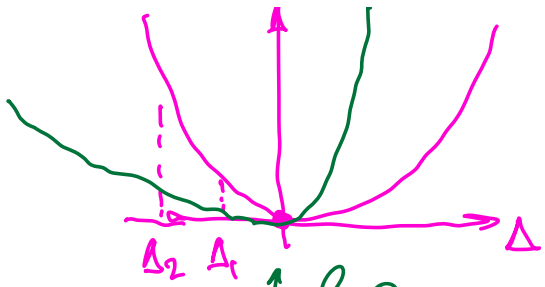
$$\|x\| = \sqrt{\sum x_i^2}$$

Т.е. мы хотим с некоторыми оговорками, чтобы $\cos(s, b_N) \uparrow \Rightarrow \angle(s, b_N) = 0$



Про a)

- $L(y, z) = g(\overbrace{y - z}^{\Delta})$
- $g(0) = 0$
- $|\Delta_1| \le |\Delta_2| \Rightarrow g(\Delta_1) \le g(\Delta_2)$



$$L(y, z) = g(y - z) \approx \overbrace{g(0)}^{0} + \underbrace{g'(0)}_{0} \cdot (y - z) + \frac{g''(0)}{2!}(y - z)^2 =$$

$$= C \cdot (y - z)^2$$

③ Почему $b_n(x)$ обычно используют неглубокие деревья?

→ глубокие деревья дольше учить

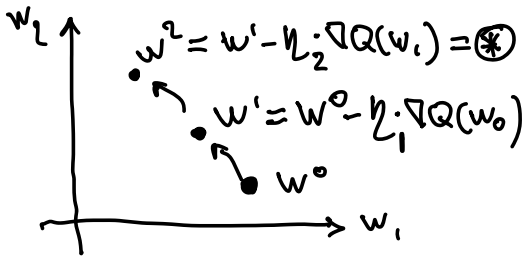
→ если базовая модель сложная, её легко переобучить.

Мы учим „направленный" ансамбль ⇒ переобучение усиливается

④ Почему говорят, что градиентный бустинг — это градиентный спуск в пространстве прогнозов композиции

на обучающей выборке?

$$w_k = w_{k-1} - \eta_k \cdot \nabla_w Q(w_{k-1})$$

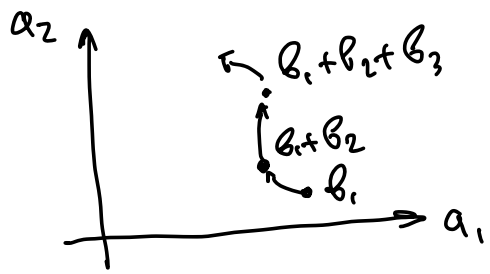


$$\circledast = w^0 - \eta_0 \nabla Q(w_0) - \eta_2 \nabla Q(w_1)$$

$$\hat{w} = \sum_{k=1}^{N} \left(-\eta_k \nabla Q(w_{k-1}) \right)$$

прогноз S_i:

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



А можно ли имитировать модификации град. спуска для бустинга?

Да, можно! [статья от 2020 года]

Momentum:

$$h_0 = 0$$

$$h_k = \alpha \cdot h_{k-1} + \eta_k \nabla_w Q(w_{k-1})$$

$$w_k = w_{k-1} - h_k$$

$a_N(x)$ – композиция алгоритмов

$b_N(x) \simeq \nabla Q$ базовые модели

$h_N(x)$ – "инерция"

$a_0(x) \quad b_0(x) \quad h_0(x)$ – init

$$b_N(x) = \underset{b_N}{\operatorname{argmin}} \sum_{i=1}^{n} \left[b_N(x_i) - \left. \frac{\partial L(y_i, z)}{\partial z} \right|_{z = a_{N-1}(x)} \right]^2$$

$$a_N(x) = a_{N-1}(x) + h_N(x)$$

$$h_N(x) = \alpha \cdot h_{N-1}(x) + \eta \cdot b_N(x)$$

Быстрее сойдётся
хранит больше моделей

(+) В статье есть пара хитростей

(5) Зачем вообще другие модели, если бустинг такой классный?

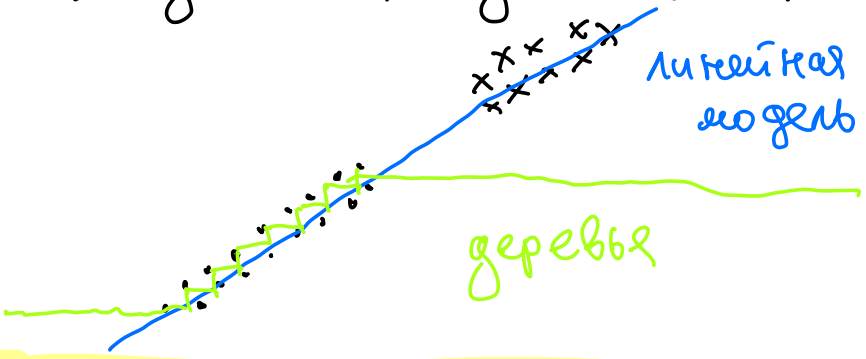
→ Линейные лучше интерпретируются

→ Если много признаков или выборка маленькая бустинг может работать плохо

→ Бустинг Vs Нейросети

ТАБЛИЦЫ	картинки/тексты/видео

→ Бустинг не умеет экстраполировать



② Бустинг для классификации

$y_i \in \{-1, 1\}$

$y_i - a_{N-1}(x_i) \in \{-2, 0, 2\}$ | ?

Давайте договоримся:

$a_N(x_i)$ — уверенность в 1ом классе (логит)

$$a_N(x_i) = \log \frac{\mathbb{P}(y=1|x)}{1 - \mathbb{P}(y=1|x)}$$

$$\mathbb{P}(y=1|x) = \frac{1}{1 + e^{-a_N(x_i)}} \qquad \sigma(t) = \frac{1}{1+e^{-t}}$$

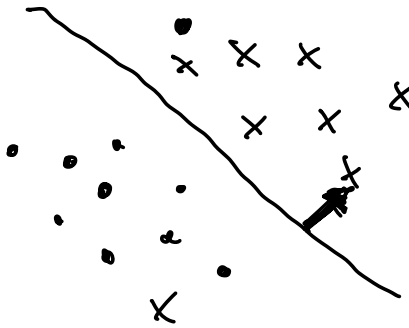
$$\sigma(a_N(x_i)) = \mathbb{P}(y=1|x)$$

Давайте учиться на logloss

$$L(y, z) = \log(1 + \exp(-yz)) \qquad y \in \{-1, 1\}$$

$$y \ln \frac{1}{1+\exp(-z)} + (1-y)\cdot(1-\dots) \qquad y \in \{0, 1\}$$

Учиться на $y_i - a_{N-1}(x_i)$ — плохая идея



$$M_i = y_i \cdot a(x_i)$$

$$\langle w, x \rangle$$

$y_i - a(x_i)$ $\quad a(x_i) \to y_i$

$y_i \, a(x_i) \to 0$

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

$$= \frac{y}{\exp(yz)+1}\Big|_{z=a_{N-1}(x_i)}$$

$\exp(-yz)$ $\quad \exp(yz)$

$$\frac{1}{n}\sum_{i=1}^{n}\left[b_N(x_i) - \frac{y_i}{1+\exp(y_i \cdot a_{N-1}(x_i))}\right]^2 \to \min_{b_N}$$

$$y_i \cdot a_{N-1}(x_i) \simeq +\infty \qquad s_i \simeq 0$$

не надо $\uparrow$ отступ

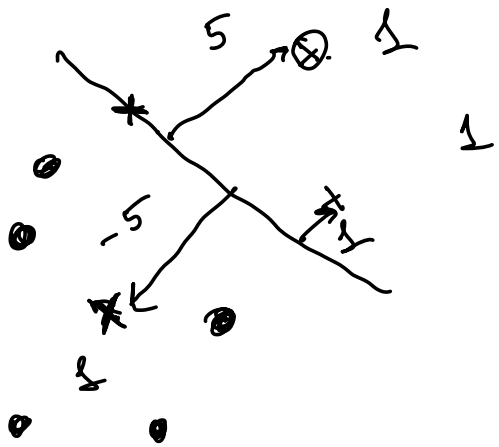
0 $\quad |s_i| = 1/2$ сомнения $\Rightarrow$ двигаем

$-\infty$ $\quad |s_i| = 1$ выброс

уверен в неверном ответе

мы не делаем на выбросах сильный акцент

$\exp(-y_i \cdot a_{N-1}(x_i))$ — мера важности объекта x_i на N-ой итерации бустинга



exp(-5)

exp(-1)

exp(5)