Введение в анализ данных

Лекция 12

Композиции моделей

Евгений Соколов

esokolov@hse.ru

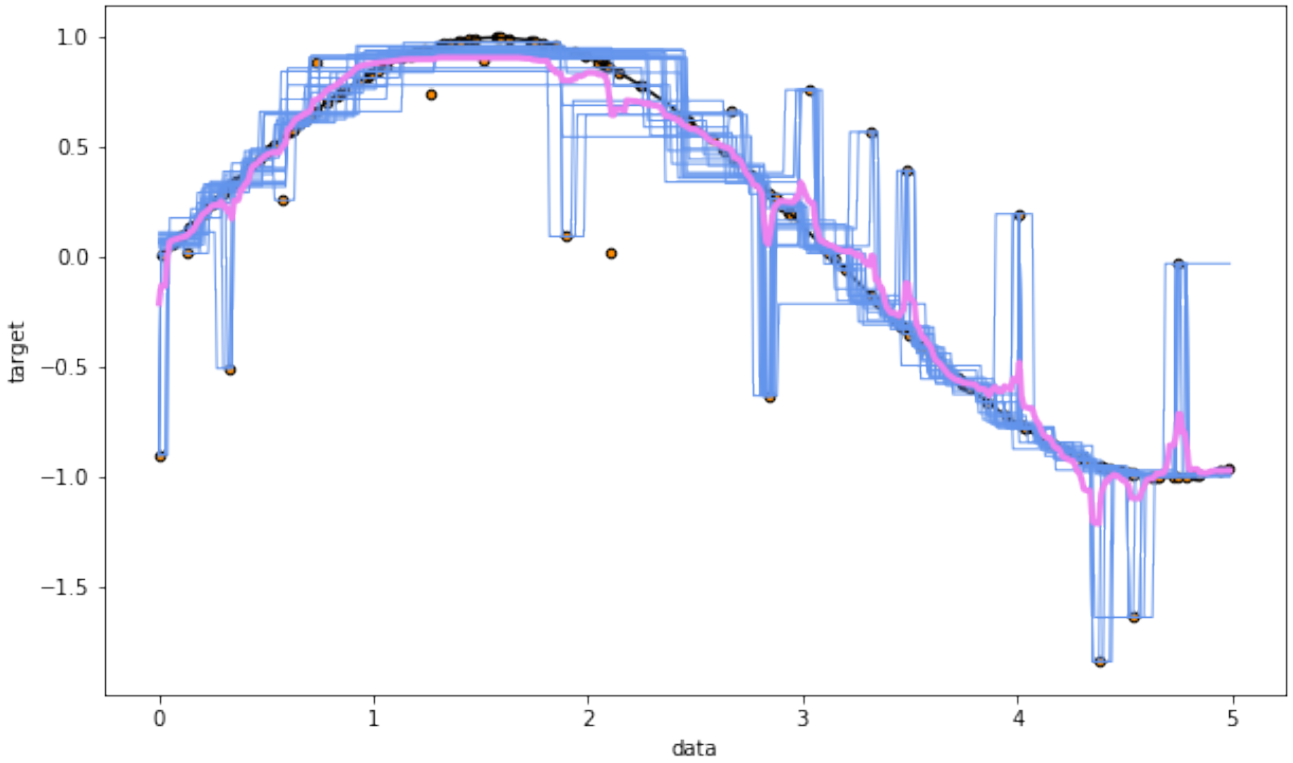
НИУ ВШЭ, 2020

Исправление ошибок моделей и идея бустинга

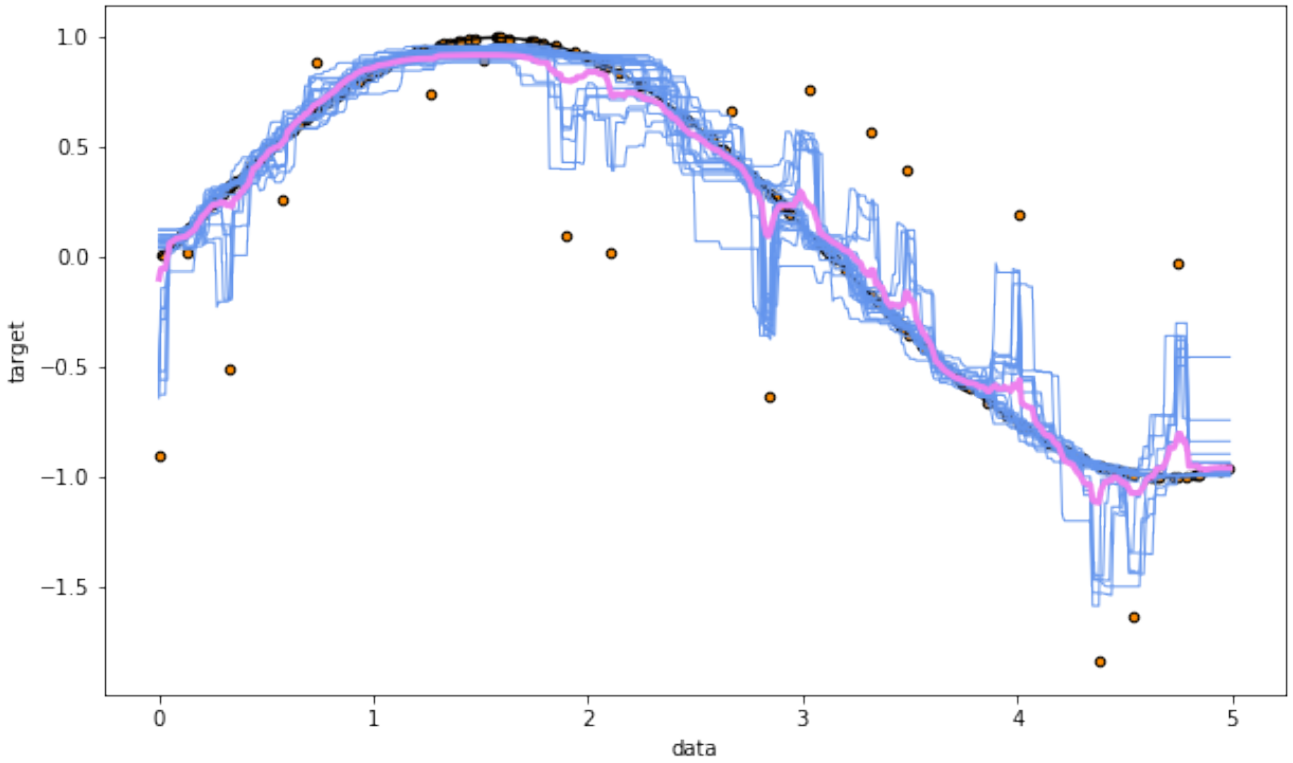
Бэггинг

- Смещение $a_N(x)$ такое же, как у $b_n(x)$
- Разброс $a_N(x)$:
- $\frac{1}{N}\left(\text{разброс } b_n(x)\right) + \text{ковариация}\left(b_n(x), b_m(x)\right)$
- Если базовые модели независимы, то разброс уменьшается в N раз!
- Чем более похожи выходы базовых моделей, тем меньше эффект от построения композиции

Смещение и разброс: деревья



Смещение и разброс: бэггинг



Проблемы бэггинга

- Если базовая модель окажется смещённой, то и композиция не справится с задачей
- Базовые модели долго обучать и применять, дорого хранить

Идея бустинга

- Возьмём простые базовые модели
- Будем строить композицию последовательно и жадно
- Каждая следующая модель будет строиться так, чтобы максимально корректировать ошибки построенных моделей

Идея бустинга

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение первой модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, b_1(x_i)\big) \to \min_{b_1(x)}$$

Идея бустинга

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, a_{N-1}(x_i) + b_N(x_i)\big) \to \min_{b_N(x)}$$

Идея бустинга

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, a_{N-1}(x_i) + b_N(x_i)\big) \to \min_{b_N(x)}$$

Идея бустинга

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, a_{N-1}(x_i) + b_N(x_i)\big) \to \min_{b_N(x)}$$

- Непонятно, как обучать дерево на такое в общем случае

Резюме

- В бустинге базовые модели обучаются последовательно
- Каждая следующая корректирует ошибки уже построенных
- В общем случае получается функционал, на который может быть сложно обучать деревья

Бустинг для среднеквадратичной ошибки

Идея бустинга

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, a_{N-1}(x_i) + b_N(x_i)\big) \to \min_{b_N(x)}$$

Бустинг для MSE

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(a_{N-1}(x_i) + b_N(x_i) - y_i)^2 \to \min_{b_N(x)}$$

Бустинг для MSE

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\Big(b_N(x_i) - \big(y_i - a_{N-1}(x_i)\big)\Big)^2 \to \min_{b_N(x)}$$

Бустинг для MSE

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\Big(b_N(x_i) - \underbrace{\big(y_i - a_{N-1}(x_i)\big)}_{s_i^{(N)}}\Big)^2 \to \min_{b_N(x)}$$

Бустинг для MSE

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(b_N(x_i) - s_i^{(N)}\right)^2 \to \min_{b_N(x)}$$

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

Первая итерация

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(b_1(x_i)-y_i)^2 \to \min_{b_1(x)}$$

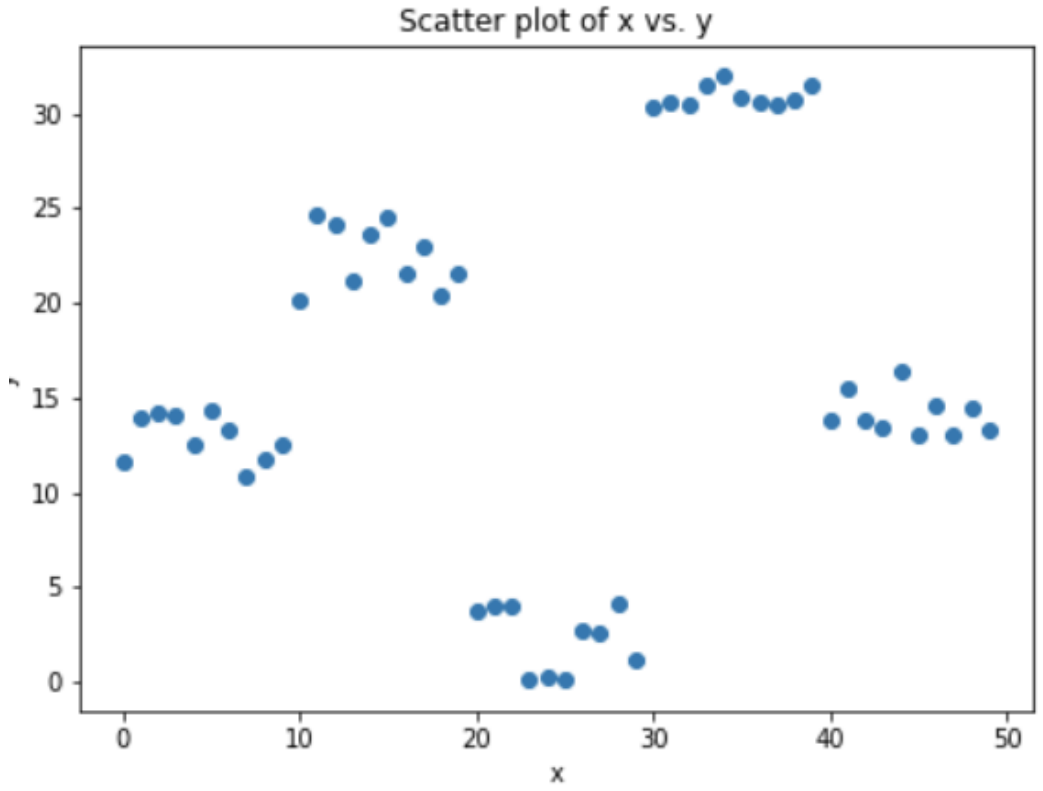
Вторая итерация

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\Big(b_2(x_i)-\big(y_i-b_1(x_i)\big)\Big)^2 \to \min_{b_2(x)}$$

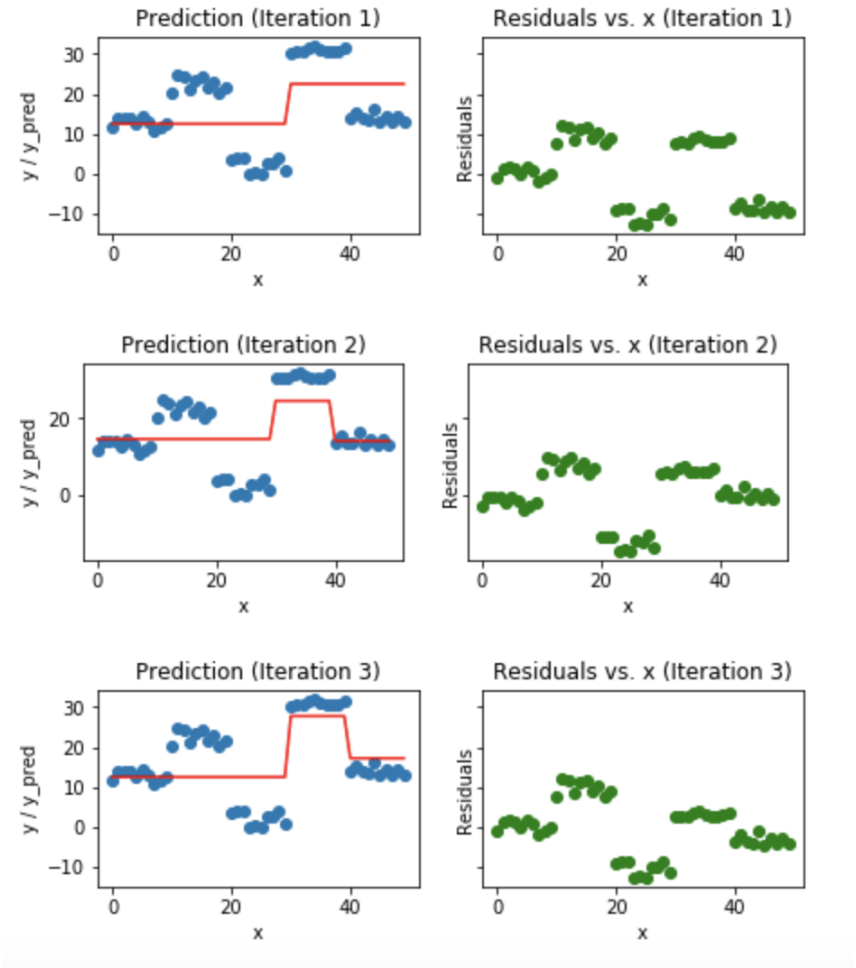
Третья итерация

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\Big(b_3(x_i)-\big(y_i-b_1(x_i)-b_2(x_i)\big)\Big)^2 \to \min_{b_3(x)}$$

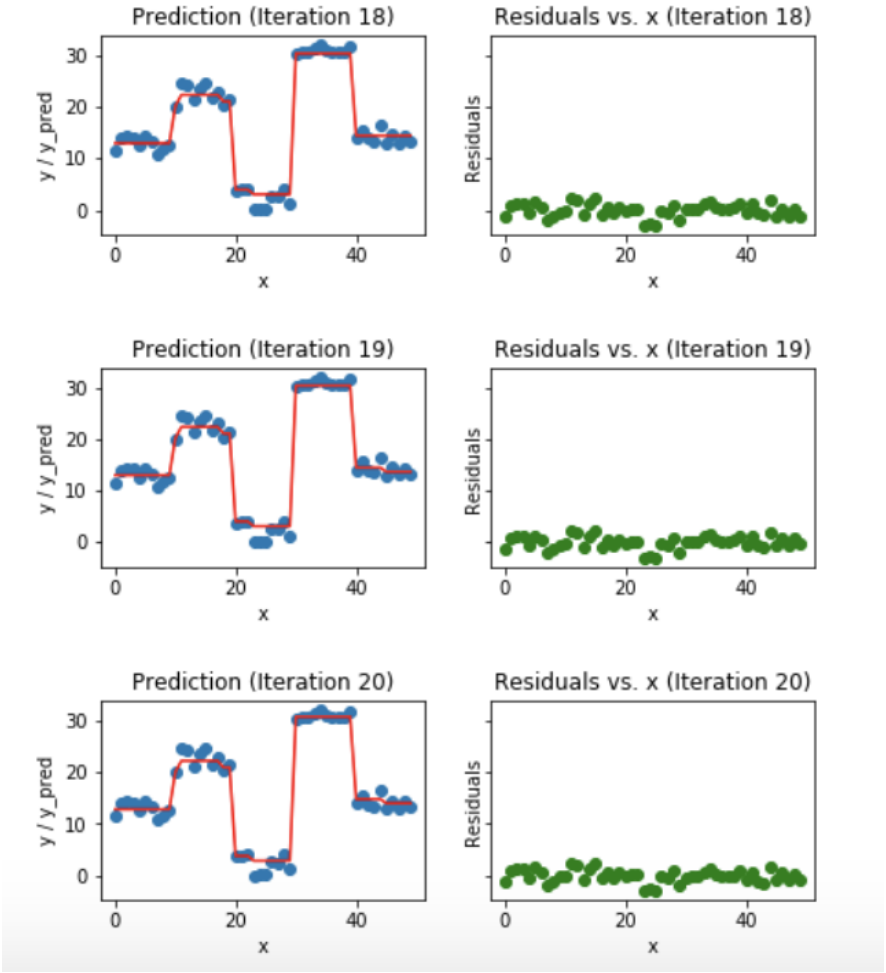
Визуализация



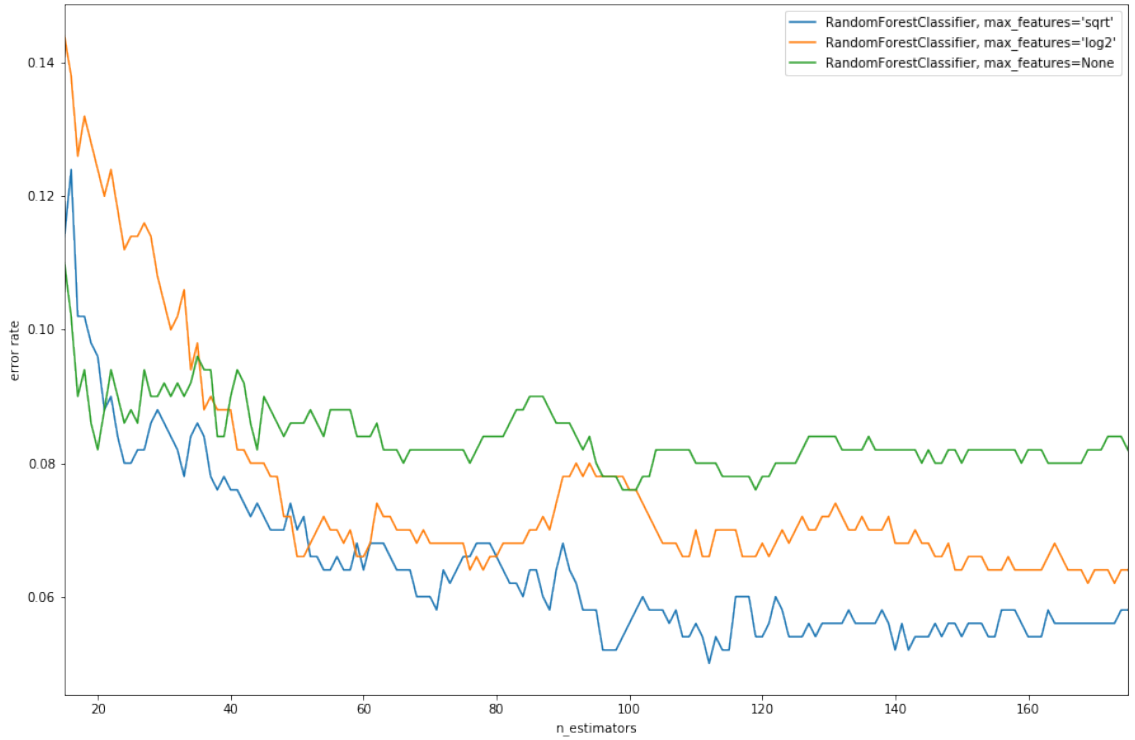
Визуализация



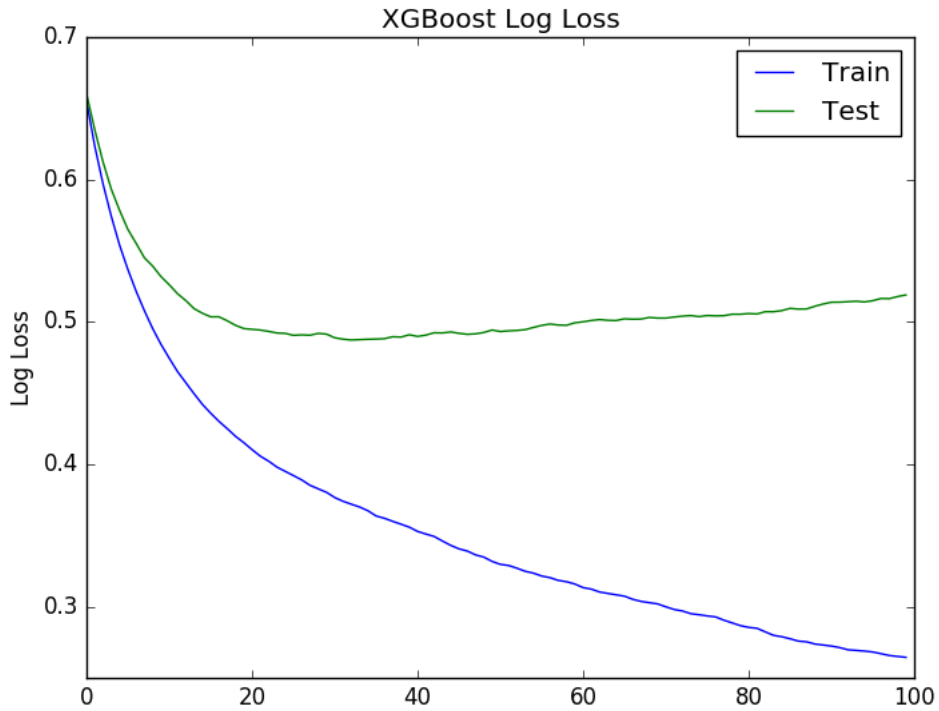
Визуализация



Random Forest



Ошибка бустинга на обучении и тесте



Резюме

- В случае с MSE обучение базовых моделей сводится к обычной процедуре обучения с заменой целевой переменной
- Бустинг может переобучаться, поэтому надо следить за ошибкой на тестовой выборке

Сложности с произвольной функцией потерь

Задача обучения базовой модели

$$a_N(x) = \sum_{n=1}^{N} b_n(x)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, a_{N-1}(x_i) + b_N(x_i)\big) \to \min_{b_N(x)}$$

Задача обучения базовой модели

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i, a_{N-1}(x_i) + b_N(x_i)\big) \to \min_{b_N(x)}$$

- Может, просто обучаться на остатки, как в MSE?

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\big(y_i - a_{N-1}(x_i), b_N(x_i)\big) \to \min_{b_N(x)}$$

Логистическая функция потерь

$$a_N(x) = \text{sign} \sum_{n=1}^{N} b_n(x)$$

$$L(y, z) = \log(1 + \exp(-yz))$$

- Может, просто обучаться на остатки, как в MSE?

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log\left(1 + \exp\left(-\left(y_i - a_{N-1}(x_i)\right) b_N(x_i)\right)\right) \to \min_{b_N(x)}$$

- Если $y_i = a_{N-1}(x_i)$, то объект не участвует в обучении
- Иначе $y_i - a_{N-1}(x_i) = \pm 2$

Логистическая функция потерь

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log\left(1+\exp\left(-\frac{y_i - a_{N-1}(x_i)}{2}\,b_N(x_i)\right)\right) \to \min_{b_N(x)}$$

- Если $y_i = a_{N-1}(x_i)$, то объект не участвует в обучении
- Если $y_i \neq a_{N-1}(x_i)$, то базовая модель учится выдавать корректный класс

Логистическая функция потерь

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log\left(1+\exp\left(-\frac{y_i - a_{N-1}(x_i)}{2}b_N(x_i)\right)\right) \to \min_{b_N(x)}$$

- $y_i = +1, \sum_{n=1}^{N-1} b_n(x_i) = -0.5 \to$ надо $b_N(x_i) > 0.5$
- $y_i = +1, \sum_{n=1}^{N-1} b_n(x_i) = -100 \to$ надо $b_N(x_i) > 100$
- Но на обоих объектах будет одинаково максимизироваться отступ
- На объектах с корректными ответами никак не контролируется выход $b_N(x)$