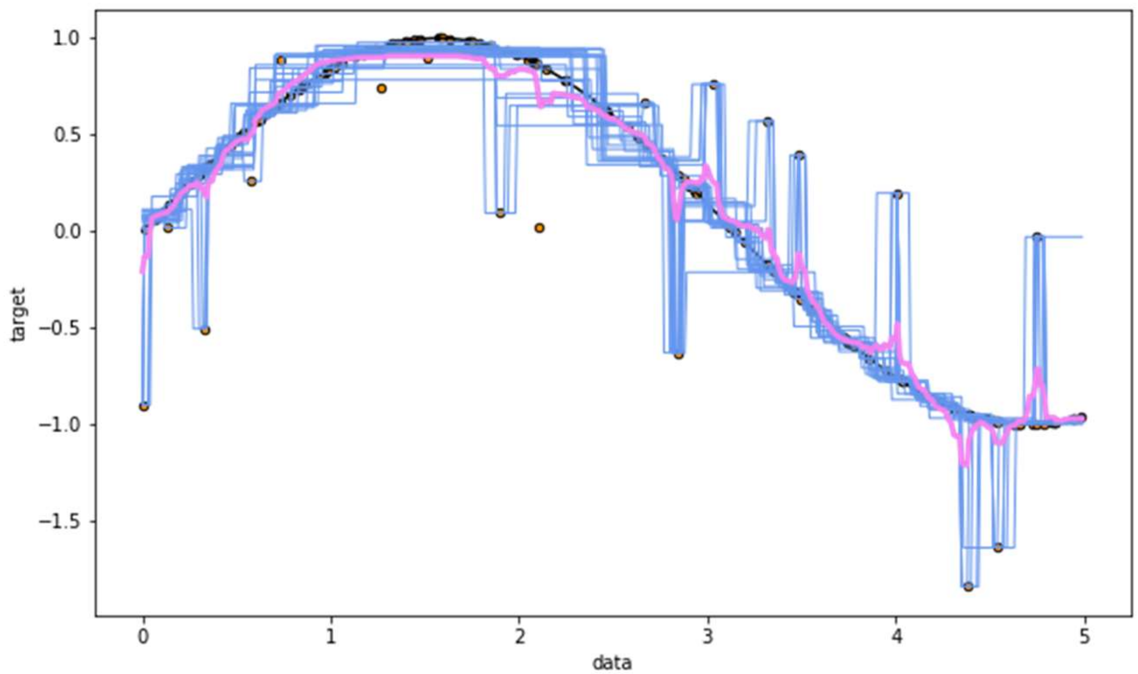
Исправление ошибок моделей и идея бустинга

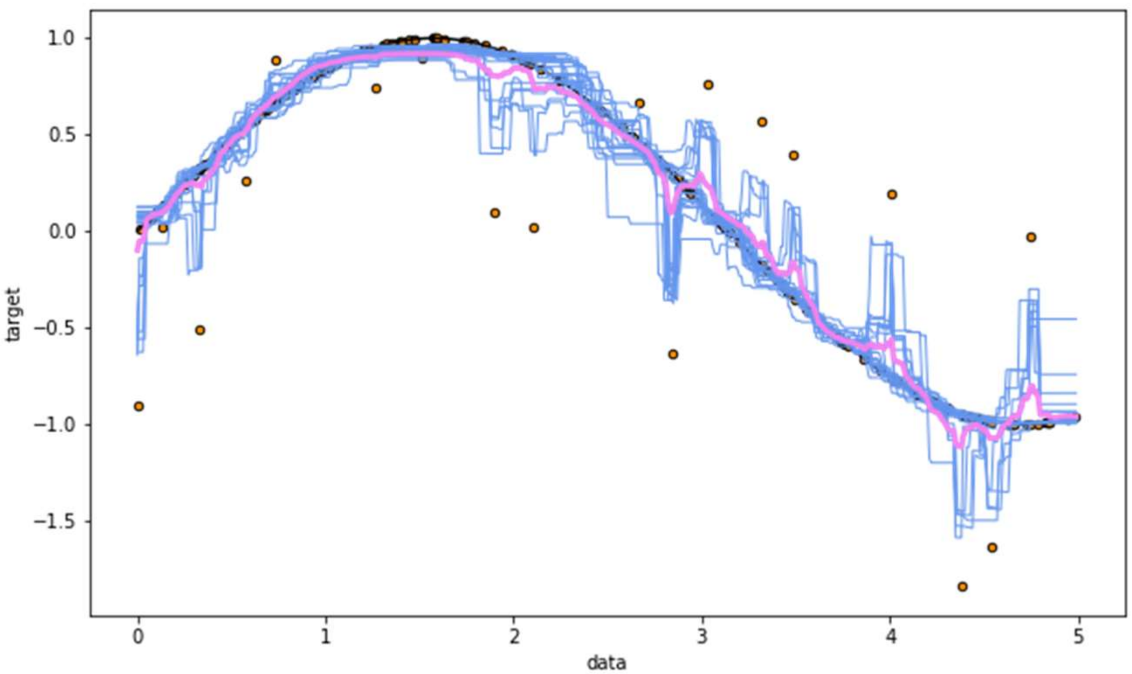
Бэггинг

- Смещение $a_N(x)$ такое же, как у $b_n(x)$
- Разброс $a_N(x)$:
- $\frac{1}{N}\big(\text{разброс } b_n(x)\big) + \text{ковариация}\big(b_n(x), b_m(x)\big)$
- 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

- Бустинг (англ. boosting — усиление)
- 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 (b_1):

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

Идея бустинга

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

- Обучение первой модели (b_1):

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

- Пытаемся подобрать модель b_1, минимизирующую ошибку

Идея бустинга

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

- Обучение N-й модели (b_N) :

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L\Big(y_i, \underbrace{a_{N-1}(x_i)}_{\text{фиксировано}} + \underbrace{b_N(x_i)}_{\text{учится}}\Big) \to \min_{b_N(x)}$$

Идея бустинга

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

- Обучение N-й модели (b_N) :

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

- Пытаемся подобрать модель b_N, минимизирующую ошибку итоговой композиции $a_N = a_{N-1} + b_N$

Идея бустинга

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

- Обучение N-й модели (b_N) :

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

Идея бустинга

- Обучение N-й модели (b_N) :

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

- Нет такого алгоритма машинного обучения, который учил бы «добавку»

- Попробуем понять, как можно сформулировать задачу с точки зрения ML

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:

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

Идея бустинга

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

- Обучение N-й модели:

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

- MSE:

$$L(y, \hat{y}) = (y - \hat{y})^2$$

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:

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

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

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

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
- Если b_N научится выдавать остатки $s_i^{(N)}$, то задача будет решена идеально

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
- Если b_N научится выдавать остатки $s_i^{(N)}$, то задача будет решена идеально

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

Пример

- $y_i = 12$
- $a_{N-1}(x_i) = 10$
- $s_i^{(N)} = ?$
- $b_N(x_i) = ?$
- $a_N(x_i) = ?$

Пример

- $y_i = 12$
- $a_{N-1}(x_i) = 10$
- $s_i^{(N)} = 2$
- $b_N(x_i) = 2$
- $a_N(x_i) = 12$

Первая итерация

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

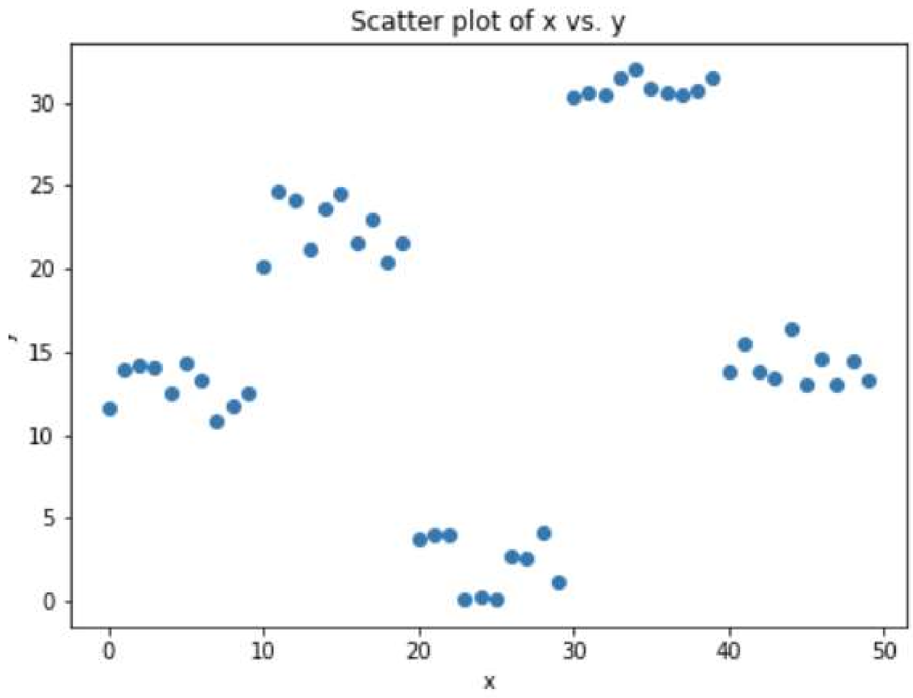
Вторая итерация

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

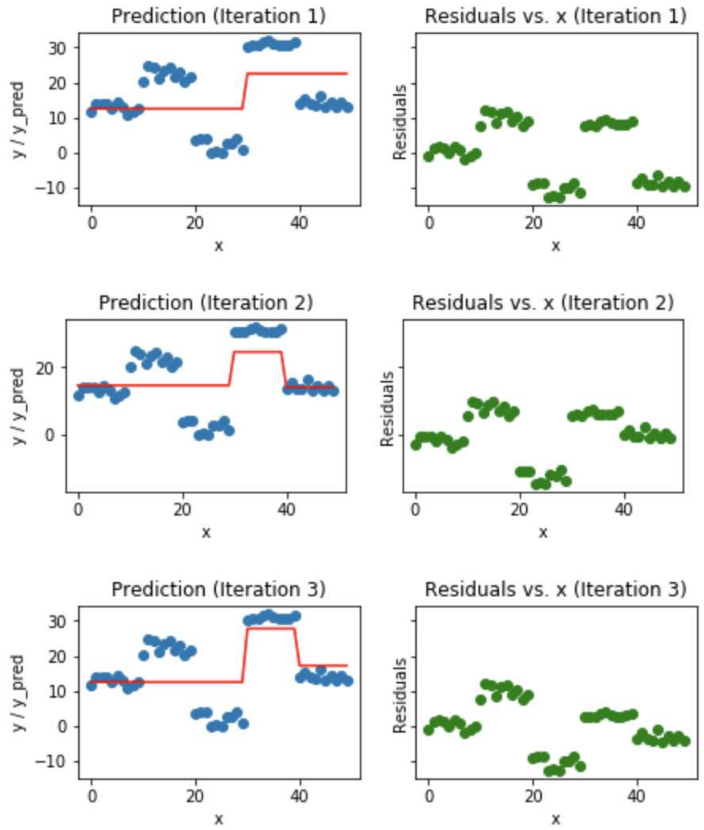
Третья итерация

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

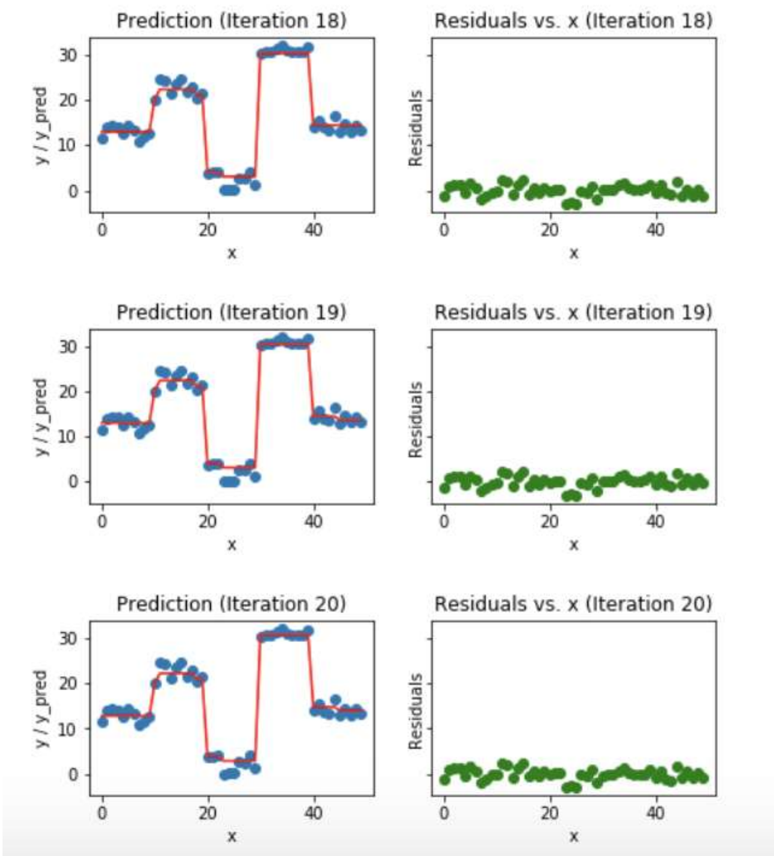
Визуализация



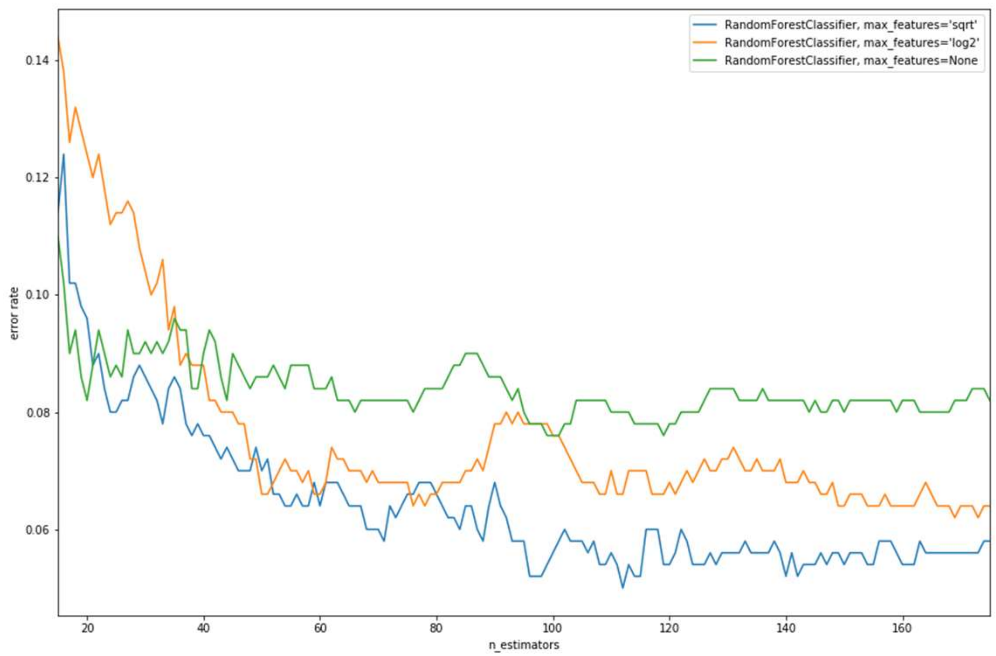
Визуализация



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?

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

Задача обучения базовой модели

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

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

- Хотим, чтобы модель b_N выдавала $y_i - a_{N-1}(x_i)$

Логистическая функция потерь

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

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

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

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

Логистическая функция потерь

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

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

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

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

Логистическая функция потерь

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

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

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

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

- Если $y_i = a_{N-1}(x_i)$, то объект не участвует в обучении
- Тогда модель $b_N(x_i)$ может выдавать что угодно и испортить композицию

Логистическая функция потерь

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

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

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

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

- 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 $b_N(x_i) > 100$

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 $b_N(x)$

MSLE

- Mean Squared Logarithmic Error (среднеквадратичная логарифмическая ошибка)

$$L(y, z) = (\log(z+1) - \log(y+1))^2$$

MSLE

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

$$L(y, z) = (\log(z + 1) - \log(y + 1))^2$$

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

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

MSLE

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

$$L(y, z) = (\log(z+1) - \log(y+1))^2$$

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

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

- Аргумент второго логарифма может оказаться отрицательным

MSLE

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

y_i	$a_{N-1}(x_i)$	$b_N(x_i)$	**Улучшение MSLE композиции**	**Улучшение функционала базовой модели**
1000	100	2	0.09	13.7
2	0	2	1.2	1.2

Резюме

- Нельзя заменить обучение добавки к композиции на обучение базовой модели на отклонение от ответов
- Не учитываются особенности функции потерь

Градиентный бустинг в общем виде

Задача обучения базовой модели

- 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

- Обучение N-й модели:

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

- Как посчитать, куда и как сильно сдвигать $a_{N-1}(x_i)$, чтобы уменьшить ошибку?

Задача обучения базовой модели

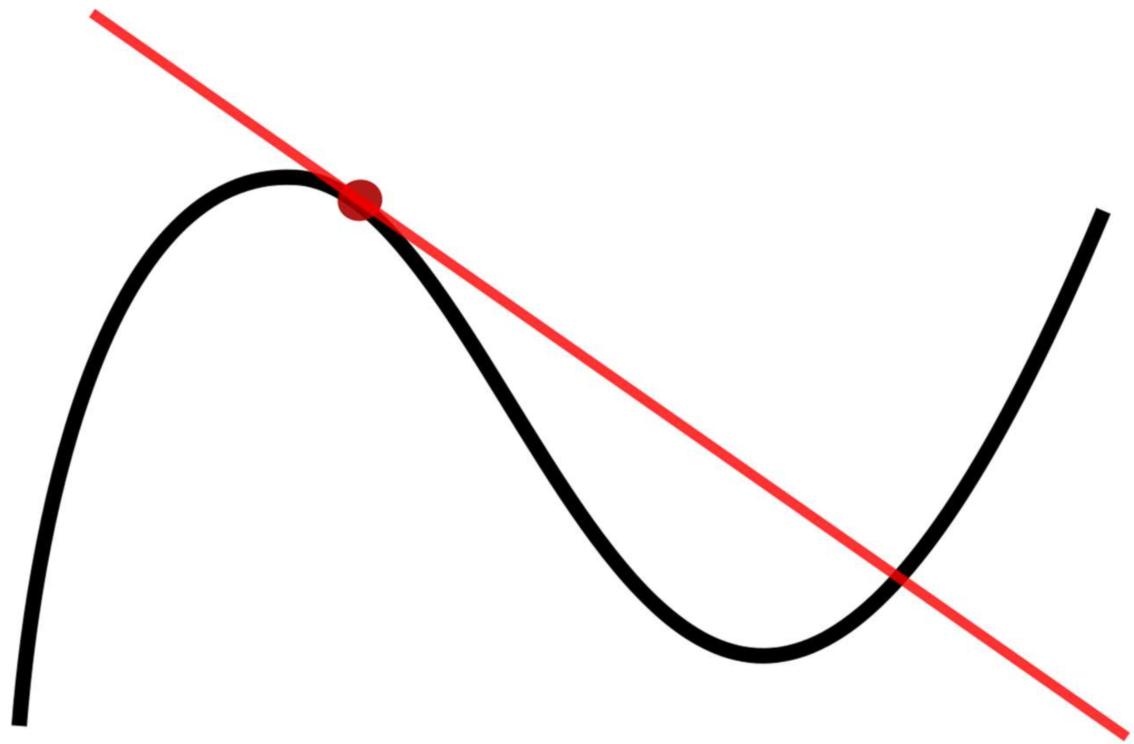
- Обучение N-й модели:

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

- Как посчитать, куда и как сильно сдвигать $a_{N-1}(x_i)$, чтобы уменьшить ошибку?

- Посчитать производную

Производная



Задача обучения базовой модели

- Обучение N-й модели:

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

- Посчитаем антипроизводную:

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

Задача обучения базовой модели

- Обучение N-й модели:

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

- Посчитаем антипроизводную:

$$s_i^{(N)} = -\underbrace{\frac{\partial}{\partial z}}_{\text{производная по z}} L(\overbrace{y_i}^{\text{правильный ответ}}, \underbrace{z}_{\text{прогноз модели}})\Bigg|_{\underbrace{z=a_{N-1}(x_i)}_{\text{в качестве z подставляем в результат предсказание композиции}}}$$

Задача обучения базовой модели

- Посчитаем антипроизводную:

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

- Знак показывает, в какую сторону сдвигать прогноз на x_i, чтобы уменьшить ошибку композиции на нём

- Величина показывает, как сильно можно уменьшить ошибку, если сдвинуть прогноз

- Если ошибка почти не сдвинется, то нет смысла что-то менять

Градиентный бустинг

- Обучение N-й модели:

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

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

Градиентный бустинг

- Обучение N-й модели:

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

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

- Таким образом, мы обучаем базовую модель b_N так, чтобы на x_i она выдавала $s_i^{(N)}$

Градиентный бустинг

- Обучение N-й модели:

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

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

- Как бы градиентный спуск в пространстве алгоритмов
- Базовая модель будет делать корректировки на объектах так, чтобы как можно сильнее уменьшить ошибку композиции
- Сдвиги учитывают особенности функции потерь

Градиентный бустинг для MSE

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z)\bigg|_{z=a_{N-1}(x_i)} = -\frac{\partial}{\partial z}\frac{1}{2}(z - y_i)^2\bigg|_{z=a_{N-1}(x_i)} =$$

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

Градиентный бустинг для MSE

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z)\Big|_{z=a_{N-1}(x_i)} = -\frac{\partial}{\partial z}\frac{1}{2}(z - y_i)^2\Big|_{z=a_{N-1}(x_i)} =$$

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

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

Градиентный бустинг для асимметричной функции

$$L(y, z) = \frac{1}{2}([z < y](z - y)^2 + 5[z \geq y](z - y)^2)$$

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

$$= [z < y](y - z) + 5[z \geq y](y - z)$$

Градиентный бустинг для асимметричной функции

$$s_i^{(N)} = [z < y](y - z) + 5[z \geq y](y - z)$$

- $y_i = 10, a_{N-1}(x_i) = 5$: $s_i = 5$
- $y_i = 10, a_{N-1}(x_i) = 15$: $s_i = -25$

Градиентный бустинг для логистической функции потерь

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

$$= -\frac{\partial}{\partial z} \log(1 + \exp(-y_i z))\bigg|_{z=a_{N-1}(x_i)} =$$

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

Градиентный бустинг для логистической функции потерь

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

$$= -\frac{\partial}{\partial z} \log(1 + \exp(-y_i z))\Big|_{z=a_{N-1}(x_i)} =$$

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

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

Градиентный бустинг для логистической функции потерь

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

- Отступ большой положительный: $\frac{y_i}{1+\exp\left(y_i a_{N-1}(x_i)\right)} \approx 0$

- Отступ большой отрицательный: $\frac{y_i}{1+\exp\left(y_i a_{N-1}(x_i)\right)} \approx \pm 1$

Градиентный бустинг для логистической функции потерь

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

- Отступ большой положительный: $\frac{y_i}{1+\underbrace{\exp\big(y_i a_{N-1}(x_i)\big)}_{\to\,\infty}} \approx 0$

- Отступ большой отрицательный: $\frac{y_i}{1+\underbrace{\mathrm{e}\quad\big(y_i a_{N-1}(x_i)\big)}_{\to\,0}} \approx \pm 1$

Градиентный бустинг для логистической функции потерь

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

- $y_i = +1, a_{N-1}(x_i) = -0.7: s_i = 0.67$
- $y_i = +1, a_{N-1}(x_i) = 2: s_i = 0.12$

Резюме

- Чтобы учесть особенности функции потерь, можно посчитать её производные в точке текущего прогноза композиции

- Базовую модель будем обучать на эти производные (со знаком минус)

Гиперпараметры и регуляризация в бустинге

Градиентный бустинг

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

- Обучение N-й модели:

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

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

Глубина деревьев

- Градиентный бустинг уменьшает смещение базовых моделей
- Разброс может увеличиться

Глубина деревьев

- Градиентный бустинг уменьшает смещение базовых моделей
- Разброс может увеличиться
- Поэтому в качестве базовых моделей стоит брать…

Глубина деревьев

- Градиентный бустинг уменьшает смещение базовых моделей
- Разброс может увеличиться
- Поэтому в качестве базовых моделей стоит брать **неглубокие** деревья

Гиперпараметры

- Глубина базовых деревьев
- Число деревьев N

Проблемы бустинга

- Сдвиги показывают направление, в котором надо сдвинуть композицию на всех объектах обучающей выборки
- Базовые модели, как правило, очень простые
- Могут не справиться с приближением этого направления

Проблемы бустинга

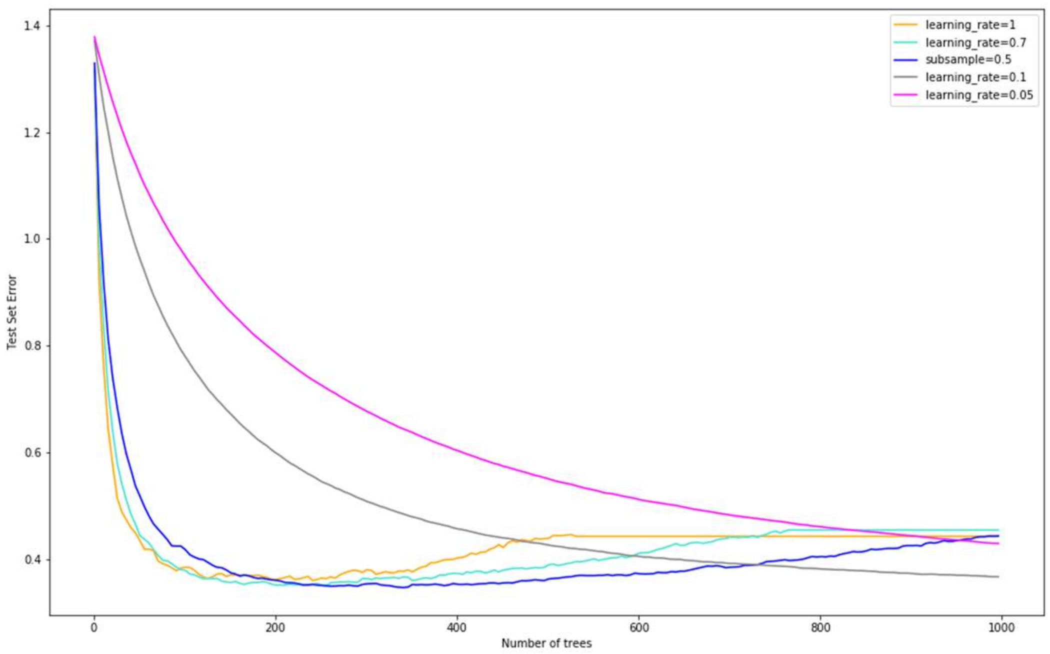
- Сдвиги показывают направление, в котором надо сдвинуть композицию на всех объектах обучающей выборки
- Базовые модели, как правило, очень простые
- Могут не справиться с приближением этого направления
- Выход: добавлять деревья в композицию с небольшим весом

Длина шага

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

- $\eta \in (0, 1]$ — длина шага (learning rate)
- Можно сказать, что это регуляризация композиции
- Снижает вклад каждой модели в композицию
- Чем меньше η, тем больше надо деревьев

Длина шага



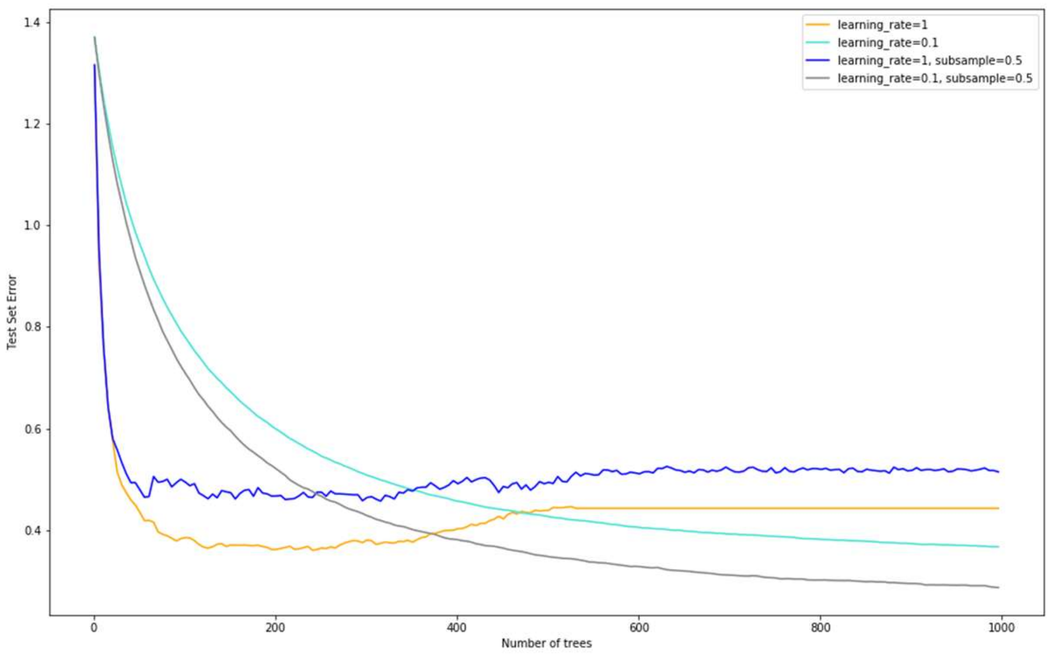
learning_rate=1
learning_rate=0.7
subsample=0.5
learning_rate=0.1
learning_rate=0.05
Test Set Error
Number of trees
1.4
1.2
1.0
0.8
0.6
0.4
0
200
400
600
800
1000

Рандомизация

- Можно обучать деревья на случайных подмножествах признаков
- Бустинг уменьшает смещение, поэтому итоговая композиция всё равно получится качественной
- Может снизить переобучение

- Можно обучать деревья на подмножествах объектов — способ борьбы с шумом в данных

Рандомизация



Гиперпараметры

- Глубина базовых деревьев
- Число деревьев N
- Длина шага
- Размер подвыборки для обучения
- и т.д.

Резюме

- Чтобы снизить переобучение, можно добавлять модели в композицию с небольшими весами
- Также может помочь обучение моделей на подвыборках

Полезные ссылки

- https://www.gormanalysis.com/blog/gradient-boosting-explained/
- https://youtu.be/3CC4N4z3GJc
- https://en.wikipedia.org/wiki/Gradient_boosting
- https://dyakonov.org/2017/06/09/градиентный-бустинг/