

Введение в анализ данных

Лекция 14

Композиции моделей

Евгений Соколов

esokolov@hse.ru

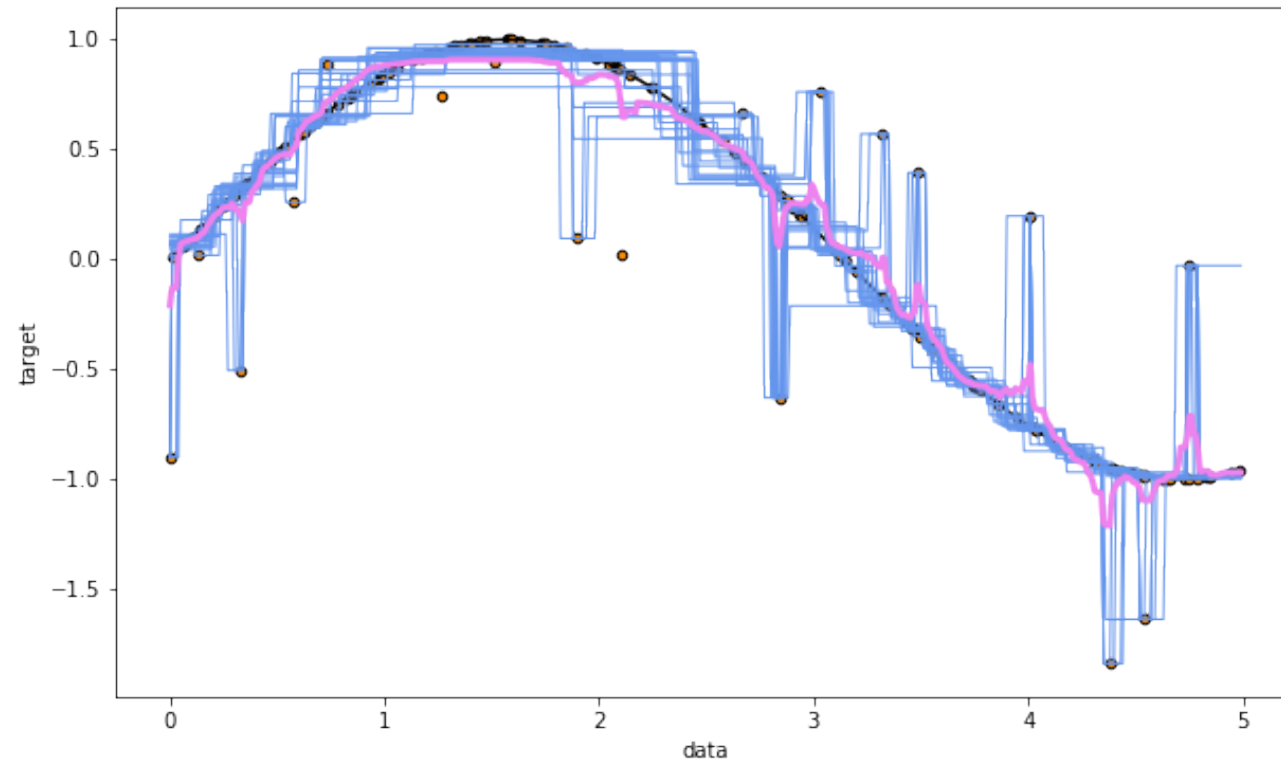
НИУ ВШЭ, 2020

Исправление ошибок моделей
и идея бустинга

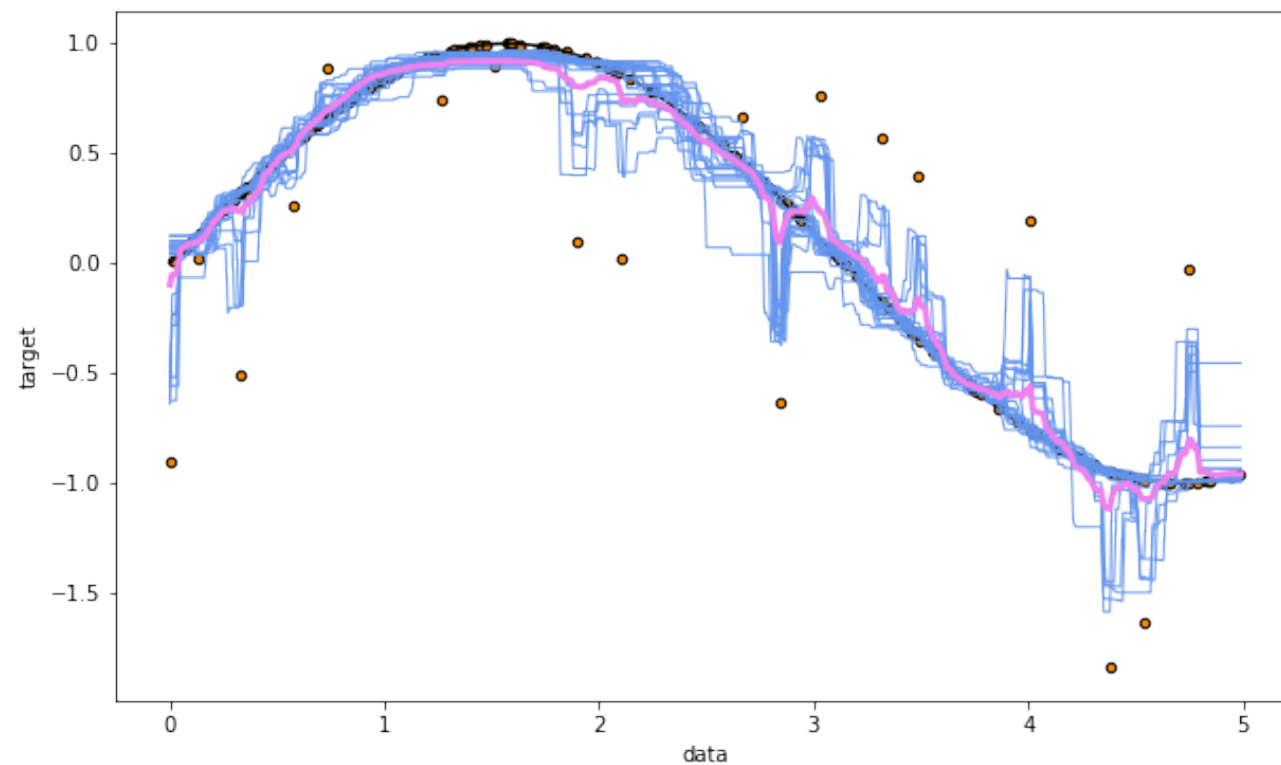
Бэггинг

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

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

Идея бустинга

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

- Обучение N -й модели:

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

Идея бустинга

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

- Обучение N -й модели:

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

Идея бустинга

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

- Обучение N -й модели:

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

- 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:

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

Бустинг для MSE

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

- Обучение N -й модели:

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

Бустинг для MSE

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

- Обучение N -й модели:

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

Бустинг для MSE

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

- Обучение N -й модели:

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

Бустинг для MSE

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

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

Первая итерация

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

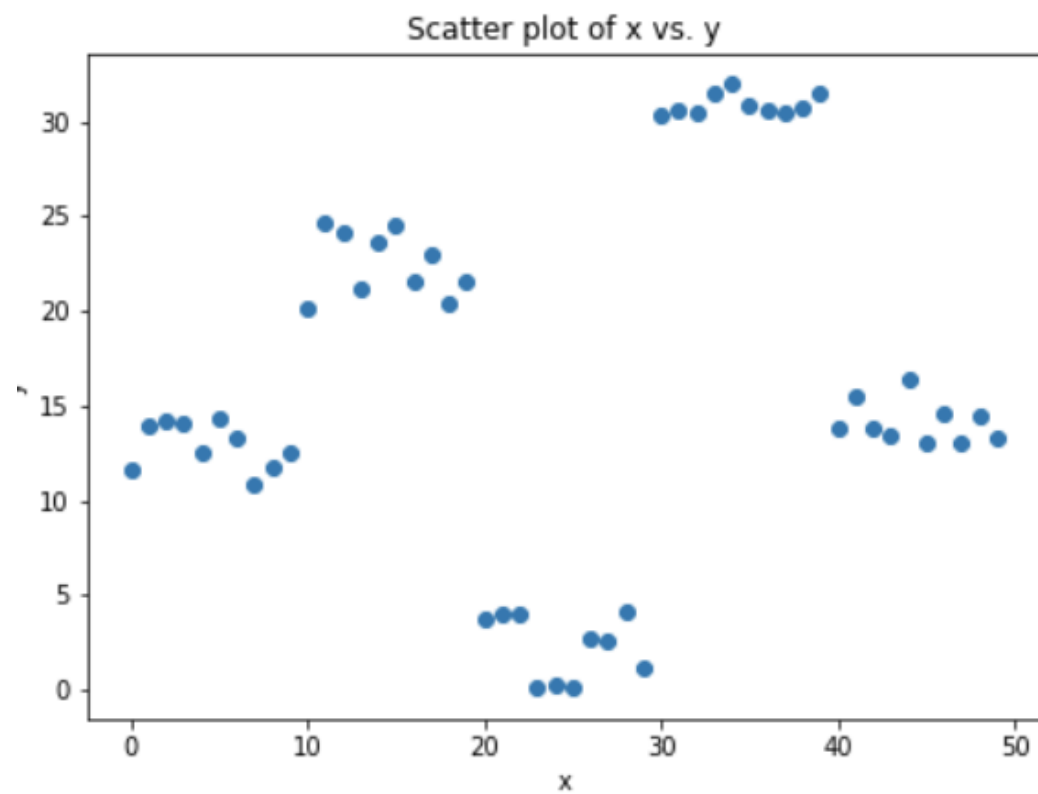
Вторая итерация

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

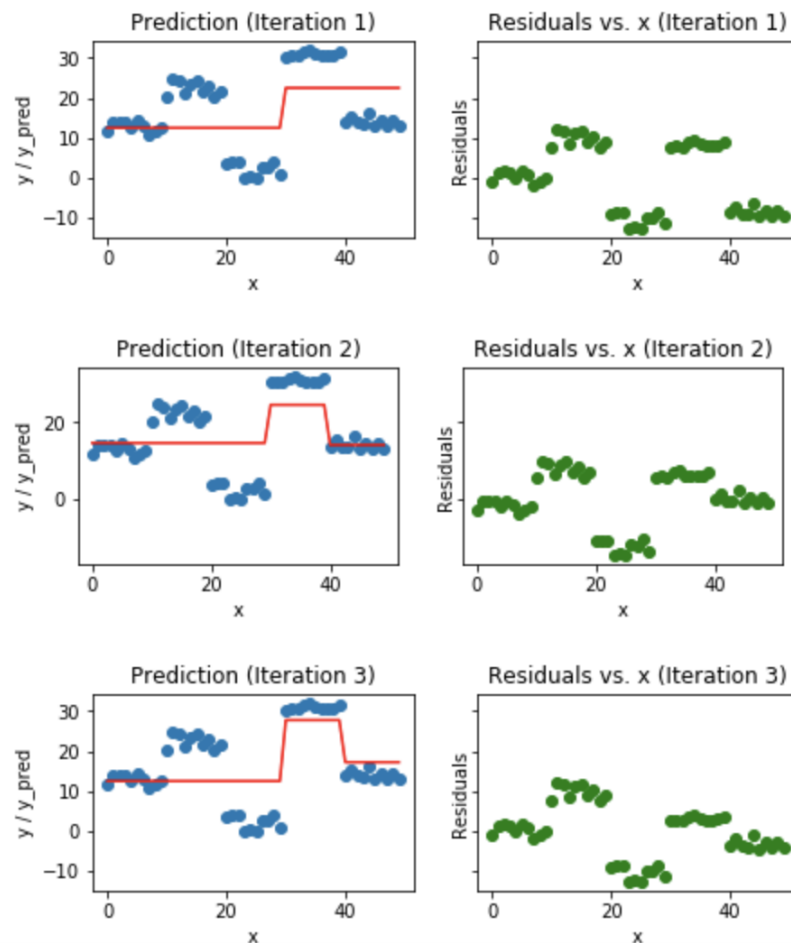
Третья итерация

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

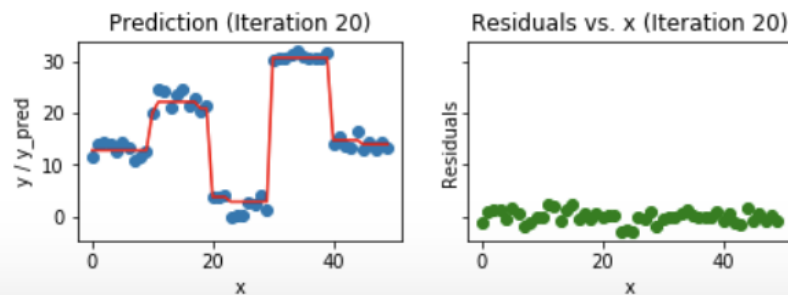
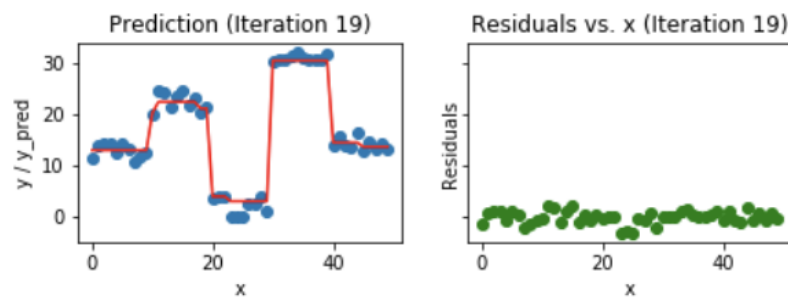
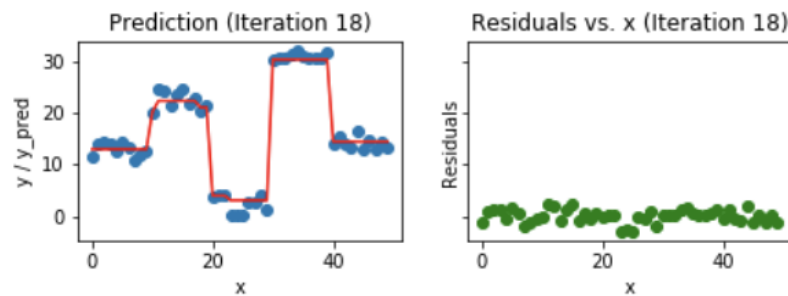
Визуализация



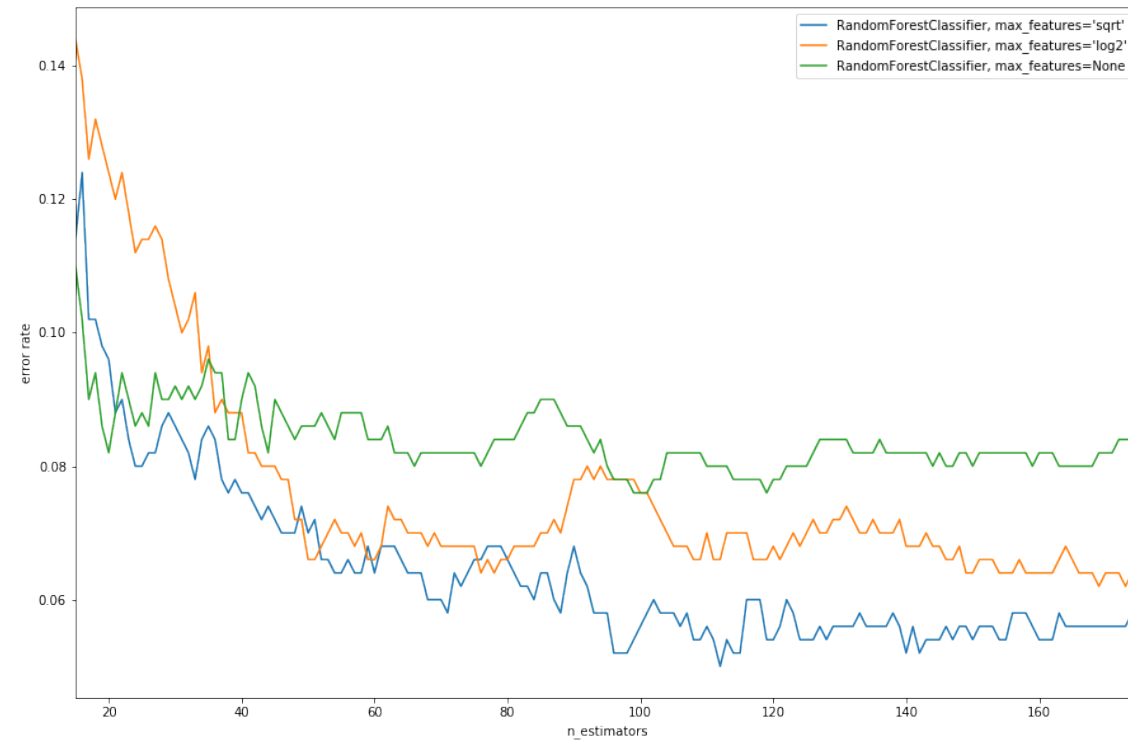
Визуализация



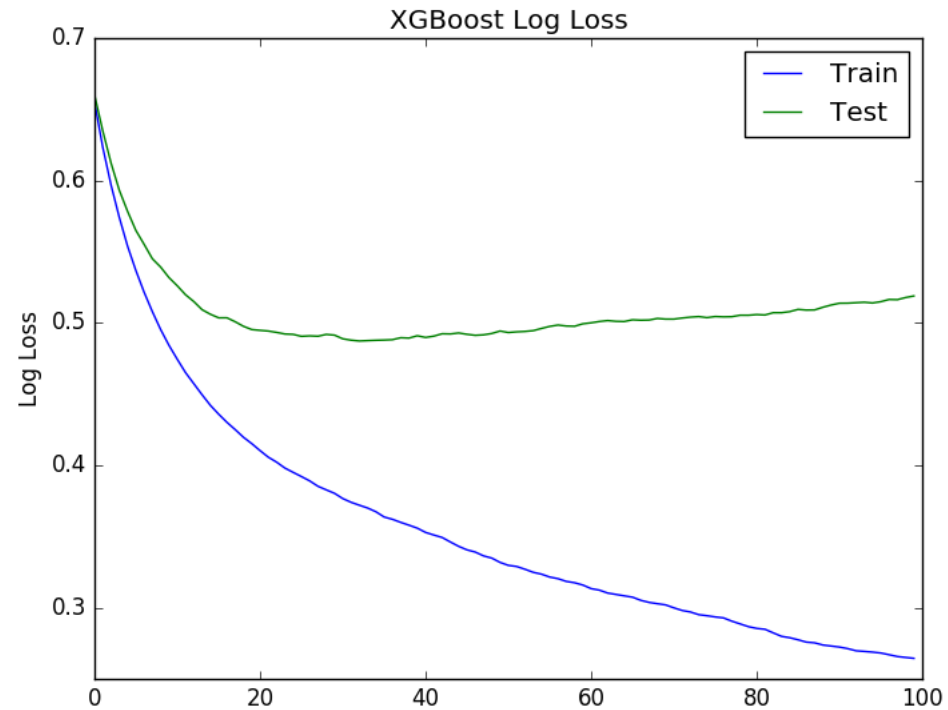
Визуализация



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:

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

Задача обучения базовой модели

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

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

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

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?

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

- 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

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

- 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

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

- $y_i = +1, \sum_{n=1}^{N-1} b_n(x_i) = -0.5 \rightarrow \text{надо } b_N(x_i) > 0.5$
- $y_i = +1, \sum_{n=1}^{N-1} b_n(x_i) = -100 \rightarrow \text{надо } b_N(x_i) > 100$
- 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?

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\log(\textcolor{red}{b}_N(\textcolor{red}{x}_i) + 1) - \log(\textcolor{blue}{y}_i - \textcolor{blue}{a}_{N-1}(\textcolor{blue}{x}_i) + 1))^2 \rightarrow \min_{b_N(x)}$$

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?

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

- Аргумент второго логарифма может оказаться отрицательным

MSLE

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\log(\textcolor{red}{b}_N(\textcolor{red}{x}_i) + 1) - \log(\textcolor{blue}{y}_i - \textcolor{blue}{a}_{N-1}(\textcolor{blue}{x}_i) + 1))^2 \rightarrow \min_{b_N(x)}$$

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:

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

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

Задача обучения базовой модели

- Обучение N -й модели:

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

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

Задача обучения базовой модели

- Обучение N -й модели:

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

- Посчитаем производную:

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

Задача обучения базовой модели

- Посчитаем производную:

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

- 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:

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

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

Градиентный бустинг

- Обучение N -й модели:

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

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

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

Градиентный бустинг для MSE

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

Градиентный бустинг для MSE

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

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

Градиентный бустинг для логистической функции потерь

$$\begin{aligned} s_i^{(N)} &= -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)} = \\ &= -\frac{\partial}{\partial z} \log(1 + \exp(-y_i z)) \Big|_{z=a_{N-1}(x_i)} = \\ &= \frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))} \end{aligned}$$

Градиентный бустинг для логистической функции потерь

$$\begin{aligned} s_i^{(N)} &= -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)} = \\ &= -\frac{\partial}{\partial z} \log(1 + \exp(-y_i z)) \Big|_{z=a_{N-1}(x_i)} = \\ &= \frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))} \end{aligned}$$

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

Градиентный бустинг для логистической функции потерь

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

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

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:

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

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

Глубина деревьев

- Градиентный бустинг уменьшает смещение базовых моделей
- Разброс может увеличиться
- 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

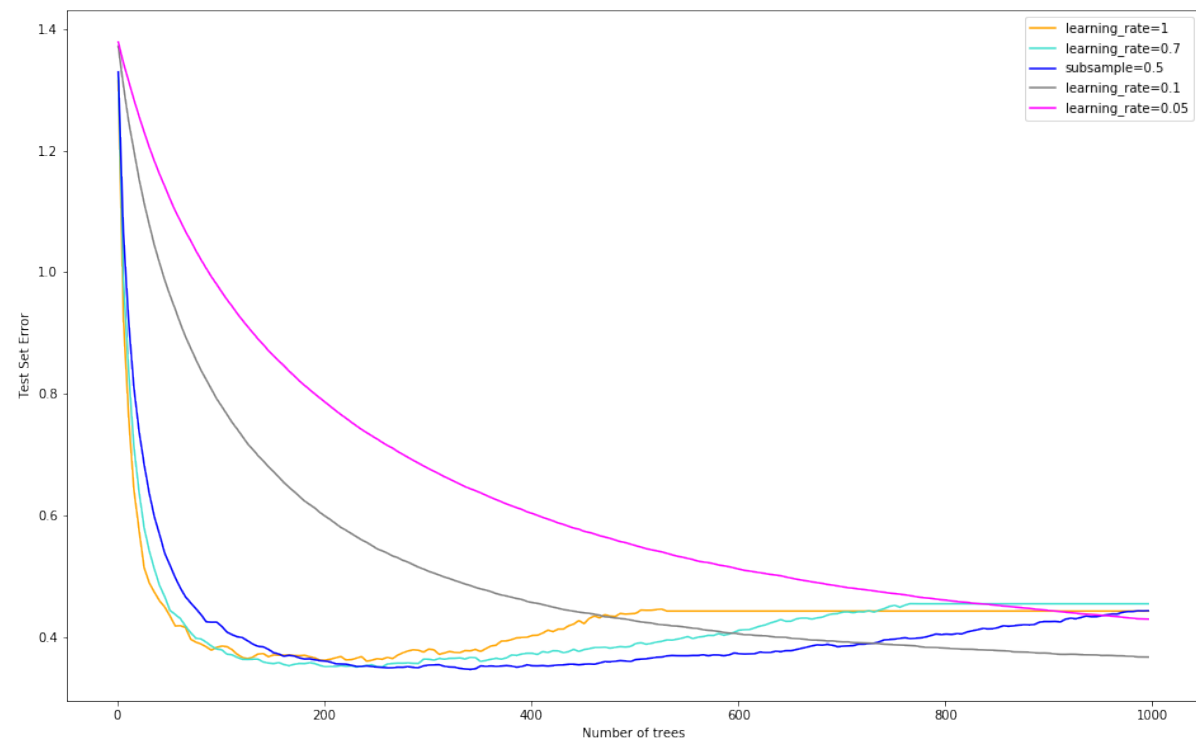
- Сдвиги показывают направление, в котором надо сдвинуть композицию на всех объектах обучающей выборки
- Базовые модели, как правило, очень простые
- Могут не справиться с приближением этого направления
- Выход: добавлять деревья в композицию с небольшим весом

Длина шага

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

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

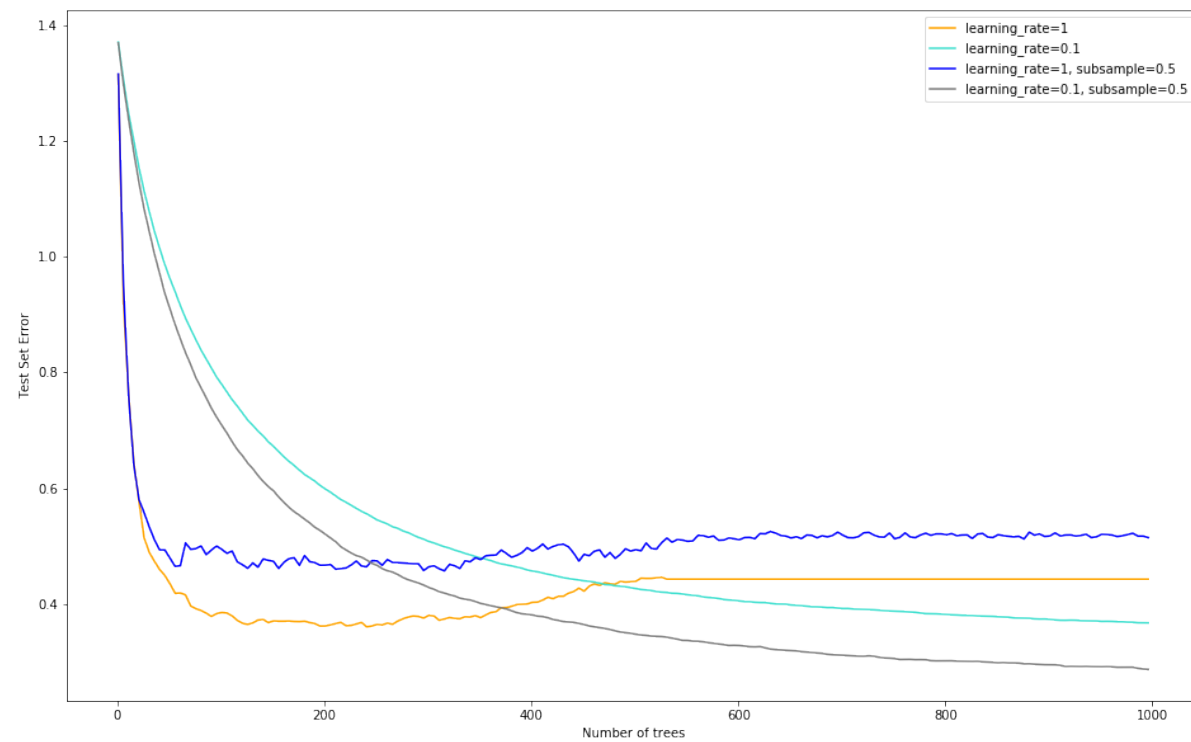
Длина шага



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.

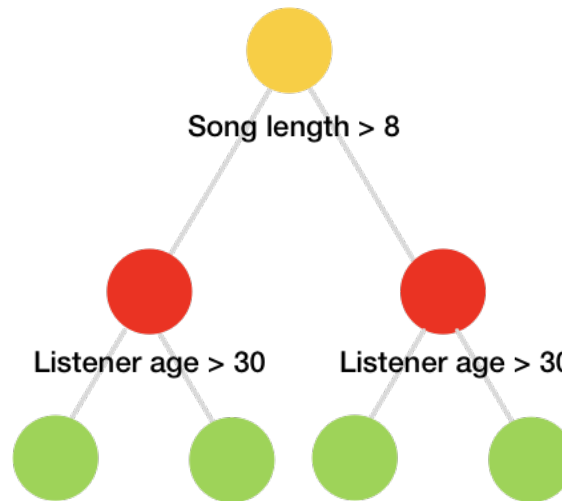
Резюме

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

Вариации бустинга

ODT

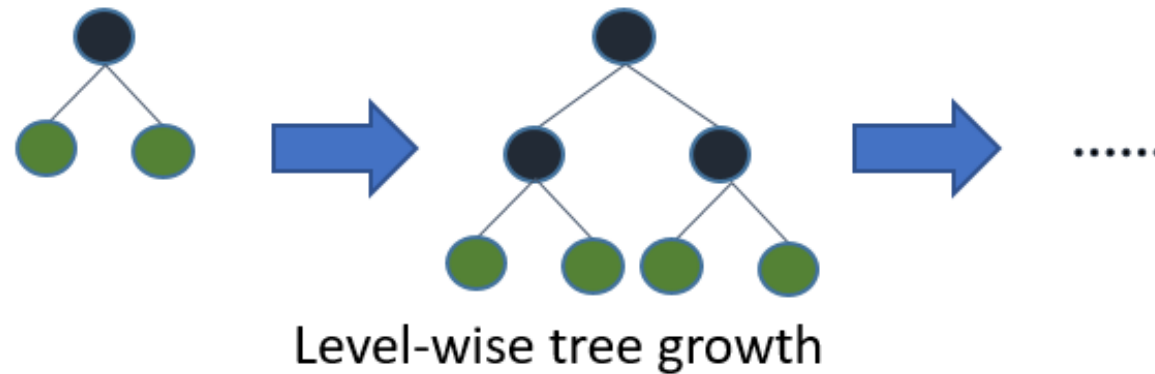
- Oblivious decision trees
- Ограничение: на одном уровне дерева используется один и тот же предикат



<https://catboost.ai/>

Способ построения дерева

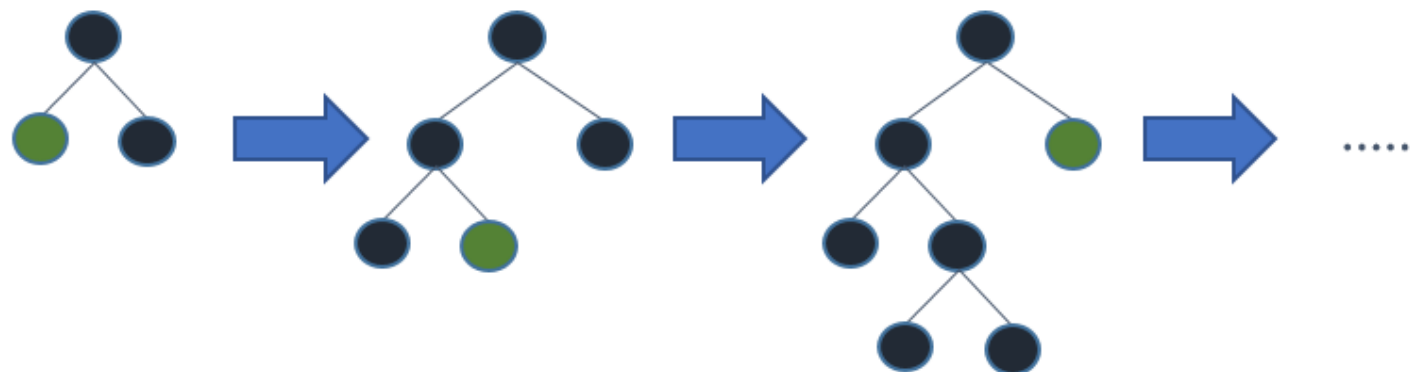
- Level-wise: дерево строится рекурсивно до тех пор, пока не достигнута максимальная глубина



<https://lightgbm.readthedocs.io/>

Способ построения дерева

- Level-wise: дерево строится рекурсивно до тех пор, пока не достигнута максимальная глубина
- Leaf-wise: среди текущих листьев выбирается тот, чьё разбиение сильнее всего уменьшает ошибку



Leaf-wise tree growth

Выбор лучшего порога для предиката

- $[x_j < t]$ — как выбрать t ?
- Вариант 1: перебрать все известные значения признака
- Вариант 2: построить гистограмму для признака и искать пороги среди границ на гистограмме
- Вариант 3: просемплировать объекты с близкими к нулю значениями производной

Регуляризация деревьев

- Базовая регуляризация: введение длины шага и семплирования признаков
- Штрафы за число листьев в дереве
- Штрафы за величину прогнозов в листьях дерева

Улучшенное обучение

- Мы обучаем деревья на сдвиги, ошибка измеряется с помощью MSE
- Когда дерево построено, можно подобрать оптимальные значения в листьях с точки зрения исходной функции потерь

Имплементации

- XGBoost
- LightGBM: leaf-wise growth, поиск порогов на основе производных
- CatBoost: ODT