Отчет по практическому заданию №3 «Композиции алгоритмов для решения задачи регрессии»

Суходоева Анна Евгеньевна
курс «Практикум на ЭВМ» ММП ВМК МГУ

Декабрь 2019

Описание работы

В данной работе требовалось реализовать методы случайных лес и градиентный бустинг и исследовать поведение алгоритмов и значение RMSE в зависимости от ряда параметров.

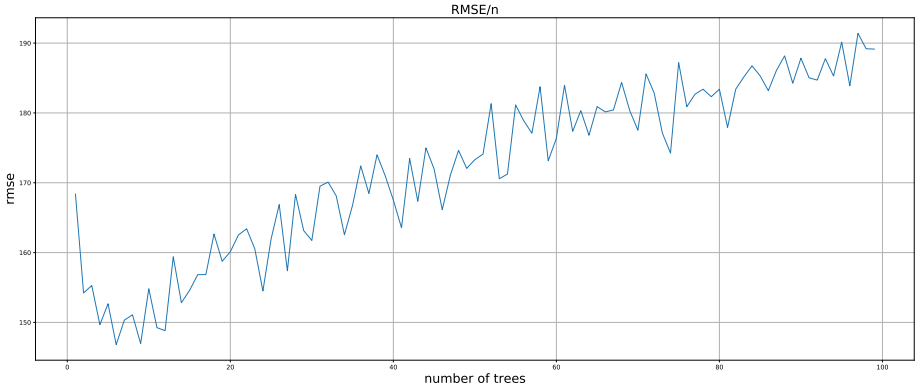
RandomForest

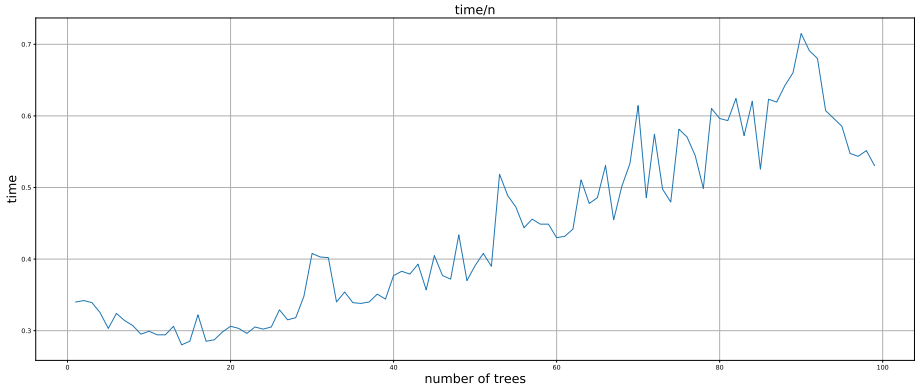
Описание алгоритма

1. Генерируется случайная подвыборка из объектов и признаков.
2. Строится решающее дерево. Обучается на данной подвыборке.
3. Итоговое предсказание берется как среднее значение ответов на всех построенных деревьях.

Эксперименты

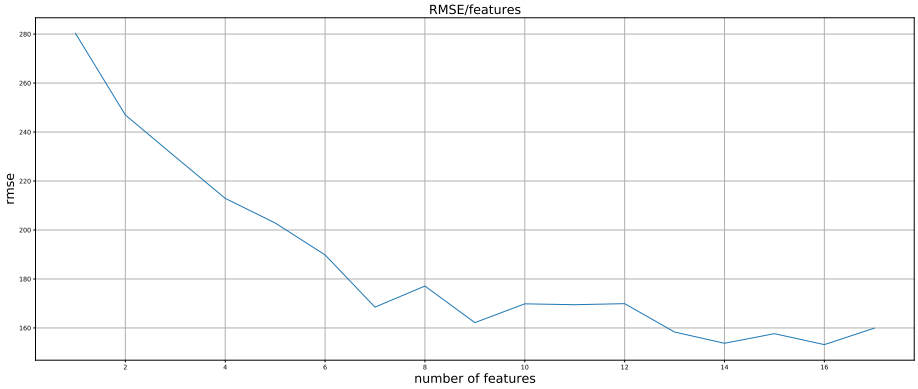
Количество деревьев

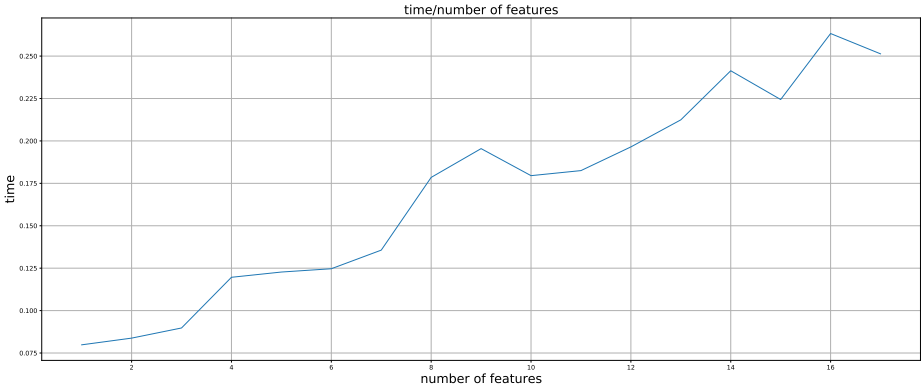




В данном эксперименте значение параметра nestimators перебиралось от 1 до 100, параметр maxdepth был принят равным 10, а остальные - по умолчанию. Как можно видеть из графика, увеличение количества деревьев ведет к увеличению ошибки, что связано с увеличением сложности модели до той степени, когда деревья уже не компенсируют ошибки друг друга, однако при небольшом увеличении их количества точность как раз увеличивается. При этом время работы растет, что связано с количеством вычислений, которое растет при увеличении количества деревьев.

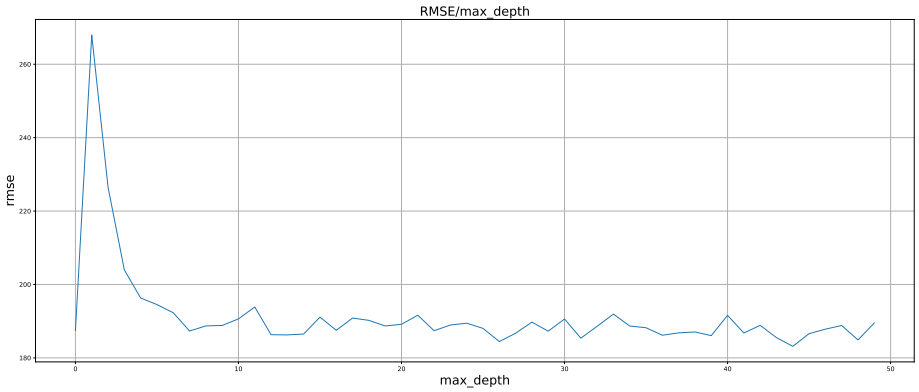
Размерность подвыборки признаков для одного дерева

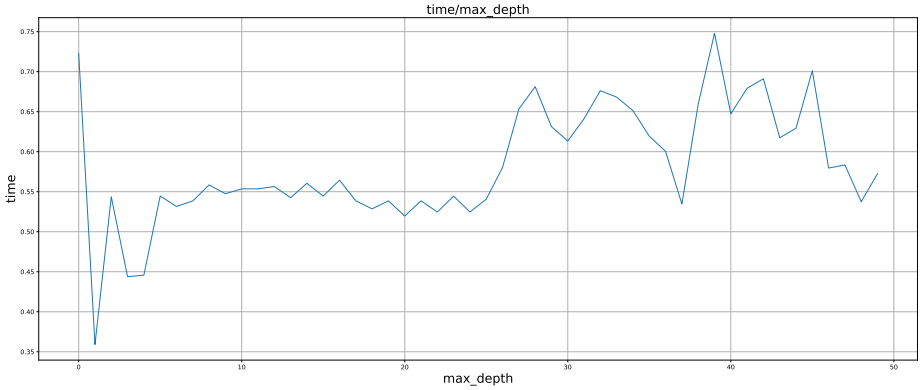




В данном эксперименте значение параметра featuresubsamplesize перебиралось от 1 до 18 - максимального количества признаков, остальные параметры были взяты по умолчанию. Увеличение размерности подвыборки признаков для одного дерева приводит к уменьшению значения ошибки, так как при стремлении данного параметра к размерности признакового пространства, модели становятся более полными и по сути приближаются к построению полного дерева. Время работы в данном случае растет линейно, так как количество признаков влияет на размеры дерева и количество вычислений.

Максимальная глубина дерева





В данном эксперименте значение параметра maxdepth перебиралось от 1 до 50, остальные параметры были взяты по умолчанию. На графике значение 0 по шкале абцисс соответствует параметру None. Поэтому можно утверждать, что отсутствие ограничения на максимальную глубину дерева ведет к увеличению ошибки, так как в данном случае каждый раз будет строиться полное дерево поиска с проходом всех признаков. Опять же заметим, что неограниченное по глубине дерево строится достаточно долго, а далее время работы возрастает с ростом глубины. Стоит заметить, что во всех экспериментах при их повторе с другими начальными параметрами вид зависимостей не менялся.

GradientBoosting

Описание алгоритма

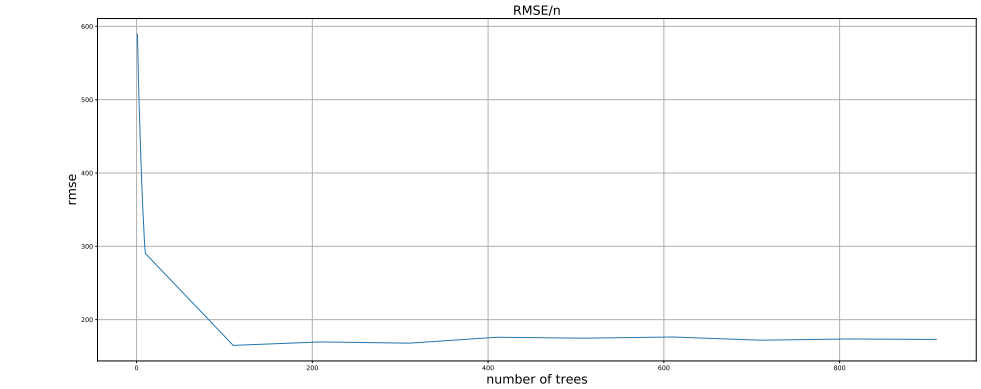
Строится ансамбль простых моделей(деревьев), таким образом, что каждая следующая модель уменьшает функцию потерь. Цель алгоритма - минимизировать функцию потерь(в данном случае MSE).

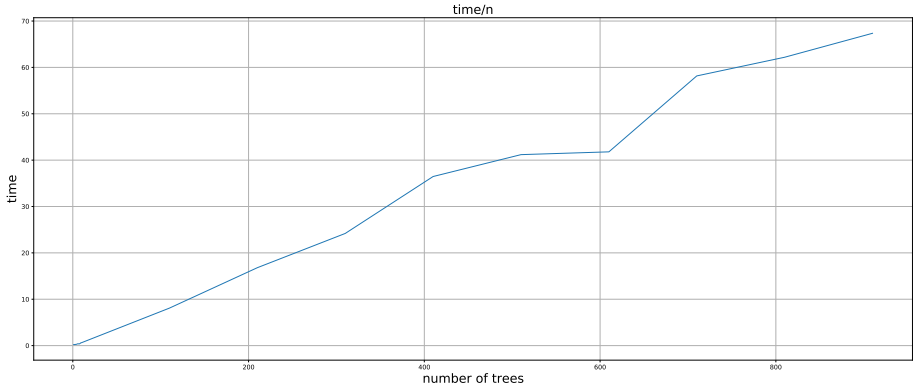
$$c_m = argmin \sum_{i=1}^{n} L(F_{m-1}(x) + c * f_m(x), y) \tag{1}$$

$$F_m(x) = F_{m-1}(x) + c_m * f_m(x) \tag{2}$$

Эксперименты

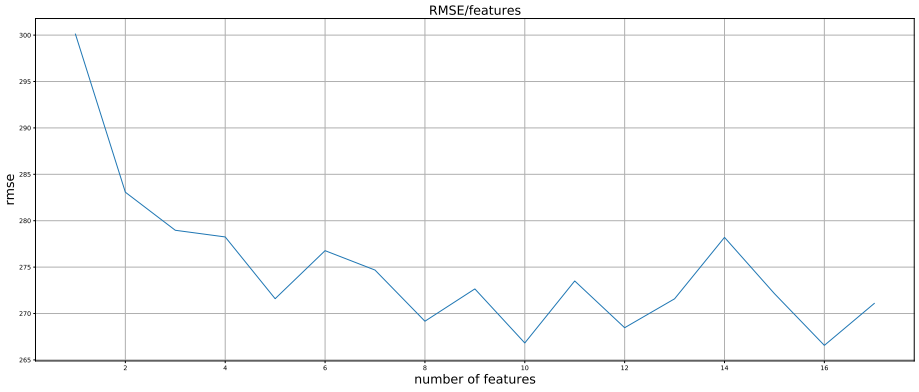
Количество деревьев

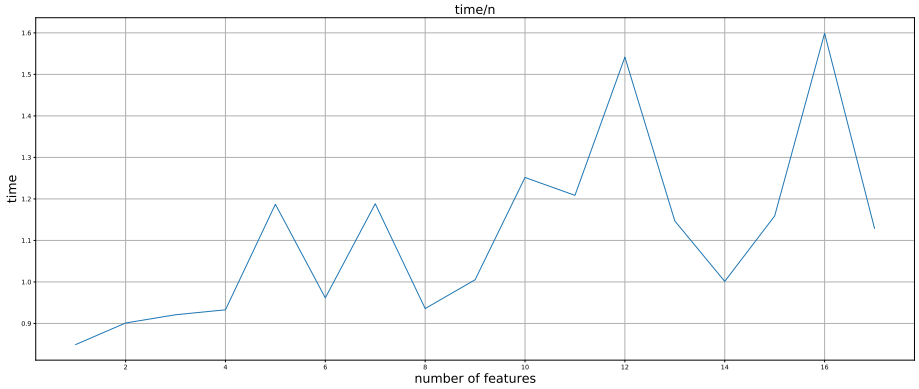




В данном эксперименте значение количество деревьев перебиралось от 1 до 10 с шагом 1 и от 10 до 100 с шагом 100, остальные параметры были взяты по умолчанию. При увеличении количества деревьев ошибка падает, так как увеличивается общая сложность моделей и ошибки деревьев компенсируются друг другом, в среднем давая приближенный к верному результат. Время растет линейно, так как количество деревьев в данном случае равно количеству итераций метода.

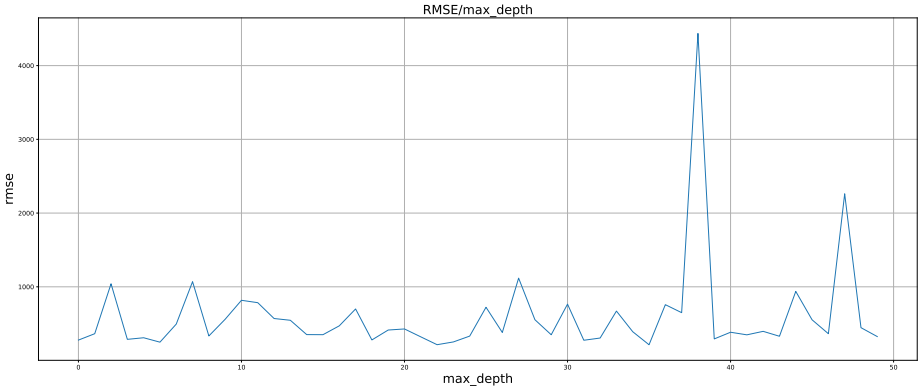
Размерность подвыборки признаков для одного дерева

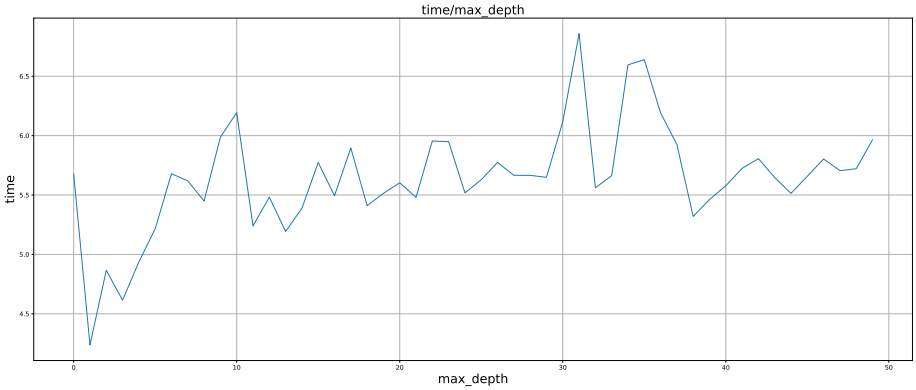




В данном эксперименте значение параметра featuresubsamplesize перебиралось от 1 до 18 - максимального количества признаков, количество деревьев было принято равным 100, а максимальная глубина - 10, остальные параметры были взяты по умолчанию. Ситуация аналогичная RandomForest - при увеличении количества признаков ошибка падает вследствие наиболее точных моделей. Время работы увеличивается, что опять же связано с размерами деревьев.

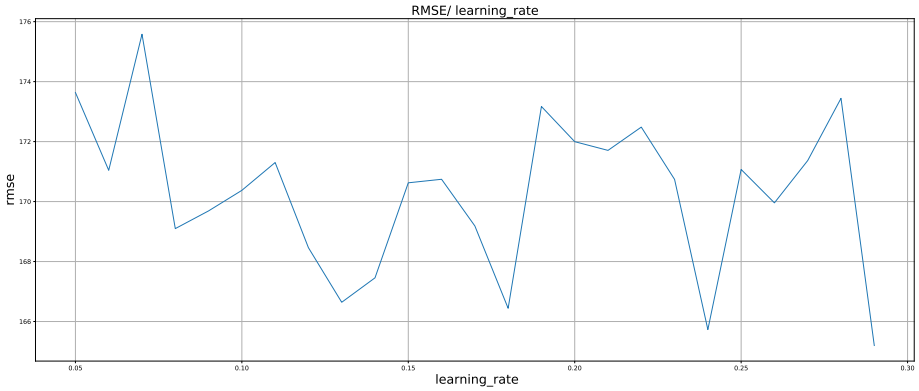
Максимальная глубина дерева

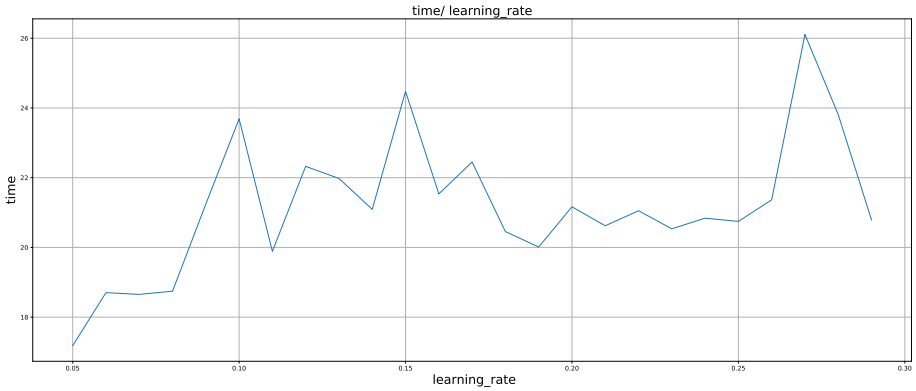




В данном эксперименте значение параметра maxdepth перебиралось от 1 до 50, количество деревьев было принято равным 100, остальные параметры были взяты по умолчанию. В данном случае опять же значение глубины = 0 соответствует значению None, так что опять можно утверждать, что наилучшие значения дает небольшая глубина дереа.

LearningRate





В данном эксперименте значение параметра LearningRate перебиралось от 0.05 до 0.3, количество деревьев было принято равным 300, остальные параметры были взяты по умолчанию. При достаточно малых значениях LearningRate ошибка уменьшается, однако при приближении к 0.15 и дальнейшем увеличении значения ошибка будет только увеличиваться. Время работы не зависит от увеличения значения, так как оно является просто множителем и никак не влияет на количество вычислений.

Вывод

После исследования поведения алгоритмов, можно сделать вывод, что в случае с количеством деревьев необходимо выбирать достаточно оптимальные параметры, чтобы при достаточно маленькой ошибке время работы также было невелико, практически при всех остальных параметрах зависимость от времени линейна, а качество модели наилучшее при небольших значениях параметров(размерности подвыборки признаков, максимальной глубине), значения LearningRate в случае использования алгоритма Градиентного бустинга не влияет на время работы, но самым оптимальным будет значение 0.13.