Отчет по заданию 3. Композиции алгоритмов для решения задачи регрессии

Зотов Арсений, 317 группа

25 декабря 2020 г.

Постановка задания

В данном задании мною была проделана следующая работа:

1. Написаны собственные реализации методов случайный лес и градиентный бустинг.

2. Проведены соответствующие эксперименты с датасетом houses.

Результаты экспериментов

Перед началом экспериментов была произведена минимальная предобработка имеющихся данных: столбец с датой был разделен на три отдельных столбца - год, месяц, день.

Данные были разделены на обучение и контроль в отношении 7:3 и переведены в numpy ndarray.

Эксперимент №1

В данном эксперименте было исследовано поведение алгоритма случайный лес. Случайный лес - это множество решающих деревьев. Все деревья строятся независимо по следующей схеме:

• Для построения каждого дерева выбираем feature_subsample_size случайных признаков

• Используем класс DecisionTreeRegressor из библиотеки scikit-learn.

Таким образом строим n_estimators деревьев. При предсказании их ответы усредняются.

Основные параметры:

n_estimators - число деревьев

feature_subsample_size - размерность подвыборки признаков для одного дерева

max_depth - максимальная глубина дерева

Изучена зависимость RMSE на отложенной выборке и время работы алгоритма в зависимости от следующих факторов:

- количество деревьев
- размерность подвыборки признаков для одного дерева
- максимальная глубина дерева

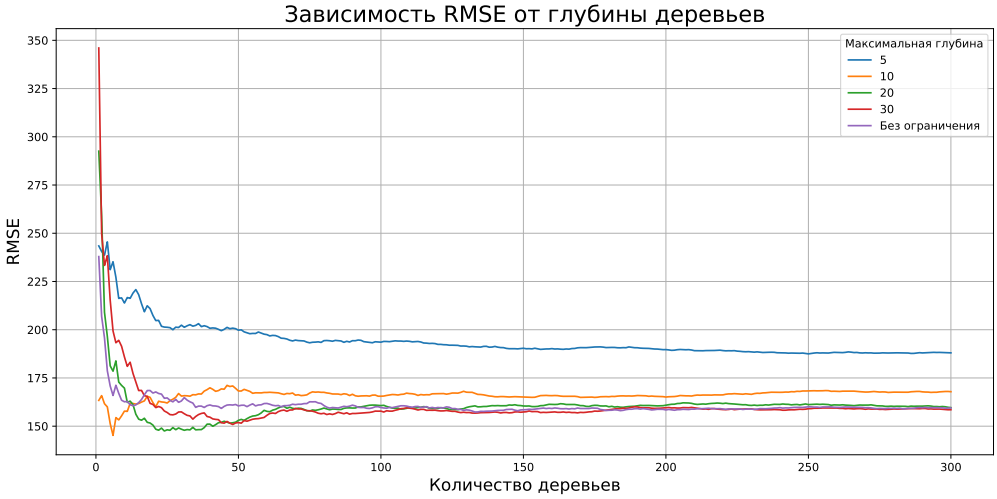

Рис. 1: Поведение случайного леса при различной максимальной глубине.

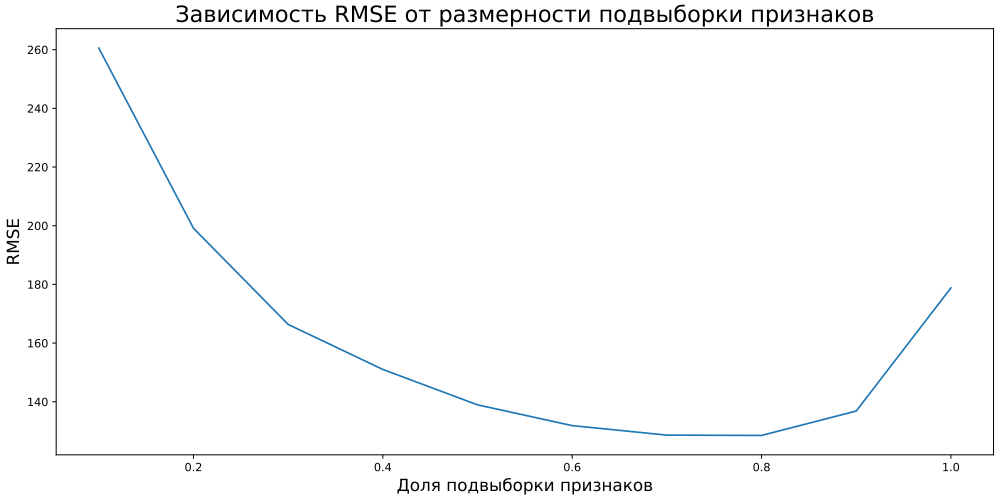

Рис. 2: Поведение случайного леса при различной размерности подвыборки признаков для одного дерева.

На рисунке 1 представлены результаты экспериментов для различной максимальной глубины. Видим, что увеличение глубины и количества деревьев приводит к улучшению качества. С увеличением количества деревьев происходит "сглаживание"графика, он принимает константную величину, зависящую от глубины деревьев.

Начиная с максимальной глубины равной 20 итоговая ошибка перестает заметно уменьшаться.

На рисунке 2 представлены результаты экспериментов для различной размерности подвыборки признаков для одного дерева. Количество деревье взято равным 300, а максимальная глубина - 20. График представляет собой выпуклую параболу. Видим, что для взятых параметров оптимальная величина доли размерности подвыборки признаков от общего числа признаков равна примерно 0.8.

Эксперимент №2

В данном эксперименте было исследовано поведение алгоритма градиентный бустинг. Градиентный бустинг - это множество решающих деревьев, построенных не независимо, а последовательно. i-ое дерево строится по следующей схеме:

• Выбираем feature_subsample_size случайных признаков

• Используя класс DecisionTreeRegressor из библиотеки scikit-learn находим $f_i(x)$. В роли целевых значений передаем $(y - F_{i-1}(X))$, где X - обучающая выборка, y - вектор действительных целевых значений на обучающей выборке, $F_{i-1} = \gamma * (c_1 * f_1 + ... + c_{i-1} * f_{i-1})$, γ - коэффициент темпа обучения.

• Используя minimize_scalar из библиотеки scikit-learn находим оптимальное значение c_i, минимизирующее функцию потерь $L = \sum_{j=1}^{n}(y_j - G_i(x_j))^2$, где $G_i = F_{i-1} + c_i * f_i$.

Таким образом строим n_estimators деревьев. $F_{n\_estimators}$ - итоговый алгоритм.

Основные параметры:

n_estimators - число деревьев

feature_subsample_size - размерность подвыборки признаков для одного дерева

max_depth - максимальная глубина дерева

learning_rate - коэффициент темпа обучения

Изучена зависимость RMSE на отложенной выборке и время работы алгоритма в зависимости от следующих факторов:

• количество деревьев

• размерность подвыборки признаков для одного дерева

• максимальная глубина дерева

• выбранный learning_rate

На рисунке 3 представлены результаты экспериментов для различных максимальной глубины и learning_rate. Видим, что увеличение глубины, количества деревьев и темпа обучения приводит к переобучению. При темпе обучения равном 1 заметно общая тенденция графиков: сначала ошибка достигает своего минимума, затем она возрастает и устремляется к константе. Данная тенденция объясняется переобучением градиентного бустинга. Лучшее качество достигается при темпе обучения равному 0.25 и небольшой глубине деревьев равной 3-5. При этом худшее качество показывает алгоритм без ограничения на максимальную глубину.

На рисунке 4 представлены результаты экспериментов для различных максимальной глубины и размерности подвыборки признаков для одного дерева. Видим,

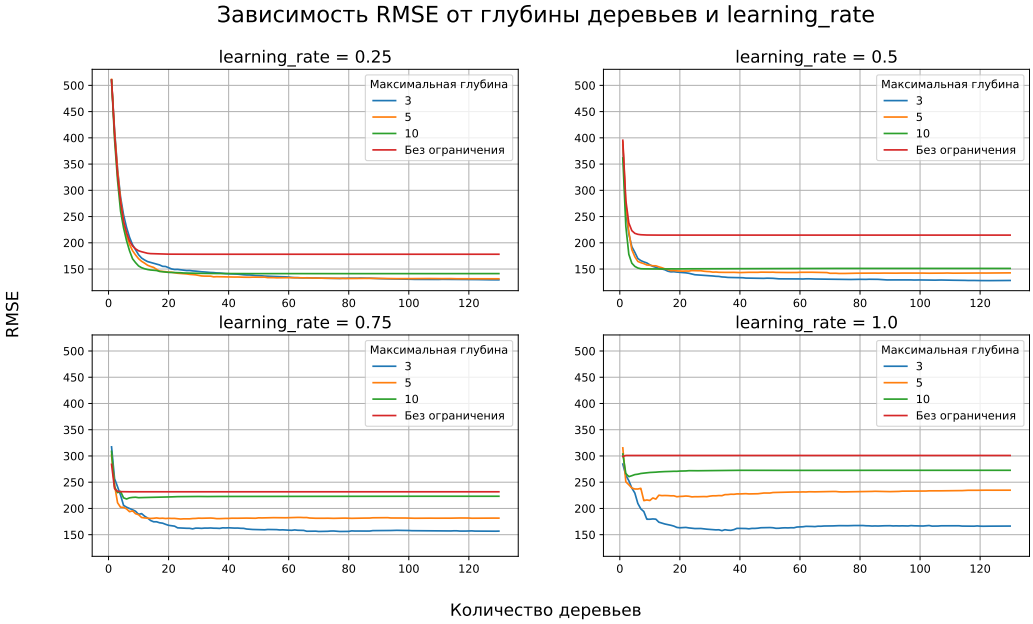

Рис. 3: Поведение градиентного бустинга при различных максимальной глубине и learning_rate.

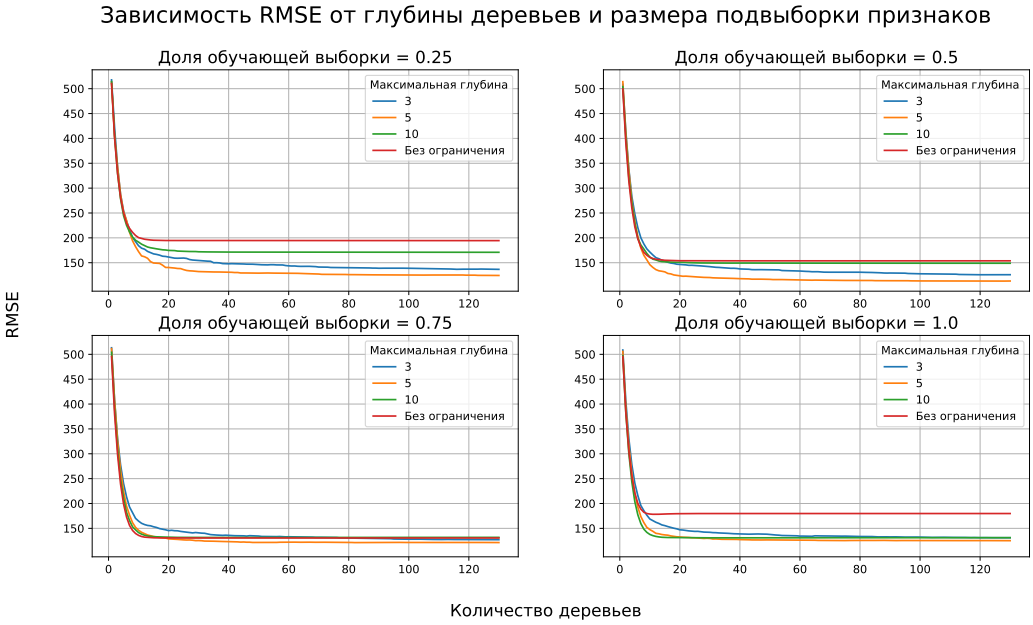

Рис. 4: Поведение градиентного бустинга при различных максимальной глубине и размерности подвыборки признаков для одного дерева.

что оптимальная величина доли размерности подвыборки признаков от общего числа признаков равна примерно 0.75. При этом наблюдаются те же тенденции переобучения при большом количестве деревьев.

На рисунке 5 представлены результаты экспериментов для различной размерности подвыборки признаков для одного дерева при количестве деревьев равном 150 и максимальной глубине - 5. График более ломанный, чем аналогичный для случай-

ного дерева, что усложняет выбор оптимальной размерности.

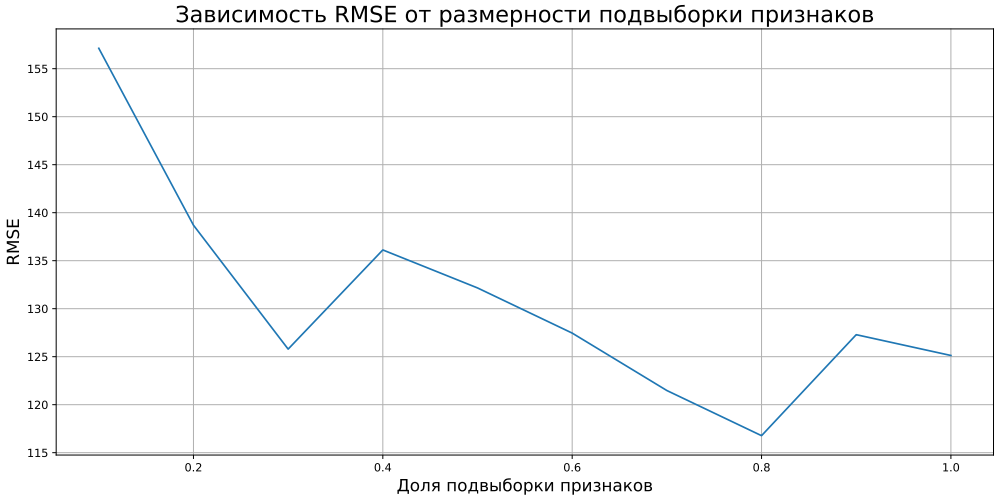

Рис. 5: Поведение градиентного бустинга при различной размерности подвыборки признаков для одного дерева.

Полученное качество при градиентном бустинге заметно превзошло качество при случайном лесе.

Выводы

По итогам задания были изучены алгоритмы случайный лес и градиентный бустинг.

Результаты экспериментов продемонстрировали, что при правильном подборе гиперпараметров с помощью рассмотренных методов можно добиться лучшего качества.