# Отчет по исследовательской работе на тему "Композиции алгоритмов для решения задачи регрессии"

Висков Василий, группа 317

Декабрь 2020

## Постановка задачи

Целью исследования является анализ зависимости значения фиксированной функции потерь на отложенной выборке и времени работы композиционных алгоритмов при решении задачи регрессии в зависимости от параметров моделей.

## Список экспериментов

В рамках этой работы будет рассмотрен:

- градиентный бустинг над решающими деревьями;
- случайный лес;

Функцией потерь в исследовании будет выступать Root Mean Squared Error, задаваемая формулой:

$$RMSE = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}}$$

Здесь $y_i$ - истинное вещественное значение, поставленное в соответствие объекту $i$, а $\hat{y}$ - предсказание модели.

Исследование будет проводиться на наборе данных, взятых предположительно с конкурса Boston Houses 2020 [1]. В данных содержались категориальный признак, описывающий номер комнаты, и порядковый признак, описывающий дату. Первый из перечисленных преобразуем с помощью dummy-кодирования, а второй переведем в количество прошедших от заданной даты дней, что будет являться монотонным преобразованиям в данных, к которым деревья решений являются инвариантными. Датасет, состоящий из 17280 прецедентов, будет разбит на обучающую и тестовую выборки в соотношении 7:3.

В качестве параметров по умолчанию, один из которых будет варьироваться, возьмем следующие значения:

- RF:
    1. n_estimators: 1000;
    2. feature_subsample_size: 0.7;

3. max_depth: 22;

- GBDT:

1. n_estimators: 1000;
2. feature_subsample_size: 0.7;
3. max_depth: 6;
4. learning_rate: 0.1;

## Эксперимент 1

Цель эксперимента - изучить зависимость RMSE на отложенной выборке и время работы алгоритма случайного леса в зависимости от следующих факторов:

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

### Дизайн эксперимента

Для эксперимента выбраны следующие значения параметров:

- n_estimators $\in$ [10, 100, 300, 500, 800, 1000, 2500];
- feature_subsample_size $\in$ [0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 1];
- max_depth $\in$ [3, 5, 9, 15, 22, 27, 32, 40, 50, None];

Специфика реализации позволяет находить оптимальное с точки зрения функции потерь на отложенной выборке количество базовых моделей для фиксированных остальных параметров - найдем это значение в рамках исследования влияния количества деревьев на качество.

### Результаты эксперимента

Как можно видеть, при увеличении количества деревьев качество увеличивается, но упирается в некоторый предел, так как перестает выполняться условие независимости построенных базовых моделей из-за ограниченности датасета. Время обучения растет при увеличении количества итераций, что очевидно (таб. 1). На основании графика 1 и подсчета найдено, что для заданных по умолчанию параметров наилучшего результата можно достигнуть при использовании 1677 базовых моделей. О качестве тут нет смысла говорить и сравнивать его, так как от запуска к запуску в силу использования операции бутстрапирования оно может варьироваться.

| n_estimators | RMSE | duration (sec) |
|---|---|---|
| 10 | 124.8002 | 0.547 |
| 100 | 119.2506 | 3.917 |
| 300 | 118.9167 | 12.211 |
| 500 | 118.1341 | 19.736 |
| 800 | 118.0323 | 32.271 |
| 1000 | 117.8181 | 39.685 |
| 2500 | 118.1271 | 98.544 |

Таблица 1: Зависимость RMSE и времени работы от количества деревьев (RF)

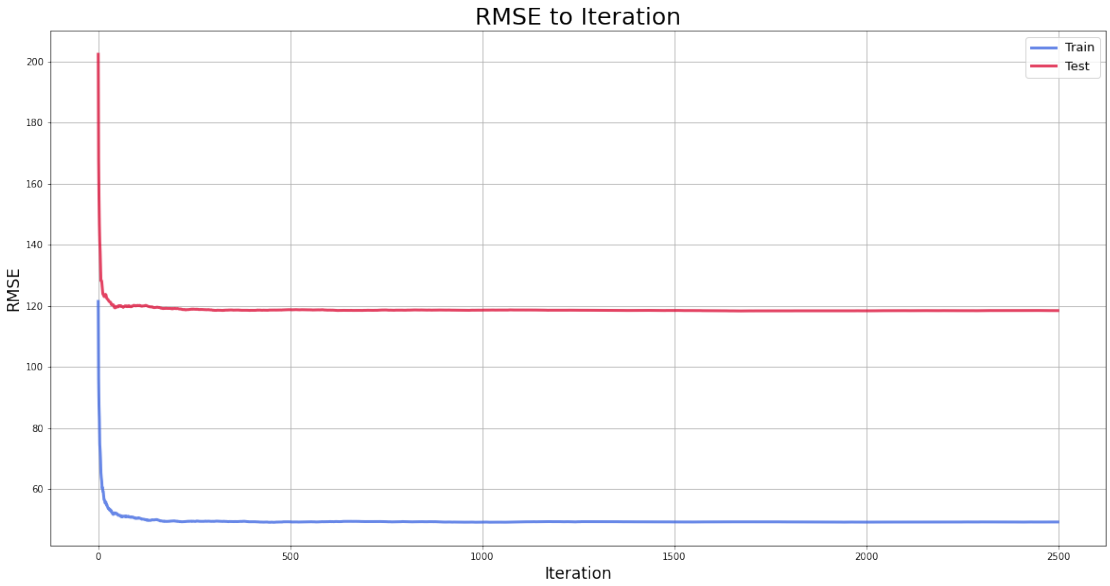


Рис. 1: RMSE в зависимости от итерации (RF)

Аналогичная ситуация обстоит и с параметром, отвечающим за размер подвыборки признаков: по мере увеличения размера признакового подпространства растет и качество, но достигает своего предела при определенном значении. В рамках эксперимента оно равно 0.9. При этом время работы алгоритма растет по мере увеличения значения параметра, но резко падает при условии, что рассматривается все пространство признаков каждый раз, что связано, видимо, с накладными расходами на сэмплирование признаков в каждой вершине (таб. 2).

| feature_subsample_size | RMSE | duration (sec) |
|---|---|---|
| 0.1 | 138.6692 | 18.465 |
| 0.3 | 122.7939 | 24.906 |
| 0.5 | 118.9703 | 32.127 |
| 0.7 | 118.7364 | 38.966 |
| 0.8 | 118.2980 | 43.305 |
| 0.9 | 118.1342 | 46.988 |
| 1 | 154.0329 | 15.142 |

Таблица 2: Зависимость RMSE и времени работы от размерности подвыборки признаков (RF)

С параметром max_depth та же картина: по мере увеличения максимальной глубины качество растет, но достигает наибольшего при определенном значении, в нашем случае - при глубине, равной 32. Время работы растет по мере увеличения значения параметра (таб. 3).

| max_depth | RMSE | duration (sec) |
|---|---|---|
| 3 | 197.2342 | 12.114 |
| 5 | 160.297 | 15.795 |
| 9 | 128.6512 | 21.223 |
| 15 | 118.837 | 32.874 |
| 22 | 118.4231 | 38.725 |
| 27 | 118.3923 | 39.691 |
| 32 | 117.8966 | 40.679 |
| 40 | 118.4197 | 42.233 |
| 50 | 118.2913 | 43.173 |
| None | 118.0025 | 43.905 |

Таблица 3: Зависимость RMSE и времени работы от максимальной глубины деревьев (RF)

## Эксперимент 2

Цель эксперимента - изучить зависимость RMSE на отложенной выборке и время работы алгоритма градиентного бустинга на деревьях в зависимости от следующих факторов:

- количество деревьев;
- размерность подвыборки признаков для одного дерева;
- максимальная глубина дерева;
- темп обучения;

### Дизайн эксперимента

Для эксперимента выбраны следующие значения параметров:

- n_estimators $\in$ [10, 100, 300, 500, 800, 1000, 2500];
- feature_subsample_size $\in$ [0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 1];
- max_depth $\in$ [3, 5, 9, 15, 22, 27, 32, 40, 50, None];
- learning_rate $\in$ [0.001, 0.01, 0.05, 0.1, 0.5, 1];

Специфика реализации, как и в случае со случайным лесом, позволяет находить оптимальное с точки зрения функции потерь на отложенной выборке количество базовых моделей для фиксированных остальных параметров - найдем это значение в рамках исследования влияния количества деревьев на качество.

### Результаты эксперимента

Как можно видеть, при увеличении количества деревьев качество увеличивается, но упирается в некоторый предел: модель переобучается под обучающую выборку. Время обучения растет при увеличении количества итераций, что очевидно (таб. 4). На основании графика 2 и подсчета найдено, что для заданных по умолчанию параметров наилучший результат достигается при использовании 626 базовых моделей (RMSE: 108.1772).

| n_estimators | RMSE | duration (sec) |
|---|---|---|
| 10 | 265.5242 | 0.607 |
| 100 | 112.8786 | 6.327 |
| 300 | 112.0387 | 17.743 |
| 500 | 110.6598 | 28.611 |
| 800 | 112.3461 | 43.275 |
| 1000 | 111.4958 | 50.626 |
| 2500 | 111.6600 | 123.285 |

Таблица 4: Зависимость RMSE и времени работы от количества деревьев (GB)

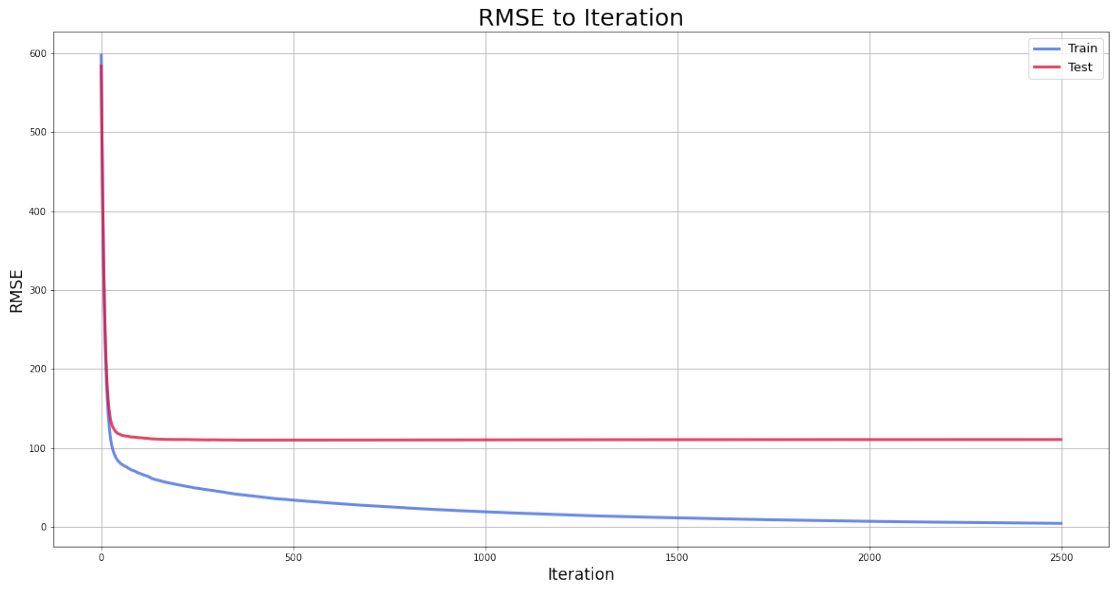


Рис. 2: RMSE в зависимости от итерации (GB)

Аналогичная ситуация обстоит и с параметром, отвечающим за размер подвыборки признаков: по мере увеличения размера признакового подпространства растет и качество, но достигает своего предела при определенном значении. В рамках эксперимента оно равняется 0.9. При этом время работы алгоритма растет по мере увеличения значения параметра, но резко падает при условии, что рассматривается все пространство признаков каждый раз, что связано, видимо, с накладными расходами на сэмплирование признаков в каждой вершине (таб. 5).

| feature_subsample_size | RMSE | duration (sec) |
|---|---|---|
| 0.1 | 122.6883 | 14.745 |
| 0.3 | 108.8641 | 24.897 |
| 0.5 | 110.7159 | 35.600 |
| 0.7 | 109.8719 | 47.492 |
| 0.8 | 110.8404 | 52.486 |
| 0.9 | 110.6577 | 58.807 |
| 1 | 128.2873 | 9.730 |

Таблица 5: Зависимость RMSE и времени работы от размерности подвыборки признаков (GB)

С параметром max_depth та же картина: по мере увеличения максимальной глубины качество растет, но достигает наибольшего при определенном значении, в нашем случае - при глубине, равной 5. Заметим, что для градиентного бустинга оказалось, что лучше брать неглубокие деревья, в отличие от деревьев из случайного леса (в предыдущем эксперименте наилучшее качество было достигнуто при значении максимальной глубины, равном 32). Время работы растет по мере увеличения значения параметра (таб. 6).

| max_depth | RMSE | duration (sec) |
|---|---|---|
| 3 | 110.7048 | 28.193 |
| 5 | 108.8505 | 40.940 |
| 9 | 115.7349 | 68.817 |
| 15 | 121.3925 | 106.287 |
| 22 | 123.2503 | 125.441 |
| 27 | 123.0779 | 125.589 |
| 32 | 123.7637 | 126.333 |
| 40 | 122.4359 | 124.326 |
| 50 | 123.1098 | 119.900 |
| None | 123.5908 | 122.479 |

Таблица 6: Зависимость RMSE и времени работы от максимальной глубины деревьев (GB)

Темп обучения оказывает такое же влияние на качество по мере его увеличения, наилучшее значение RMSE достигнуто при значении 0.05. При этом на время работы алгоритма оно влияния не оказывает, что ожидаемо: принципиального значения величина этого параметра не имеет в рамках измерения времени выполнения операции умножения выхода базовых моделей на некоторое вещественное число (таб. 7).

| learning_rate | RMSE | duration (sec) |
|---|---|---|
| 0.001 | 275.9895 | 46.942 |
| 0.01 | 111.6654 | 49.603 |
| 0.05 | 109.4831 | 50.18 |
| 0.1 | 111.1591 | 50.582 |
| 0.5 | 129.2203 | 49.699 |
| 1 | 177.6525 | 48.862 |

Таблица 7: Зависимость RMSE и времени работы от темпа обучения (GB)

# Выводы

- параметры модели можно сравнить с веществами, которые при малых дозах используются в лечебным целях, а при больших становятся ядом, - нужно обязательно включать в пайплайн работы над проектом процесс их подбора;

# Список литературы

[1] https://drive.google.com/file/d/1IQOuOojfp7UHc9J1PbrsQB2S9SQ_BWN3/view;