Третий бейзлайн в контесте по временным рядам

курс ML2, OzonMasters

Попов Артём, Камиль Сафин

Постановка задачи

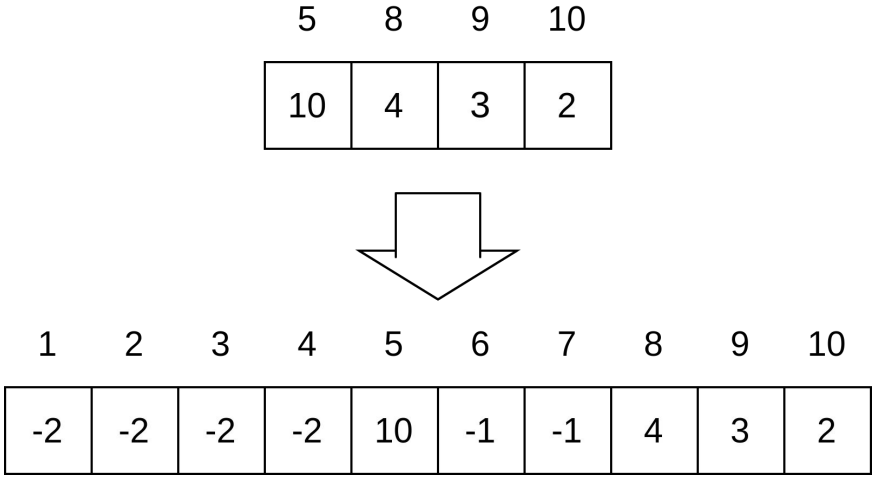
Даны временные ряды (статистика, связанная с количеством пользователей в компьютерной игре).

Необходимо для каждого ряда предсказать три следующих значений.

Критерий качества: MASE - MAE с весами, вес вычисляется как MAE по обучающей выборке при предсказании последним элементом ряда

Этап 1: заполнение пропущенных значений

- “-2” для пропущенных значений в начале ряда (моменты времени, когда игра ещё не продавалась)
- “-1” для пропущенных значений внутри ряда (моменты времени, когда игра снималась с продажи?)



Этап 2: сэмплирование объектов для обучения

Из каждого ряда длины > 4 будем брать сэмплы.

Один сэмпл - срез ряда [0; 3 + end], end выбирается случайно равномерно.

Количество сэмплов из одного ряда = max(1, int(k * length * weight))

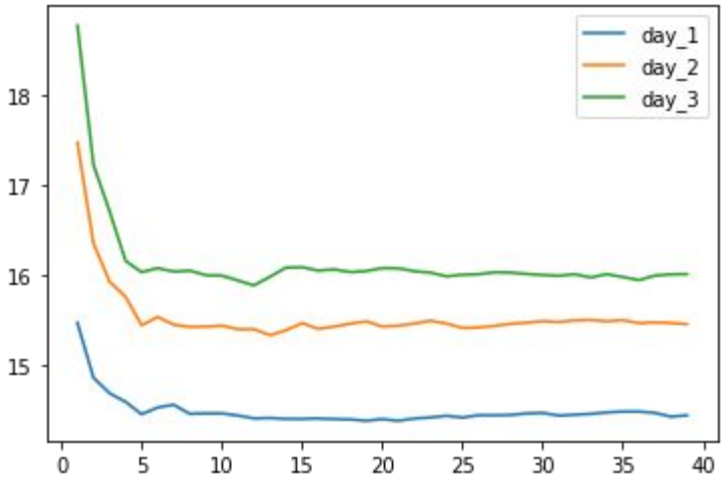
- k - гиперпараметр, для безйлайна выбрали k = 2
- length - количество элементов в ряду до заполнения пропусков
- weight - вес ряда из функционала

Этап 3: подбор оптимального окна

Хотим по значениям ряда $y_{t-s}, \ldots, y_{t-1}$ предсказывать y_t, y_{t+1}, y_{t+2}.

Для каждой задачи будем использовать свою модель градиентного бустинга.

Для каждой задачи подберём свой оптимальный размер окна (s).



Для каждого значения s обучаем бустинг и смотрим качество функционала MASE на валидации.

Валидация — последние значения в каждом ряду.

Этап 4: подбор гиперпараметров модели

```
mdl = lgbm.LGBMRegressor()

gridParams = {
    'learning_rate': [0.05, 0.005],
    'n_estimators': [20, 80, 100],
    'num_leaves': [150, 200, 250],
    'boosting_type' : ['gbdt'],
    'objective' : ['regression'],
    'random_state' : [501]
}
```

Валидация - стандартное k-fold разбиение объектов (4 фолда).

Этап 5: финальное обучение

Теперь всё готово!

Обучаем модель на всех днях, предсказываем следующие значения.

Т.к контест принимает только целые числа, округляем результат.

Итоговый результат:

- 0.51997 (как бы 3 место) / 0.47523 (как бы 7 место)
- немного переобучились под паблик

Что ещё пробовали

- увеличение количества сэмплов (замедляет обучение, k=4 улучшает качество, k=6 почти не меняет)
- добавление в качестве признаков алгоритмов первого бейзлайна — медианы по последним элементам и взвешенного среднего (почти не меняет результат)
- добавление категориальных признаков из дополнительных данных как Bag of words признаков (улучшает качество)
- использовать MAE в качестве функционала обучения (ухудшает качество по сравнению с MSE и замедляет обучение)

Что можно было бы сделать лучше при решении

- Гиперпараметры подбирались без учёта того, что модель выдаёт целые числа. Если написать кастомную метрику под бустинг, можно было бы подбирать гиперпараметры с учётом целочисленных выходов.
- Не попробовали добавлять в ряд признаки из других рядов (сложно было встроить в существующий пайплайн).