Курс по методам машинного обучения

Практическое задание № 7

Градиентный бустинг деревьев

Юлиан Сердюк

1 Характеристики задания

- **Длительность:** 2 недели (до жесткого дедлайна)
- **Кросс-проверка:** 20 баллов; в течение 1 недели после жесткого дедлайна; нельзя сдавать после жесткого дедлайна
- **ML-решение:** 20 баллов; можно сдавать после жесткого дедлайна; публичная и приватная часть
- **Почта:** ml.cmc@mail.ru
- **Темы для писем на почту:** ВМК.ML[Задание 7][peer-review], ВМК.ML[Задание 7][unit-tests], ВМК.ML[Задание 7][ML]

Кросс-проверка: После окончания срока сдачи, у вас будет еще неделя на проверку решений как минимум **3х других студентов** — это **необходимое** условие для получения оценки за вашу работу. Если вы считаете, что вас оценили неправильно или есть какие-то вопросы, можете писать на почту с соответствующей темой письма

2 Описание задания

Внимание! Это описание ML задания про градиентному бустингу. Также на cv-gml.ru имеется ноутбук по соответствующему заданию, в котором вы можете найти дополнительную информацию по методам, которые будут полезны для решения данного задания.

Привет, ребятушки! Сегодня мы с вами будем заниматься благодарным делом, а именно предсказывать популярность фильмов. Описание присутствует в основном ноутбуке по градиентному бустингу, поэтому просто скопирую сюда самое важное.

В некотором царстве, некотором государстве была развита кинопромышленность. Новые фильмы в этом государстве показывают по интернету, а пользователи после просмотра могут дать фильму некоторую "награду". Наша цель - предсказать число наград для фильма.

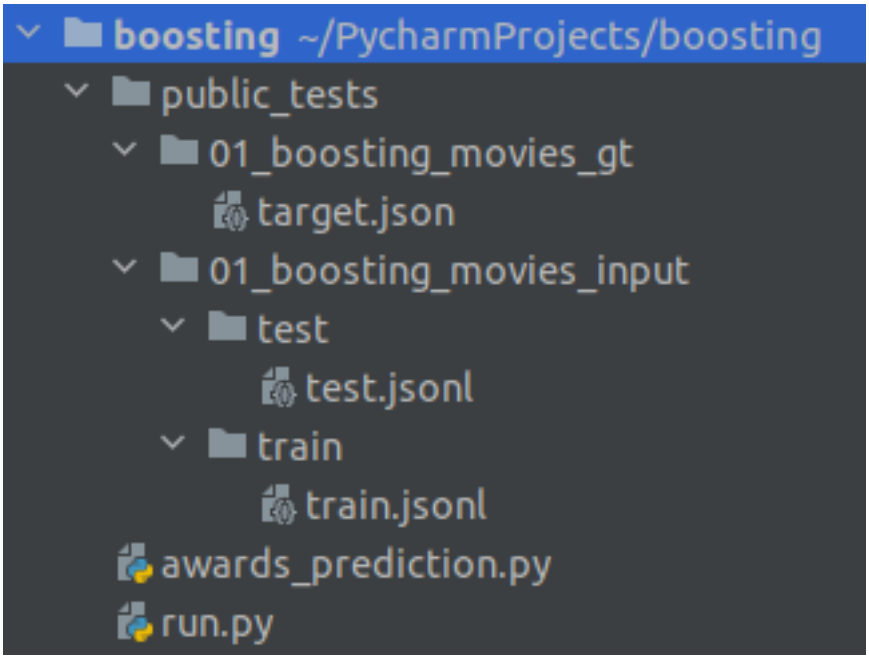
В нашем распоряжении имеются следующие данные:
awards - количество наград, полученных фильмом от пользователей (целевое значение)
potions - количество магических зелий, потраченных на создание спец-эффектов
genres - жанры созданного фильма
questions - количество вопросов, заданных пользователями на соответствующих форумах об этом фильме до премьеры
directors - режиссеры фильма (если неизвестны, то unknown)
filming_locations - области, в которых снимался фильм
runtime - продолжительность фильма в некоторых единицах, принятых в этом государстве

critics_liked - количество критиков из 100, присудивших награды фильму на предварительных закрытых показах
pre-orders - количество зрителей, заранее купивших билеты на первый показ
keywords - ключевые слова, описывающие содержание фильма
release_year - год, во котором фильм был показан (конечно, в летоисчислении этого государства)

Следующие поля появляеются несколько раз с разными значениями i:
actor_i_known_movies - количество известных фильмов актера i (i от 1 до 3)
actor_i_postogramm - количество подписчиков в социальной сети "по сто грамм"актера i (i от 1 до 3)
actor_i_gender - количество пол актера i (i от 1 до 3)
actor_i_age - возраст актера i (i от 1 до 3)

3 Файлы и папки

Структура файлов и папок описываемых заданий должна быть такой:



Если всё сделано правильно, то при переходе в соответствующую папку в консоли и запуске команды ‘python3 run.py‘ Вы не должны получать сообщений об ошибках. Учтите, что после запуска скрипта будет создано несколько дополнительных файлов и директорий (это связано с работой тестирующей системы).

4 Исходные данные

Данные разбиты на три части: тренировочная часть, публичные тестовые команды (которые вы можете скачать) и приватные тестовые файлы.

Данные храняться в виде jsonl файлов, то есть файлов, в которых каждая строка - это json. Каждая строка соответствует одному объекту в датасете. Ключи этого json соответствуют названиям переменных. В примере решения Вы можете увидеть как читать такие файлы без проблем.

В тестовом файле сохраняйте даннные в том же порядке, в котором они записаны в файл!

5 Решение

Внимание! Подбирать оптимальные параметры стоит только на локальном компьютере! В решении Вы должны использовать регрессор с оптимальными параметрами, которые вы нашли путём перебора по сетке.

В шаблонном файле ‘awards_prediction.py‘ Вы должны реализовать функцию train_model_and_predict, которая получает на вход папку для обучения и теста. На обучении вы обучаете ваш алгоритм, а затем возвращаете предсказания значений awards для всех фильмов из теста. Предсказания должны быть расположены в том же порядке, в котором они находятся в тесте (то есть, не примените где-то случайно shuffle).

В этом файле вы можете создавать любые дополнительные функции и методы, которые нужны вам для решения. Главное – сохранить интерфейс функции train_model_and_predict.

6 Советы по решению

В этом задании Вы можете добиться лучшего качества при помощи:

1. Предобработки датасета, выбора категориальных переменных и дополнительной фильтрации.

2. Выбора лучшего метода обучения и подбора оптимальных параметров с использованием кросс-валидации.

7 Тестирование решения

После реализации ‘awards_prediction.py‘ Вам необходимо запустить из консоли файл run.py. Если всё верно, то Вы увидите что-то вроде ‘Mark: 1 OK, mae = [2287.341688095019]‘, т.е. Вашу оценку и значение MAE на публичном датасете.

8 Разрешенные методы и библиотеки

В качестве метода обучения предлагается использовать любой регрессор, основанный на градиентном бустинге деревьев. Разрешается пользоваться библиотеками sklearn, xgboost, lightgbm, catboost.

Жесткого требования использовать градиентный бустинг нет! Градиентный бустинг является одним из лучших методов обучения на сегодняшний день, поэтому будет даже интересно, удастся ли кому-то получить макс. балл альтернативными методами.

Также разрешается пользоваться библиотекой hyperopt для подбора параметров модели.

9 Ограничения для скрипта на cv-gml.ru

В этом задании стоит ограничение по времени: 10 минут. Также Вы не можете использовать памяти больше, чем 1024 мб.

10 Используемая метрика

В качестве метрики качества используется значение MAE, которое вычисляется по следующей формуле:

$$MAE = \sum_{i=1}^{N} \frac{|a(x_i) - y_i|}{N},$$

где N - число объектов в тестовой выборке, x_i - вектор признаков i-го объекта, $a(x_i)$ - предсказание на i-ом объекте, y_i - значение целевой переменной для i-го объекта.

11 Оценивание

Баллы выставляются по следующим правилам:

5 баллов : mae $\in$ [0, 2100],

3 балла: mae $\in$ (2100, 2200],

1 балл: mae $\in$ (2200, 2300],

0 баллов: mae $\in$ (2300, $+\infty$]

Значение mae будет посчитано раздельно на публичной и приватной выборках. Количество полученных баллов на приватной выборке будет дополнительно умножено на 3. Таким образом за это задание Вы можете получить до 20 баллов (5 на публичной и 15 на приватной выборках).

12 Возможная проблема с catboost

Обучайте регрессор со следующим параметром train_dir:
“CatBoostRegressor(train_dir=’/tmp/catboost_info’)”

Важно: *перед сдачей проверьте, пожалуйста, что не оставили в ноутбуке где-либо свои ФИО, группу и т.д. — кросс-рецензирование проводится анонимно.*

Важно: *задания, в которых есть решения, содержащие в каком-либо виде взлом тестов и прочие нечестные приемы, будут автоматически оценены в 0 баллов без права пересдачи задания.*