ВМК МГУ

Задание 6. Деревья решений

Курс по методам машинного обучения, 2023, Юлиан Сердюк

1 Характеристики задания

- **Длительность:** 2 недели
- **Кросс-проверка:** 20 баллов + 2 бонусных балла; в течение 1 недели после дедлайна; нельзя сдавать после жесткого дедлайна
- **Юнит-тестирование:** 5 баллов (1 public + 4 private); можно сдавать после дедлайна со штрафом в 40%; публичная и приватная части; PEP8
- **ML-задание:** 15 баллов (5 public + 10 private); нельзя сдавать после дедлайна; публичная и приватная части; PEP8
- **Почта:** ml.cmc@mail.ru
- **Темы для писем на почту:** ВМК.ML[Задание 6][peer-review], ВМК.ML[Задание 6][unit-tests], ВМК.ML[Задание 6][ml-task]

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

Привет, ребятушки!
В этом файле описываются сразу ДВА задания, которые относятся к теме решающих деревьев!

- Оценка разбиений (юнит-тесты)
- Предсказание потенциалов (ML-решение)

3 Кросс-проверка

- **Ссылка на задание:** ссылка тут

Замечание: После отправки ноутбука убедитесь, что все графики сохранены корректно и правильно отображаются в системе.

Замечание: Перед сдачей проверьте, пожалуйста, что не оставили в ноутбуке где-либо свои ФИО, группу и так далее — кросс-рецензирование проводится анонимно.

Замечание: В ноутбуке можно использовать дополнительные импорты.

4 Юнит-тестирование (Оценка разбиений)

В этом задании вам нужно реализовать функцию, которая оценивает качество разбиения множества объектов. На лекции вам рассказали о трех метриках: gini, entropy и classification_error. В этом задании вам нужно реализовать их все. Для вычислений разрешается пользоваться библиотекой numpy.

Замечание: При вычислениях используйте **натуральный** логарифм (подробнее можно посмотреть в FAQ).

4.1 Формат отправки

В шаблонном файле split_measures.py, вам необходимо реализовать функцию evaluate_measures. На вход эта функция получает список меток классов объектов, которые попали в одно из получившихся разбиений. Эта функция возвращает словарь, в котором содержатся значения всех трёх метрик. Вовращаемый словарь должен содержать три ключа: “gini”, “entropy” и “error”, которым должны быть сопоставлены значения метрик gini, entropy и classification error в типе float ().

Ниже показан схематический пример такого скрипта.

```
def evaluate_measures(sample):
    return {"gini": float(np.sum(sample)),
            "entropy": float(np.min(sample)),
            "error": float(np.max(sample))}

print(evaluate_measures([1, 2, 3, 2, 3, 1, 2, 0]))
```

4.2 Используемая метрика

Для успешного выполнения задания необходимо, чтобы Ваше полученное значение отличалось от истинного не более, чем на 0.001 (абсолютное значение разности).

4.3 Оценивание

Задание разбито на две части: публичную (один пример) и приватную (содержащую несколько выборок). Для получения оценки в каждой из частей вам необходимо пройти все тесты без ошибок вычислить ответ и получить значение, удовлетворяющее метрике.

Если вы сдаёте задание до дедлайна, то вы можете получить максимум 5 баллов (1 балл за публичную часть и 4 балла за приватную). После дедлайна вам будет зачтено лишь 60% от вашего полного балла.

4.4 Дополнительные материалы

Если вы хотите почитать что это за метрики и как их вычислять, Вы можете обратиться к лекциям Китова или Дьяконова. Воронцов, к сожалению, слишком поверхностно упоминает эти метрики :(

Замечание: Запрещается пользоваться библиотеками, импорт которых не объявлен в файле с шаблонами функций.

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

5 ML-задание (Предсказание потенциалов)

5.1 Описание задания

Внимание! АХТУНГ! АЛЯРМ! Дедлайн у задания жесткий! После оглашения результатов посылки будут запрещены.

Имеется некоторая размером 256 x 256, характеризующая некоторый потенциал. Для каждого потенциала нам необходимо предсказать величину его энергии используя деревья решений и все виды ансамблей над решающими деревьями, исключая те, что используют градиентный бустинг (хотя бустинг тут и не особо и помогает, почему-то). Со списком ансамблей, релизованных в sklearn, можно ознакомиться по ссылке https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble. Убедитесь, что выбранный Вами ансамбль не использует бустинг. Поскольку этот датасет описан в соответствующем ноутбуке, повторять его описание не буду.

Добиться хорошего качества в этом задании необходимо при помощи трех вещей:

- нахождения лучшего алгоритма обучения

- нахождения оптимальных параметров (размер леса, максимальная глубина, размер признаков или прочего). Рекомендуется ознакомиться с параметрами, используемыми в случайных лесах, чтобы знать, какие параметры Вы можете подбирать (СОВЕТ: Вы можете изменять не только свойства деревьев в лесу, но и оптимизируемый функционал)
- нахождения лучшего способа преобразования 2d матрицы в одномерный вектор признаков (с возможным перемещением/изменением потенциала).

Внимание! В качестве обучаемого регрессора (который делает предсказания) можно выбрать любой метод, основанный на деревьях или лесах. Ансамбли с применением бустинга использовать нельзя.

Для тестирования решения Вы должны скачать публичный датасет, поместить его в папку public_tests и, используя скрипт run.py, протестировать решение так, как это будет сделано проверяющей системой.

5.2 Вид решения

В Вашем решении в шаблонном файле potential_prediction.py необходимо реализовать(дополнить) функцию
train_model_and_predict(train_dir, test_dir), которая тренируется на потенциалах из тренировочной папки, а потом возвращает словарь вида {файл:предсказание} для каждого файла из тестовой директории. Датасеты и пример скрипта, реализующий необходимый функционал, Вы можете найти на страницах соответствующего задания на cv-gml.ru.

5.3 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.

5.4 Время обучения

Внимание! В проверяющей системе установлено ограничение на время работы Вашего скрипта, поэтому настройку методов и подбор параметров Вы должны провести локально (например, в ноутбуке), а в систему загружать уже метод с установленными параметрами.

Время решения ограничено 10 минутами (суммарно время трансформации, обучения и предсказания). Поскольку время обучения напрямую зависит от размера леса и выбранных параметров, подбирайте их таким образом, чтобы лес обучался не слишком долго.

При формировании GridSearch учтите, что время обучения напрямую зависит от количества узлов в сетке. В то же время, слишком большие значения параметров тестировать не стоит, если время обучения на них превышает лимит в 10 минут.

Pro tip! Как это часто бывает в машинном обучении, время обучения напрямую зависит от сложности оптимизационной задачи. Если зависимости, которые Вы хотите найти, достаточно “очевидны”, то леса будут строиться относительно быстро. Если Ваши зависимости очень сложны (или, еще хуже, они отсутствуют), то леса будут тратить много времени на поиск оптимального разбиения. Поэтому для ускорения решения рекомендуется не только подбирать правильные параметры, но и найти то преобразование, которое упростит решение задачи.

5.5 Видимость результатов

Тестирование проводится на двух тестовых выборках: открытой и закрытой. Открытую вы можете загрузить себе и тестировать свой метод на ней. После загрузки Вашего решения в тестирующую систему оно изначально будет протестировано на открытой выборке. Когда наступит дедлайн задания, Вы увидите результаты

своего метода на закрытой тестовой выборке, и по значению целевой метрики вы получите оценку за это задание.

5.6 Оценивание

Баллы выставляются по следующим правилам:
5 баллов : mae $\in$ [0, 0.007],
4 баллов: mae $\in$ (0.007, 0.01],
3 баллов: mae $\in$ (0.01, 0.02],
2 балла: mae $\in$ (0.02, 0.05],
0 баллов: mae $\in$ (0.05, $+\infty$]

Значение mae будет посчитано раздельно на публичной и приватной выборках. Количество полученных баллов на приватной выборке будет дополнительно умножено на 2. Таким образом за это задание Вы можете получить до 15 баллов.

5.7 Советы по решению

Внимание! В функции train_model_and_predict рекомендуется менять лишь строку, в которой определяется регрессор, а именно

```
regressor = Pipeline([('vectorizer', PotentialTransformer()), \
('decision_tree', DecisionTreeRegressor())])
```

Здесь Вы можете переопределить Pipeline, реализовать свой собственный трансформер и поставить предпочтительный алгоритм обучения.

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

Важно: *В этом задании в файле с шаблоном* potential_prediction.py *можно использовать дополнительные импорты (например, для обработки признаков), однако модель, которую вы обучаете, должна быть основана на деревьях или лесах (**но бустинги использовать нельзя!**)*

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.*

6 Стиль программирования

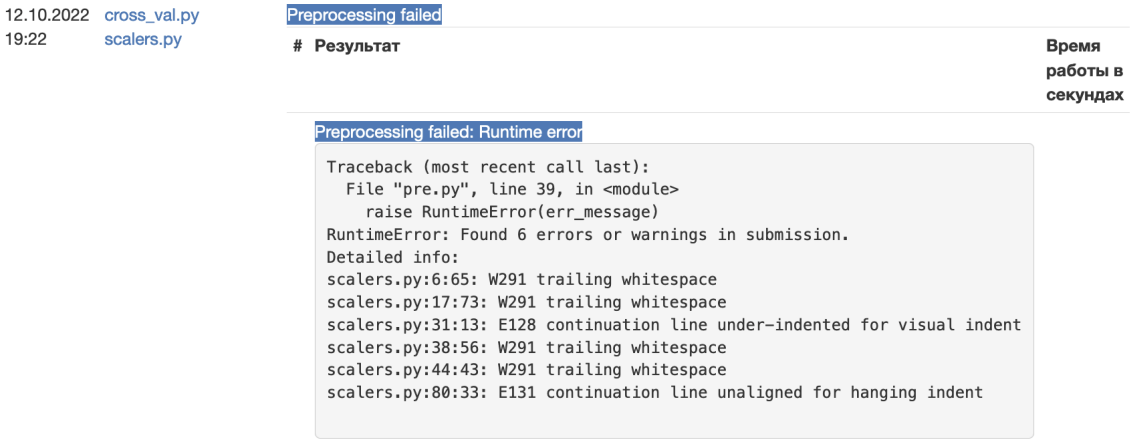
При выполнении задач типа unit-tests, ML-задания вам необходимо будет соблюдать определенный стиль программирования (codestyle). В данном случае мы выбирали PEP8 как один из популярных стилей для языка Python. Зачем мы это вводим? Хорошая читаемость кода – не менее важный параметр, чем работоспособность кода :) Единый стиль позволяет быстрее понимать код сокомандников (в командных проектах, например), упрощает понимание кода (как другим, так и вам). Также, привыкнув к какому-либо стилю программирования, вам будет проще переориентироваться на другой.

Полезные при изучении PEP8 ссылки, если что-то непонятно, дополнительный материал можно найти самостоятельно в интернете:

- Официальный сайт PEP8, на английском
- Небольшое руководство по основам на русском

Требования к PEP8 мы вводим только для заданий с авто-тестами, требований к такому же оформлению ноутбуков нет. Но улучшение качества кода в соответствии с PEP8 в них приветствуется!

В проверяющей системе, при несоответствии прикрепляемого кода PEP8, будет высвечиваться вердикт `Preprocessing failed`. Более подробно посмотреть на ошибки можно, нажав на них:



Проверить стиль программирования локально можно при помощи утилиты pycodestyle с параметром максимальной длины строки (мы используем 160 вместе дефолтных 79):

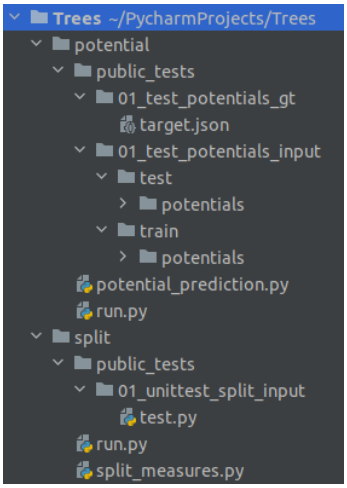
```
pycodestyle --max-line-length=160 your_file_with_functions.py
```

7 Тестирование

В cv-gml можно скачать все файлы, необходимые для тестирования, одним архивом. Для этого просто скачайте zip-архив во вкладке **шаблон решения** соответствующего задания и разархивируйте его. Далее следуйте инструкциям по запуску тестирования.

Внимание! Эти задания используют разные файлы run.py! Из-за технических особенностей тестируемой системы все запускающие файлы называются run.py, которые имеют различные реализации в различных заданиях. Поэтому рекомендуем создать разные папки для каждого задания и в каждом хранить соответствующий файл run.py.

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).

Запуск тестов производится командой

```
python3 run.py
```

Успехов!