

Задание 3. К ближайших соседей (kNN) Research

Курс по методам машинного обучения, 2025-2026, Драгунов Никита

1 Характеристики задания

Уровень Research

- **Длительность:** 2 недели
- **Кросс-проверка:** 7 баллов + 0.5 бонусных баллов; в течение 1 недели после дедлайна; Нельзя сдавать после дедлайна.
- **Юнит-тестирование:** 3 баллов; Можно сдавать после дедлайна со штрафом в 40%; Публичная часть.

Почта

- **Почта:** ml.cmc@mail.ru
- **Темы для писем на почту:** ВМК.ML[Задание 3][unit-tests], ВМК.ML[Задание 3][peer-review]

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

В настоящем задании вы познакомитесь с алгоритмом k ближайших соседей (k-NN, KNN, kNN) для решения задач классификации и регрессии.

Приведем здесь краткое напоминание о принципе работы kNN. Пусть дана обучающая выборка $X = (x_i, y_i)$ и функция расстояния ρ . Требуется классифицировать новый объект u . Алгоритм k ближайших соседей относит объект u к тому классу, представителей которого окажется больше всего среди k его ближайших по ρ соседей: $a(u; X, k) = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1}^k w_i [y_u^{(i)} = y]$, где $y_u^{(i)}$ — метка класса i-го соседа объекта u . В классическом методе k ближайших соседей все объекты имеют единичные веса: $w_i = 1$. Альтернативой данному подходу служат веса, обратно пропорциональные расстоянию между объектами. При решении задачи регрессии ответом алгоритма служит средневзвешенное значение меток $y_u^{(i)}$ среди k ближайших соседей.

3 Кросс-проверка

Внимание! Отправлять задание нужно в систему во вкладку с пометкой (notebook).

Внимание! Отправлять задание нужно только с расширением `ipynb`! После отправки проверьте корректность загруженного задания в систему, просмотрев глазами загруженное решение (оно автоматически сконвертируется в `html`). Как это сделать, можно найти в руководстве по проверяющей системе на сайте курса.

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

4 Юнит-тестирование

- В ходе выполнения задания из jupyter-notebook вам необходимо реализовать две функции в модуле `cross_val.py`. Первая функция `kfold_split` выполняет генерацию индексов обучающей и валидационной выборки для кросс-валидации. Вторая функция `knn_cv_score` выполняет кросс-валидацию для KNN-модели.

Файл `cross_val.py` можно найти в архиве из **шаблона решения** из соответствующего задания в системе проверки. Более подробное описание входных и выходных данных вы найдете в этих файлах. После реализации ваш код можно протестировать локально (описано в конце pdf), а затем его необходимо сдать в проверяющую систему.

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 Стиль программирования

Внимание! Обновление!!!

Начиная с этого задания 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!

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

12.10.2022 `cross_val.py`
19:22 `scalers.py`

`Preprocessing failed`

Результат

Время
работы в
секундах

`Preprocessing failed: Runtime error`

```
Traceback (most recent call last):
  File "pre.py", line 39, in <module>
    raise RuntimeError(err_message)
RuntimeError: Found 6 errors or warnings in submission.
Detailed info:
scalers.py:6:65: W291 trailing whitespace
scalers.py:17:73: W291 trailing whitespace
scalers.py:31:13: E128 continuation line under-indented for visual indent
scalers.py:38:56: W291 trailing whitespace
scalers.py:44:43: W291 trailing whitespace
scalers.py:80:33: E131 continuation line unaligned for hanging indent
```

Также послышки, упавшие по code style, считаются за попытку сдачи и идут в счет общего количества попыток за день.

Проверить стиль программирования локально можно при помощи утилиты `pycodestyle` (в окружении, которое вы ставили, эта утилита уже есть) с параметром максимальной длины строки (мы используем 160 вместо дефолтных 79):

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

6 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.

Тесты запускаются с помощью команд:

```
python run.py unittest cv
```