Задания для выполнения расчетно-графической работы по дисциплине “Математическое моделирование прикладных задач”

РГР должна быть оформлена в печатном виде на листах формата А4 и подшита в папку. Электронная версия РГР, ссылка на ваш код в Google Colaboratory (либо файл Jupyter Notebook) должна быть отправлена через сервис ДО ЧГУ moodle .

Состав: титульный лист, содержание, введение, техническое задание, описание проделанной работы с подробным объяснением и приложением всех требуемых скриншотов работы программ (Глава 1 и Глава 2), а также ссылок на Google Colaboratory, заключение, список используемых источников, Приложение А (с указанием кода первого задания), Приложение Б (с указанием кода второго задания).

Навыки, знания и умения, наличие которых оценивает РГР:

- умение работать с библиотеками Pandas, Numpy, Scikit Learn, MatPlotLib.
- Умение работать с текстовыми файлами форматов .txt и .csv.
- работа с различными коллекциями в Python - списками, словарями.
- знания базового синтаксиса Python.
- Знания базовых алгоритмов кластеризации, основ кластерного анализа и анализа данных, алгоритмов классификации и регрессии.

Задание 1. Кластерный анализ данных.

Вам предлагается датасет. Необходимо изучить описание данных, при его наличии. Вам необходимо провести кластерный анализ, разбить данные на группы (кластеры).

Порядок действий:

1. Загрузить данные. Если данные разбиты на файлы, необходимо объединить их в одну общую базу.
2. Осуществить предобработку данных.
    a. Удаление лишних неинформативных признаков
    b. Удаление (либо заполнение пропущенных значений). Обоснуйте ваш метод.
    c. Обработка категориальных данных, нормализация количественных признаков.
3. Проведите кластерный анализ 2-мя алгоритмами: методом k-means а также иерархической кластеризацией. Определите оптимальное число кластеров. Опишите ход ваших мыслей и алгоритм вашей работы.
4. Дайте развернутое описание и характеристику каждому кластеру.

Задание 2. Задача классификации.

Вам предлагается датасет. Необходимо изучить описание данных, при его наличии.
Необходимо разработать модель классификации для предсказания класса целевой переменной по вектору признаков.

Для расчета точности применяется метрика accuracy.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score
Она вычисляется как доля верно предсказанных классов. Вы можете рассчитать ее самостоятельно, или использовать встроенную функцию в Scikit Learn.

Порядок действий:

1. Загрузить данные. Если данные разбиты на файлы, необходимо объединить их в одну общую базу.
2. Осуществить предобработку данных.
    a. Удаление лишних неинформативных признаков
    b. Удаление (либо заполнение пропущенных значений). Обоснуйте ваш метод.
    c. Обработка категориальных данных, нормализация количественных признаков.
    d. Создание обучающей и тестовой выборки.
3. Разработайте модель методом KNN, подберите оптимальное значение k такое, чтобы точность модели была максимальная при отсутствующем переобучении. Выведите метрики полученной модели. Постройте графики зависимости точности классификатора от числа k.

Задание 3. Задача регрессии.

Вам предлагается датасет. Необходимо изучить описание данных, при его наличии.
Необходимо разработать модель регрессии для предсказания значения количественной целевой переменной по вектору признаков.

Самостоятельно изучите построение регрессионной моде и KNN:

1. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html
2. https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/
3. https://www.machinelearningmastery.ru/the-basics-knn-for-classification-and-regression-c1e8a6c955/
4. https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_k-%D0%B1%D0%BB%D0%B8%D0%B6%D0%B0%D0%B9%D1%88%D0%B8%D1%85_%D1%81%D0%BE%D1%81%D0%B5%D0%B4%D0%B5%D0%B9
5. https://russianblogs.com/article/6216957921/

В качестве метрики качества используйте MSE – среднеквадратичую ошибку, либо MAE – среднюю абсолютную ошибку предсказанных значений от целевых. Вы можете рассчитать ее самостоятельно, либо применить встроенную функцию в библиотеке Scikit Learn.

1. https://www.machinelearningmastery.ru/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0/
2. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

Порядок действий:

1. Загрузить данные. Если данные разбиты на файлы, необходимо объединить их в одну общую базу.

2. Осуществить предобработку данных.
    a. Удаление лишних неинформативных признаков
    b. Удаление (либо заполнение пропущенных значений). Обоснуйте ваш метод.
    c. Обработка категориальных данных, нормализация количественных признаков.
    d. Создание обучающей и тестовой выборки.
3. Разработайте модель методом KNN, подберите оптимальное значение k такое, чтобы точность модели была максимальная при отсутствующем переобучении. Выведите метрики полученной модели. Постройте графики зависимости величины ошибки от числа k соседей. Выведите итоговую ошибку на тренировочной и тестовой выборке.

По итогам работы оформить отчет в печатном и электронном виде. Электронный отчет загрузить в систему ДО ЧГУ moodle .