Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный Исследовательский Университет ИТМО"
Мегафакультет Компьютерных Технологий и Управления
Факультет Программной Инженерии и Компьютерной Техники

Модуль №2
по дисциплине
'Системы искусственного интеллекта'

Выполнил Студент группы P33102
Лапин Алексей Александрович
Преподаватель:
Авдюшина Анна Евгеньевна

г. Санкт-Петербург
2023г.

Содержание

1 Введение:

1.1 Описание целей проекта и его значимости.

2 Анализ требований:

Лабораторная 4. Линейная регрессия

Задание

- Выбор датасетов: Студенты с нечетным порядковым номером в группе должны использовать про обучение студентов
- Получите и визуализируйте статистику по датасету (включая количество, среднее значение, стандартное отклонение, минимум, максимум и различные квантили).
- Проведите предварительную обработку данных, включая обработку отсутствующих значений, кодирование категориальных признаков и нормировка.
- Разделите данные на обучающий и тестовый наборы данных.
- Реализуйте линейную регрессию с использованием метода наименьших квадратов без использования сторонних библиотек, кроме NumPy и Pandas (для использования коэффициентов использовать библиотеки тоже нельзя). Использовать минимизацию суммы квадратов разностей между фактическими и предсказанными значениями для нахождения оптимальных коэффициентов.
- Постройте три модели с различными наборами признаков.
- Для каждой модели проведите оценку производительности, используя метрику коэффициент детерминации, чтобы измерить, насколько хорошо модель соответствует данным.
- Сравните результаты трех моделей и сделайте выводы о том, какие признаки работают лучше всего для каждой модели.
- Бонусное задание - Ввести синтетический признак при построении модели

Лабораторная 5. Метод k-ближайших соседей

Задание

- Выбор датасета: Нечетный номер в группе - Датасет про диабет
- Проведите предварительную обработку данных, включая обработку отсутствующих значений, кодирование категориальных признаков и масштабирование.
- Реализуйте метод k-ближайших соседей без использования сторонних библиотек, кроме NumPy и Pandas.
- Постройте две модели k-NN с различными наборами признаков:
  - Модель 1: Признаки случайно отбираются .

  - Модель 2: Фиксированный набор признаков, который выбирается заранее.

- Для каждой модели проведите оценку на тестовом наборе данных при разных значениях k. Выберите несколько различных значений k, например, k=3, k=5, k=10, и т. д. Постройте матрицу ошибок.

Лабораторная 6. Деревья решений

Задание

- Для студентов с четным порядковым номером в группе – датасет с классификацией грибов, а нечетным – датасет с данными про оценки студентов инженерного и педагогического факультетов (для данного датасета нужно ввести метрику: студент успешный/неуспешный на основании грейда)
- Отобрать случайным образом sqrt(n) признаков
- Реализовать без использования сторонних библиотек построение дерева решений (numpy и pandas использовать можно, использовать списки для реализации дерева - нельзя)
- Провести оценку реализованного алгоритма с использованием Accuracy, precision и recall
- Построить AUC-ROC и AUC-PR (в пунктах 4 и 5 использовать библиотеки нельзя)

Лабораторная 7. Логистическая регрессия

Задание

- Выбор датасета: Датасет о диабете: Diabetes Dataset
- Загрузите выбранный датасет и выполните предварительную обработку данных.
- Разделите данные на обучающий и тестовый наборы в соотношении, которое вы считаете подходящим.
- Реализуйте логистическую регрессию "с нуля"без использования сторонних библиотек, кроме NumPy и Pandas. Ваша реализация логистической регрессии должна включать в себя:
  - Функцию для вычисления гипотезы (sigmoid function).
  - Функцию для вычисления функции потерь (log loss).
  - Метод обучения, который включает в себя градиентный спуск.
  - Возможность варьировать гиперпараметры, такие как коэффициент обучения (learning rate) и количество итераций.
- Исследование гиперпараметров: Проведите исследование влияния гиперпараметров на производительность модели. Варьируйте следующие гиперпараметры:
  - Коэффициент обучения (learning rate).
  - Количество итераций обучения.

  - Метод оптимизации (например, градиентный спуск или оптимизация Ньютона).

- Оценка модели: Для каждой комбинации гиперпараметров оцените производительность модели на тестовом наборе данных, используя метрики, такие как accuracy, precision, recall и F1-Score.

- Сделайте выводы о том, какие значения гиперпараметров наилучшим образом работают для данного набора данных и задачи классификации. Обратите внимание на изменение производительности модели при варьировании гиперпараметров.

3 Лабораторная 4. Линейная регрессия

3.1 Реализация:

Реализация линейной регрессии GitHub

4 Лабораторная 5. Метод k-ближайших соседей

4.1 Реализация:

Реализация метода k-ближайших соседей GitHub

5 Лабораторная 6. Деревья решений

5.1 Реализация:

Реализация деревьев решений GitHub

6 Лабораторная 7. Логистическая регрессия

6.1 Реализация:

Реализация логистической регрессии GitHub