Обработка данных на языке Python

28.09.2024

Какие библиотеки используются

- NumPy
- Pandas
- Matplotlib
- Seaborn
- opencv*
- NLTK*

Какие бывают данные

Числовые

Категориальные

Текстовые

Временные ряды

Изображения

Аудио

Видео

Сетевые данные

Пространственные данные

Числовые (Numerical) данные

1. Непрерывные (Continuous): Значения могут принимать любое значение на определенном интервале. Примеры: температура, вес, доход.
2. Дискретные (Discrete): Значения могут быть только определенными точками в пределах интервала. Примеры: количество детей, количество машин.

Категориальные (Categorical) данные

1. Номинальные (Nominal): Категории не имеют естественного порядка. Примеры: цвета (красный, зеленый, синий), типы животных (собака, кошка, птица).
2. Порядковые (Ordinal): Категории имеют естественный порядок, но расстояние между ними не определено. Примеры: уровни образования (начальное, среднее, высшее), оценки (плохо, удовлетворительно, хорошо, отлично).

Текстовые (Textual) данные

Обычный текст, который может быть представлен в виде строк.
Примеры: отзывы клиентов, статьи, сообщения в социальных сетях.

Временные ряды (Time Series)

Данные, собранные в течение определенного времени, обычно с фиксированными интервалами. Примеры: цены акций, погодные данные, данные сенсоров.

Изображения (Image Data)

Пиксельные данные, представляющие визуальную информацию.
Примеры: фотографии, рентгеновские снимки.

Аудио (Audio Data) / Видео (Video Data)

- Звуковые волны, которые могут быть представлены в виде временных рядов или спектрограмм. Примеры: музыка, речь.
- Последовательность изображений (кадров), возможно вместе с аудио. Примеры: видеозаписи, потоковое видео.

Сетевые данные (Graph Data)

Данные, представляющие структуры в виде узлов и ребер. Примеры: социальные сети, сети дорог, молекулярные структуры.

Пространственные данные (Spatial Data)

Данные, которые включают географические координаты или другую пространственную информацию. Примеры: карты, GPS-данные.

Предварительная обработка данных

1. Обработка отсутствующих значений
2. Обнаружение и удаление дубликатов
3. Преобразование типов данных
4. Очистка данных

Обработка отсутствующих значений

Отсутствующие значения (NaN, null) могут негативно сказаться на анализе и моделировании данных. Методы решения проблемы:

1. Удаление строк/столбцов с отсутствующими значениями.
2. Замена отсутствующих значений.
3. Могут использоваться более сложные методы, такие как интерполяция или модели машинного обучения, чтобы предсказать и заполнить отсутствующие значения.

Обнаружение и удаление дубликатов

Дубликаты данных могут исказить результаты анализа и моделей. В Pandas существуют методы для решения проблемы:

1. Обнаружение дубликатов: `duplicated()`: возвращает булев массив, указывающий на дубликаты.
2. Удаление дубликатов: `drop_duplicates()`: удаляет дубликаты из DataFrame.

Очистка данных

Очистка данных включает в себя различные методы для удаления пробелов, исправления ошибок и приведения данных к нужному формату.

- Удаление пробелов.
- Исправление ошибок.
- Удаление или замена некорректных значений.

Манипуляции с данными с помощью Pandas

- Основные структуры данных: DataFrame и Series
- Индексация и выборка данных
- Фильтрация, сортировка и группировка данных
- Объединение и слияние DataFrame

Анализ данных

- Описание и статистический анализ данных
- Визуализация данных с помощью Matplotlib и Seaborn (Примеры использования: гистограммы, боксплоты, тепловые карты)

Описание и статистический анализ данных

На этом этапе важно понять структуру данных, их типы и основные характеристики:

1. Описание данных: Вычисление основных статистических показателей (среднее, медиана, мода, стандартное отклонение и т.д.).
2. Инферентная статистика: Делает выводы о генеральной совокупности на основе выборки (доверительные интервалы, гипотезы, тесты значимости).
3. Распределение вероятностей: Анализ распределения данных (нормальное, биномиальное, пуассоновское и т.д.).

Визуализация данных

Matplotlib — это библиотека для создания статических, анимационных и интерактивных визуализаций в Python. Она предоставляет гибкие средства для построения графиков разнообразных типов.

Линейные графики Гистограммы Графики рассеяния	Графики с областями Круговые диаграммы 3D-графики

Что на следующих этапах?

Разделение данных на тренировочные и тестовые наборы;

Выбор модели;

Обучение модели;

Тонкая настройка гиперпараметров;

Оценка модели;

Интерпретация и объяснение модели;

Тестирование модели на новых данных;

Внедрение модели;

Мониторинг и обновление модели;

Рассмотренные сегодня темы в контексте проектных задач