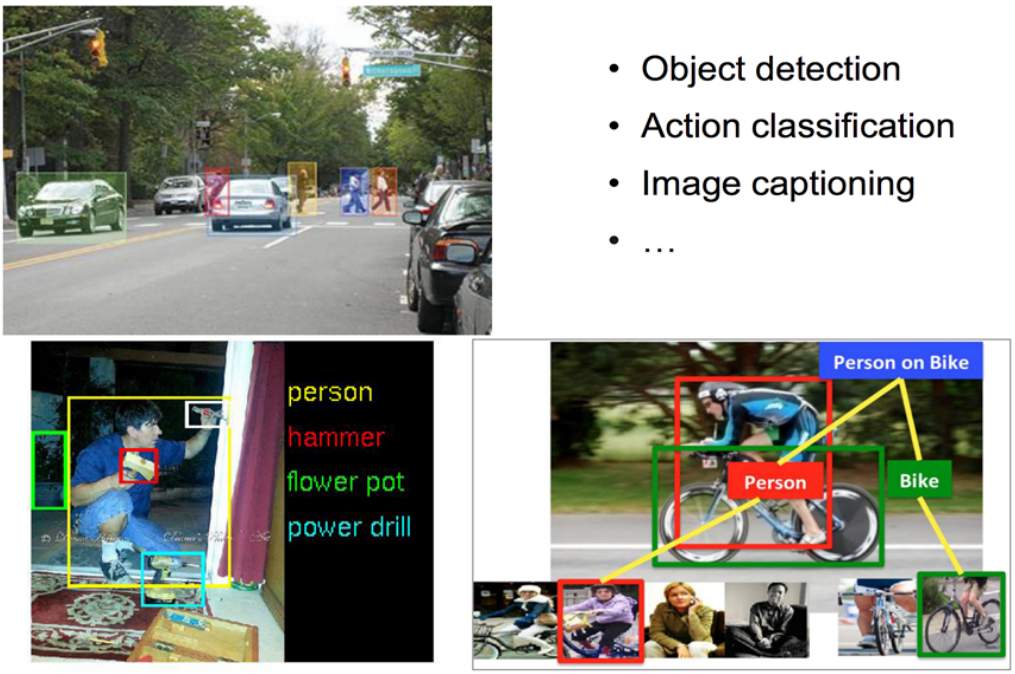
Введение в анализ данных и машинное обучение

В современном мире, где данные стали неотъемлемой частью нашей жизни, умение анализировать и извлекать ценную информацию из них является критически важным навыком. Анализ данных и машинное обучение — это области, которые позволяют нам делать это эффективно.

Мы рассмотрим в этой части курса сферы применения машинного обучения и основные библиотеки, используемые для анализа данных. Подробно рассматривать библиотеки для машинного обучения вы будете во втором семестре программы.

1. Сферы применения машинного обучения

Теория машинного обучения за последние 60 лет оформилась в самостоятельную математическую дисциплину, которая находится на стыке линейной алгебры, теории вероятности, прикладной статистики, численных методов оптимизации, дискретного анализа, программирования.



Сферы применения машинного обучения

Многие из нас ежедневно используют приложения, в основе которых лежат технологии искусственного интеллекта (ИИ) и машинного обучения. Мы пользуемся виртуальными помощниками, чтобы включить будильник и узнать прогноз погоды. И рекомендательными системами, чтобы решить, какой фильм посмотреть и какую еду приготовить. Сосредоточением решений машинного обучения можно считать беспилотный автомобиль, в котором внедрены системы компьютерного зрения, предсказания действий объектов на дороге, распознавания голоса и другие.

Помимо использования машинного обучения, человечество пытается с ним соревноваться. Так, в 2016 году корейски профессиональный игрок в Го Ли Седоля был побежден в этой игре программой AlphaGo. С тех пор прогресс в этой сфере зашел так далеко, что подобная программа считается простой. Ее уже можно давать студентам в качестве домашнего или семинарского задания.

Человек пытается соревноваться с искусственным интеллектом

Технологии машинного обучения активно используются не только для решения прикладных задач, но и в науке.

До недавнего времени единственным видом, который занимался порождением знаний из данных, был человек. Вот только делали это ученые самостоятельно. Например, Ньютон и Кеплер проанализировали огромное количество данных, определи взаимосвязи и сформулировали свои законы.

Только с момента открытия машинного обучения мы перепоручили эту задачу компьютеру. Получив определенное количество данных и наблюдений и найдя логические паттерны, он может получить модель и сформулировать выводы. То есть технологии машинного обучения пытаются автоматизировать извлечение знаний из полученных данных. К сожалению, пока что это не просто рычаг, который можно дернуть и получить модель. Для этого специалистам машинного обучения приходится использовать сложные математические методы, сделать огромное количество априорных предположений.

2. Обзор библиотек для анализа данных

Python часто используют в практических задачах анализа данных. Одна из причин — большое количество мощных библиотек для решения прикладных задач.

В этой подборке мы собрали некоторые полезные модули, библиотеки и сервисы, сгруппировав их по области решаемых задач.

2.1 Математика

Модуль `math`

Модуль `math` включает математические константы и функции. Функции, включенные в модуль, делятся на группы:

- математические (`ceil`, `floor`, `factorial` и др.);
- степенные и логарифмические (`log`, `log10` и др.);
- тригонометрические (`sin`, `cos`, `asin` и др.);
- для преобразования единиц измерения углов (`degrees`, `radians`);
- гиперболические (`sinh`, `cosh`, `asinh` и др.).

Модуль работает с вещественными значениями. Для работы с комплексными числами используйте модуль `cmath`.

Полный перечень функций языка содержится в документации.

Дополнительные материалы:

1. math — Mathematical functions. Python 3.11.1 documentation
2. Python math Module – AskPython
3. Математические функции и модуль math в Python

Библиотека `NumPy`

`NumPy` — сокращение от Numeric Python. Эта библиотека — основной пакет для научных вычислений. Библиотека предоставляет объект многомерного массива `ndarray`, методы для быстрых манипуляций с массивами, а также функции для выполнения дискретного преобразования Фурье, операций линейной алгебры, основных статистических операций, случайного моделирования и многое другое.

Дополнительные материалы:

1. NumPy documentation
2. NumPy в Python. Часть 1
3. NumPy в Python. Часть 2
4. NumPy в Python. Часть 3
5. NumPy в Python. Часть 4

Библиотека `SciPy`

Пакет `SciPy` (сокращение от Scientific Python) содержит набор инструментов, предназначенных `для` выполнения научных вычислений — интерполяция, интеграция, оптимизация, обработка изображений, статистика, специальные функции и т. д. Библиотека SciPy основана на NumPy и расширяет возможности последней.

SciPy состоит из подмодулей для конкретных задач, с полным перечнем которых и их описанием вы можете ознакомиться в документации.

Например, `scipy.integrate` включает подпрограммы численного интегрирования и решения дифференциальных уравнений, `scipy.optimize` содержит алгоритмы оптимизации функций, `scipy.stats` обеспечивает статистические инструменты и вероятностные описания случайных процессов.

Дополнительные материалы:

1. SciPy
2. SciPy Lecture Notes
3. SciPy — интегрирование и дифференцирование, обработка изображений и сигналов

2. Анализ данных

Библиотека Pandas

Библиотека `Pandas` включает мощный функционал для анализа, очистки, изучения и манипулирования данными. По сути, `Pandas` — основная библиотека для анализа данных.

Библиотека предоставляет собственные объекты для работы с данными: `DataFrame` (двумерная таблица) и `Series` (одномерный массив).

Из возможностей `Pandas` стоит отметить:

- гибкие возможности для манипуляции данными;
- развитые средства индексирования;
- набор функций для упорядочивания, сортировки, фильтрации, агрегирования данных;
- инструменты для подготовки и очистки данных.

Часто `Pandas` используют в связке с библиотеками визуализации — например, `matplotlib`.

Дополнительные материалы:

1. pandas — Python Data Analysis Library
2. Learn Pandas Tutorials
3. Моя шпаргалка по pandas

Модуль `statsmodels`

Модуль `statsmodels` предоставляет классы и функции для оценки множества различных статистических моделей, а также для проведения статистических тестов и исследования статистических данных.

Функции модуля `statsmodels` обычно используют совместно с другими библиотеками и модулями Python. Одно из частных применений — создание статистических моделей и анализ временных рядов.

Дополнительные материалы:

1. Introduction — statsmodels

2. DevDocs — Statsmodels documentation
3. Использование статистических методов для анализа временных рядов

3. Визуализация

Библиотека `matplotlib`

Библиотека `matplotlib` — популярный инструмент Python для визуализации данных. С ее помощью вы можете строить различные виды диаграмм:

- линейные графики;
- столбчатые диаграммы и гистограммы;
- круговые диаграммы;
- диаграммы рассеяния;
- диаграммы размаха и др.

Библиотека позволяет настроить визуальное оформление графика: оси, сетку, аннотации.

Дополнительные материалы:

1. Matplotlib — Visualization with Python
2. Matplotlib cheatsheets — Visualization with Python
3. 50 оттенков matplotlib — The Master Plots (с полным кодом на Python)

Библиотека `seaborn`

Библиотека визуализации данных `seaborn` основана на другой библиотеке визуализации — `matplotlib`.

`Seaborn` имеет высокоуровневый интерфейс для рисования привлекательных и информативных статистических графиков. Кроме того, библиотека содержит более сложные способы визуализация по сравнению с `matplotlib`.

Дополнительные материалы:

1. seaborn: statistical data visualization — seaborn 0.12.2 documentation
2. Data Visualization Using Python Seaborn | Edureka
3. Как строить красивые графики на Python с Seaborn

Библиотека `plotly`

`Plotly` — это библиотека визуализации данных, которая предоставляет широкий набор инструментов для создания интерактивных графиков, диаграмм и визуальных отчетов. Одним из ключевых преимуществ Plotly является его интерактивность: пользователи могут взаимодействовать с графиками, изменять масштаб, фильтровать данные и получать дополнительную информацию при наведении курсора на элементы графика.

Библиотека поддерживает широкий спектр типов графиков, включая линейные графики, столбчатые диаграммы, круговые диаграммы, точечные графики, тепловые карты и многое другое.

Дополнительные материалы:

1. Plotly Python Graphing Library
2. Plotly Python Tutorial for Machine Learning Specialists (neptune.ai)