Data science и Machine learning

Тема 1 Основные понятия

Data Science — деятельность, связанная с анализом данных и поиском лучших решений на их основе. В Data Science входят методы обработки больших данных (Big Data), интеллектуального анализа данных (Data Mining), статистические методы, методы искусственного интеллекта, в т.ч машинное обучение (Machine Learning).

Big Data – данные большого объема, высокой скорости накопления или изменения и/или разновариантные информационные активы, которые требуют экономически эффективных, инновационных формы обработки данных, которые позволяют получить расширенное понимание информации, способствующее принятию решений и автоматизации процессов.

Data Mining – процесс поиска в сырых необработанных данных интересных, неизвестных, нетривиальных взаимосвязей и полезных знаний, позволяющих интерпретировать и применять результаты для принятия решений в любых сферах человеческой деятельности.

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Различают два типа обучения. Обучение по прецедентам, или индуктивное обучение, основано на выявлении общих закономерностей по частным эмпирическим данным. Дедуктивное обучение предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний.

Пять основных этапов в работе с данными

1. Сбор. Поиск каналов, где можно собирать данные, и выбор методов их получения.
2. Проверка. Валидация, нивелирование аномалий, которые не влияют на результат и мешают дальнейшему анализу.
3. Анализ. Изучение данных, подтверждение предположений.
4. Визуализация. Представление информации в понятном для восприятия виде: графики, диаграммы.
5. Реакция. Принятие решений на основе данных. Например, изменение маркетинговой стратегии, увеличение бюджета компании.

Тема 2 Модели машинного обучения

Линейная регрессия – это алгоритм, машинного обучения который моделирует отношение между входной переменной (x) и выходной (y), выражается в виде уравнения:

|  |  |
| --- | --- |
|  |  |

где – точка пересечения;

– наклон линии.

Линейная регрессия применяется со скалярными переменными. Целью линейной регрессии – максимально близкое расположение к результирующей линии. На рисунке 1 представлено визуальное отображение линейной регрессии.

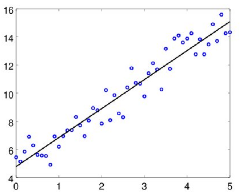


Рисунок 1 – Линейная регрессия

Логистическая регрессия – применяется, когда зависимая переменная является непрерывной, а характер линии линейный. Данный метод подходит для двоичных классификаций. Для решения логистической регрессии применяется регрессионное уравнение.

|  |  |
| --- | --- |
|  |  |

где – вероятность события;

– основание натуральных логарифмов;

– уравнение регрессии.

Логистическая регрессия использует для своего описания S-образную кривую, которая принимает значения от 0 до 1. На рисунке 2 представлено визуальное изображение логистической регрессии.

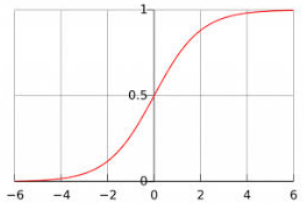


Рисунок 2 – Логистическая регрессия

K-ближайших соседей – данный метод использует весь набор данных в качестве учебника, а не разделяет их. Когда в набор поступают новые данные, алгоритм просматривает весь набор и подбирает K-ближайших соседей, наиболее похожих на новую запись, а затем выводит средний результат для задач регрессии, и наиболее частый класс для задач классификации.

Сходство между соседями определяется с помощью таких уравнений, как евклидово расстояние

|  |  |
| --- | --- |
|  |  |

Метод K-ближайших соседей применяется в статистике и распознавании образов. На рисунке 4 представлено графическое отображение метода K-ближайших соседей.

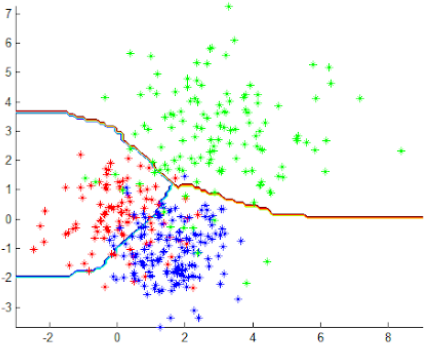


Рисунок 4 – Метод K-ближайших соседей

Метод опорных векторов – это контролируемый метод обучения, который используется и для задач классификации и для задач регрессии. Метод заключается в создании линии или гиперплоскости, которая разделяет данные на несколько плоскостей. На рисунке 5 показан метод опорных элементов.

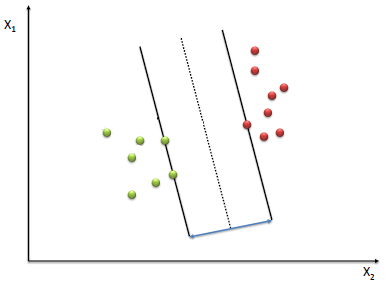


Рисунок 5 – Метод опорных векторов

Наивный байесовский классификатор – представляет набор алгоритмов классификации, реализованных на теореме Байеса

|  |  |
| --- | --- |
|  |  |

где – вероятность гипотезы h с учетом данных d;

–вероятность данных d учитывая гипотезу h верной;

– вероятность того, что гипотеза h верна;

– вероятность данных.

Данный метод предполагает, что все переменные не зависят друг от друга, что является наивным предположением. Наивный байесовский метод применяется для прогнозирования вероятности различных классов на основе атрибутов.

Анализ основных компонентов (PCA) – используется для упрощения изучения и визуального представления данных при помощи сокращения переменных. Анализ основных компонентов использует ортогональное преобразование, которое преобразует набор зависимых данных, в набор не зависимых данных. Данный метод не учитывает информацию об атрибутах. Чем выше разброс атрибутов, тем выше уровень разделения классов и меньше размерность. На рисунке 6 показан пример сокращения данных.

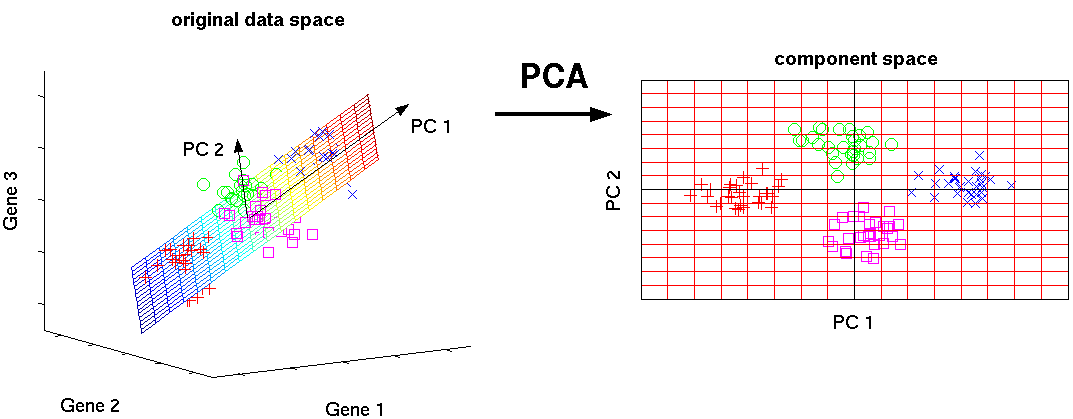


Рисунок 6 - Анализ основных компонентов

Тема 3 Python Библиотеки для анализа данных

Pandas — это библиотека Python для обработки и анализа структурированных данных. Pandas берет данные в файле CSV или TSV или базу данных SQL и создает объект Python со строками и столбцами, который называется фреймом данных. Фрейм данных очень похож на таблицу в статистическом программном обеспечении, скажем, в Excel или SPSS.

Данная библиотека может:

1. Индексирование, манипулирование, переименование, сортировка, объединение фрейма данных;

2. Обновить, добавить, удалить столбцы из фрейма данных;

3. Восстановить недостающие файлы, обработать недостающие данные или NAN.

NumPy - один из самых фундаментальных пакетов в Python - универсальный пакет для обработки массивов. Он предоставляет высокопроизводительные объекты многомерных массивов и инструменты для работы с массивами. NumPy - это эффективный контейнер универсальных многомерных данных.

Данная библиотека выполняет:

1.Основные операции с массивами: добавление, умножение, срез, выравнивание, изменение формы, индексирование массивов;

2. Расширенные операции с массивами: стековые массивы, разбиение на секции, широковещательные массивы;

3. Работа с DateTime или линейной алгеброй.

Scikit Learn, представляет собой надежную библиотеку машинного обучения для Python. Он включает в себя алгоритмы ML, такие как SVM, random forests, k-means кластеризацию и многие другие.

Scikit Learn может:

1. Классификация: обнаружение спама, распознавание изображений;

2. Кластеризация: воздействия лекарственных препаратов, цена акций;

3. Регрессия: сегментация клиентов, группировка результатов эксперимента;

4. Уменьшение размерности: визуализация, повышенная эффективность;

5. Выбор модели: повышенная точность благодаря настройке параметров.

Keras - это высокоуровневый API TensorFlow для создания и обучения кода глубоких нейронных сетей. С Keras статистическое моделирование, работа с изображениями и текстом намного легче с упрощенным кодированием для глубокого обучения.

Тема 4 Визуализация

Визуализация данных – это процесс представления данных в графическом формате, главной функцией визуализации является преобразование данных в формат более удобный для восприятия. Информация, представленная в графическом виде, более запоминающаяся и вызывает больший интерес.

Matplotlib - самая популярная библиотека Python для визуализации данных. Ее можно использовать в оболочках Python и IPython, скриптах Python, серверах веб-приложений и т.д. Библиотеку применяют для различных целей, таких как создание графиков, гистограмм, спектров мощности, ствол-лист диаграмм, круговых диаграмм и многого другого.

Seaborn использует все возможности, которые дает Matplotlib, для создания красивых графиков при помощи всего нескольких строк кода. Главное, что отличает Seaborn, это дефолтные стили и цветовые палитры: они более эстетичны и современны. Seaborn построен на базе Matplotlib.

Plotly — это библиотека для создания интерактивных визуализаций и управления ими. В этой написанной на JavaScript библиотеке есть множество встроенных приложений для машинного обучения и анализа данных, что упрощает реализацию и визуализацию разных алгоритмов.