**АНАЛИЗ ДАННЫХ. УПРАВЛЕНИЕ ДАННЫМИ**

**Базы анализа данных. Методы сбора, классификации и прогнозирования. Деревья решений. Обработка больших объемов данных. Методы и этапы интеллектуального анализа данных. Задачи интеллектуального анализа данных. Визуализация данных.**

1. Введение в данные
2. Процесс обработки и анализа данных
3. Сбор данных

3.1 Методы сбора данных

1. Обработка данных
2. Визуализация данных

5.1 Библиотеки Python для построения графиков

1. Анализ данных

6.1 Классификация

6.2 Регрессия

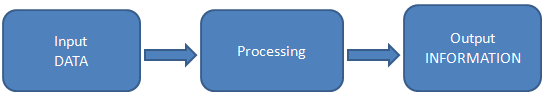
6.3 Дерево решений

1. Управление данными

**Введение в данные**

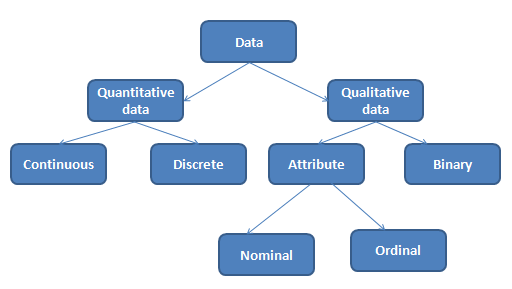
Данные — это совокупность необработанных фактов и чисел, которые необходимо преобразовать для получения информации или сбора сведений. Данные используются в качестве «ресурса» компьютерной системы и являются базовым блоком, из которого составляется информация.

Информация в целом — это лишь обработанные данные, под конкретной же информацией зачастую понимается обработанная, организованная и структурированная форма данных.



|  |  |
| --- | --- |
| Input DATA | Входные ДАННЫЕ |
| Processing | Обработка |
| Output INFORMATION | Выходная ИНФОРМАЦИЯ |

Существует два типа данных: количественные и качественные. Качественные данные — это данные, которые не поддаются измерению. Они включают в себя чувства, личный опыт, отношения и намерения. Эти данные получают из таких источников, как фокус-группы, опросы, повествовательные тексты и отчеты. Количественные данные — это числовые, поддающиеся измерению данные, которые позволяют объективно оценивать ситуации, чтобы сравнить их между собой, а также отслеживать изменение условий в рамках одной ситуации с течением времени.



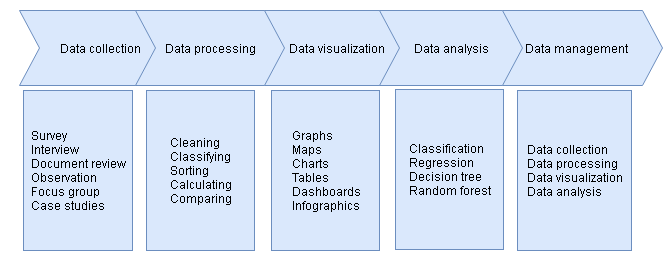
|  |  |
| --- | --- |
| Data | Данные |
| Quantitative Data | Количественные данные |
| Continuous | Непрерывные |
| Discrete | Дискретные |
| Attribute | Атрибутивные |
| Binary | Двоичные |
| Nominal | Номинальные |
| Ordinal | Порядковые |

Примеры типов данных

|  |  |  |
| --- | --- | --- |
| Тип | Определение | Пример |
| Непрерывные | Данные, которые можно измерить | Возраст, рост, вес человека  Деньги  Время |
| Дискретные | Данные, основанные на подсчете | Количество студентов  Количество животных в зоопарке |
| Номинальные | Размеченные и именованные данные, которые можно разделить на несколько непересекающихся групп | Казахский, русский, английский  Мужчины, женщины  Белый, черный, красный |
| Порядковые | Данные находятся в некотором порядке по отношению друг к другу в упорядоченном виде | Очень хорошо, хорошо, плохо  Удовлетворенный, неудовлетворенный  Счастливый, несчастный, нормальный |
| Двоичные | Данные, которые могут принимать только два возможных значения | Да, нет  Успех, неудача  Правда, ложь |

**Процесс обработки и анализа данных**

Процесс обработки и анализа данных условно включает в себя пять этапов: сбор, обработка, визуализация, анализ и управление. Чтобы начать работать с данными, во-первых, мы должны их собрать. Это включает в себя идентификацию и доступ ко всем рассматриваемым данным. Затем нам нужно изучить данные и обработать их. Первым этапом в подготовке данных является фактический просмотр данных с целью понять их происхождение, значение, качество и формат. Процесс обработки включает в себя очистку данных, фильтрацию данных или разбивку их на подмножества, создание данных, которые могут быть считаны и распознаны программами, например, моделирование из необработанных данных более определенной модели данных, или упаковку данных в определенный формат. Следующий этап – визуализация результатов. Затем подготовленные данные передаются на этап анализа, который включает в себя выбор аналитических методов, построение модели данных и анализ результатов. После анализа вы можете управлять данными по своему усмотрению.



|  |  |
| --- | --- |
| Data collection | Сбор данных |
| Data processing | Обработка данных |
| Data visualization | Визуализация данных |
| Data analysis | Анализ данных |
| Data management | Управление данными |
| Survey Interview  Document review  Observation  Focus group  Case studies | Опрос Интервью  Пересмотр документов  Наблюдение  Фокус-группа  Анализ конкретных примеров |
| Cleaning  Classifying  Sorting  Calculating  Comparing | Очистка  Классификация  Сортировка  Расчет  Сравнение |
| Graphs  Maps  Charts  Tables  Dashboards  Infographics | Диаграммы  Карты  Графики  Таблицы  Информационные панели  Инфографика |
| Classification  Regression  Decision tree  Random forest | Классификация  Регрессия  Дерево решений  «Случайный лес» |
| Data collection  Data processing  Data visualization  Data analysis | Сбор данных  Обработка данных  Визуализация данных  Анализ данных |

**Сбор данных**

Сбор данных представляет собой систематический подход к сбору и измерению информации из различных источников для получения полной и точной картины рассматриваемой области. Существует множество методов сбора информации, а также большое разнообразие источников информации. Ниже перечислены некоторые из наиболее распространенных методов сбора данных:

1. *Опрос, анкетирование*. Опрос — это выбор наиболее подходящей информации или сведений в качестве ответа на вопросы автора.

2. *Интервью*. Интервью — это отличный способ собрать подробную информацию у одного человека или нескольких людей. Это полезно, если необходимо получить экспертное мнение по рассматриваемому вопросу или поговорить с кем-то хорошо осведомленным в этой теме.

3. *Пересмотр документов*. Автор может собирать некоторые важные данные из различных баз данных, протоколов собраний, отчетов, журналов посещаемости, финансовой документации, информационных бюллетеней и т. д.

4. *Наблюдение*. Автор может проводить прямые наблюдения на местах, с разных позиций, это помогает понять поведение систем в реальной обстановке на местах.

5. *Фокус-группа*. Упрощенное групповое интервью с людьми, которые имеют между собой что-то общее.

6. *Анализ конкретных примеров*. В этом методе рассматриваются специфические и любопытные случаи.

**Обработка данных**

В процессе обработки данные собирают и преобразуют в полезную информацию. На этом этапе необработанные данные преобразуют в читаемый формат, который можно интерпретировать, анализировать и использовать для различных целей. Это подразумевает фактические методы обработки данных, такие как очистка, классификация, сортировка, расчет, суммирование, сравнение и т. д., которые превращают данные в информацию.

Специалисты обычно тратят много времени на этап очистки данных. Он включает:

1. Удаление ненужных наблюдений

2. Исправление структурных ошибок

3. Фильтрация экстремальных значений

4. Решение проблемы недостающих данных

**Визуализация данных**

Лучшие данные - это те данные, которые каждый может увидеть и понять. Поэтому специалисты по обработке и анализу данных предпочитают создавать и представлять наиболее полные и понятные визуализации. Это не всегда просто: требуется найти данные, просмотреть их, очистить, отфильтровать, а затем использовать правильный инструмент для ее визуализации.

Визуализация данных — это графическое представление информации. Она выполняется с помощью визуальных элементов, таких как диаграммы, графики и карты, которые облегчают просмотр и понимание динамики, экстремальных значений и закономерностей в данных.

Python является одним из самых популярных языков программирования, используемых при анализе и обработке данных, и поэтому располагает большим количеством полезных библиотек, разработанных его широким сообществом разработчиков открытого ПО. Python предлагает несколько библиотек построения графиков для визуализации данных, которые включают в себя большое количество различных функций. Это matplotlib, seaborn, ggplot, plotly, pandas visualization.

**Анализ данных**

Анализ данных - это процесс оценки данных с использованием аналитических и логических методов для изучения всех компонентов предоставленных данных. Более продвинутые методики анализа данных включают интеллектуальный анализ данных, который предусматривает целенаправленную сортировку больших массивов данных с целью выявления динамики, закономерностей и взаимосвязей.

Интеллектуальный анализ данных — это этап анализа, который включает методы на пересечении машинного обучения, статистики и систем баз данных. Он включает в себя задачи предварительной обработки, такие как извлечение данных, очистка данных, слияние данных, сокращение объема данных и формирование признаков, а также этапы предварительной обработки, такие как интерпретация шаблонов и моделей, подтверждение и генерация гипотез.

Машинное обучение - это метод анализа данных, который автоматизирует построение аналитических моделей. Оно предоставляет системам возможность автоматически учиться и совершенствоваться с опытом, без программирования в явном виде. Пять основных шагов, используемых при разработке алгоритмов машинного обучения, сводятся к следующему:

1. Выбор признаков и сбор обучающих выборок

2. Выбор метрики производительности

3. Выбор классификатора и алгоритма оптимизации

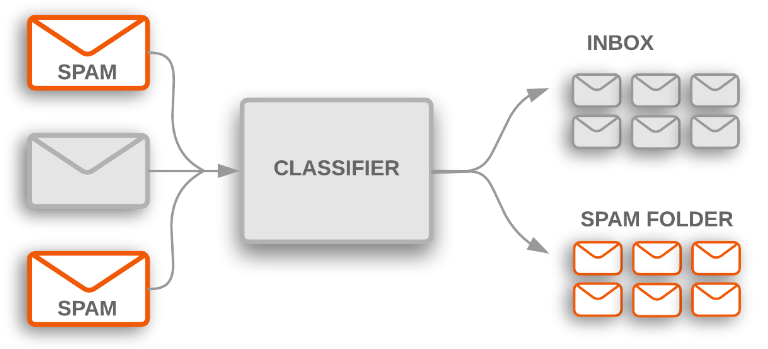
4. Оценка производительности модели

5. Настройка алгоритма

Существует два типа машинного обучения: контролируемое и неконтролируемое обучение.

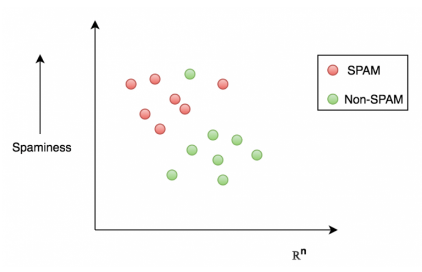
Основная цель контролируемого обучения — изучить по размеченным обучающим данным модель, которая позволит делать прогнозы о невидимых или последующих данных.

Рассматривая пример фильтрации нежелательной электронной почты (спама), мы обучаем контролируемый алгоритм машинного обучения модели на собрании размеченных писем, чтобы он мог правильно предсказывать, следует ли новое письмо помещать в папку «Входящие» или папку «Спам». Контролируемая задача обучения с использованием дискретных меток классов, как например в предыдущем примере фильтрации нежелательной почты, также называется классификацией.



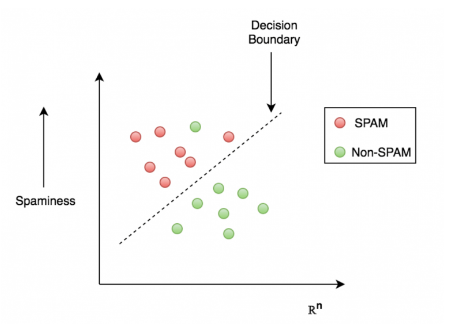
|  |  |
| --- | --- |
| SPAM | СПАМ |
| CLASSIFIER | КЛАССИФИКАТОР |
| INBOX | ПАПКА «ВХОДЯЩИЕ» |
| SPAM FOLDER | ПАПКА «СПАМ» |

Как показано на рисунке, почтовая программа использует текстовый классификатор, чтобы определить, помещать ли входящее письмо в папку «Входящие» или отфильтровать в папку «Спам». Этот пример представляет собой типичную задачу двоичной классификации, где алгоритмы машинного обучения обучаются набору правил, чтобы различать два возможных класса: «спам» и «не-спам» письма.



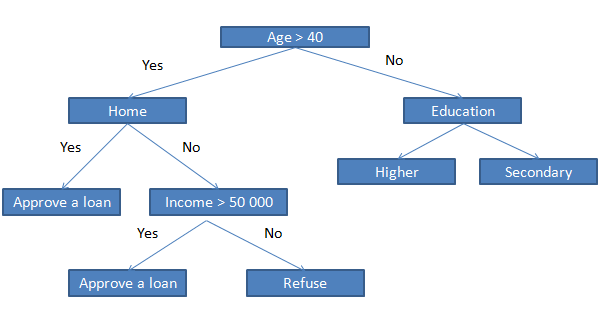
|  |  |
| --- | --- |
| Spaminess | Спамность |
| SPAM | СПАМ |
| Non-SPAM | Не-СПАМ |

Другая категория контролируемого обучения - регрессия, которая пытается оценить или предсказать целевое значение для числовых переменных. В регрессии задается ряд прогностических (объясняющих) переменных и непрерывная (целевая) переменная, и нам необходимо найти связь между этими переменными, которая позволит нам предсказывать результат.



|  |  |
| --- | --- |
| Spaminess | Спамность |
| Decision Boundary | Граница принятия решений |
| SPAM | СПАМ |
| Non-SPAM | Не-СПАМ |

Одним из алгоритмов классификации в машинном обучении является дерево решений. Дерево решений - это модель, которая разбивает наши данные, принимая решение, основанное на постановке ряда вопросов.



|  |  |
| --- | --- |
| Age >40 | Возраст > 40 лет |
| Yes | Да |
| No | Нет |
| Home | Дом |
| Education | Образование |
| Approve a loan | Одобрить кредит |
| Income > 50000 | Доход > 50000 |
| Higher | Высшее |
| Secondary | Среднее |
| Refuse | Отказать |

Исходя из признаков в нашей обучающей выборке, модель на основе дерева решений обучается ряду вопросов, по которым она делает заключение о метке класса для образца. На предыдущем рисунке показана идея предоставления банковского кредита в виде дерева решений, основанного на категориальных переменных и действительных числах. При использовании алгоритма принятия решений, мы начинаем с корня дерева и разделяем данные по признакам, после чего мы можем повторять эту процедуру разделения на каждом дочернем узле, пока не дойдем до конечных листьев. Это означает, что образцы на каждом узле принадлежат к одному классу.

Кластеризация использует неконтролируемое обучение для группировки данных в отдельные кластеры или сегменты. Другими словами, кластеризация пытается найти в данных естественную группировку.

**Управление данными**

Управление данными - это процесс, который включает в себя все процессы работы с данными, используемые для удовлетворения текущих потребностей жизненного цикла данных: сбор данных, обработка данных, визуализация данных, анализ данных и т. д.