Лекция №1

Анализ данных - область математики и информатики, занимающаяся построением и исследованием общих матемематических методов и вычислительных алгоритмов, извлечение знаний из эксперементнтальных в широком смысле знаний.

Анализ данных - процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятие решений.

[в узком смысле] - обработка информации после её сбора.

Выделяют несколько **типов** анализа данных:

1. Статистический анализ данных;
2. Интелектуальный анализ данных - набор методов из различных областей познания используемые для анализа данных;

Анализ данных стоит на стыке наук:

Данные - это совокупность различных объективных фактов.

Информация - это данные, сопровждающаяся смысловой нагрузкой, помешённые в некоторый контекст. Данные, как либо оцениваемые приёмником информации. Как правило, получение информации связывают с уменьшением неопределённости существующего выбора. Также информация - это ответ на какой заданный вопрос.

Знание - это комбинация опыта, ценностей, контексной информации, экспертных оценок, которая дает общие рамки для оценки и инкорпорирования нового опыта и информации. Знание существует в сознании тех, кто знает.

Объект - набор атрибутов;

Атрибут - свойство характеризующий объект;

Генеральная совокупность (population) - это вся совокупность изучаемых объектов, интересующая исследователя.

Выборка - часть генеральной совокупности, определённым способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.

Существует два **способа** составления выборки:

- Повторности;
- Рандомизация;

Гипотеза - это предположение относительно параметров совокупности объектов, которое должно быть проверено на её части.

Гитотеза - это частино обоснованная закономерность знаний, служущая, либо для связи между различными эмпирическими фактами, либо для объяснения какого-то факта или группы фактов.

При выдвижении гипотезы в выборке появляются **зависимые** и **независимые** переменные.

Типы наборов данных, шкалы используемые в наборах данных

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно опредённому правилу.

Шкала - это правило в соотвествие с которым объектам присваиваются числа.

Шкала:

- Категориальная
  - Дихтоническую
  - Порядковую
  - Номинальная
- Количественная
  - Интревальную
  - Отношений
  - Непрерывную

Этапы:

1. Постановка задачи
   1.1. Определение цели исследования
   1.2. Определения состава данных
   1.3. Сбор данных
   1.4. Выбор средства анализа данных
   1.5. Формализация данных
2. Ввод данных в обработку
   2.1. Ввод данных в памать компьютера
   2.2. Работа с архивом данных
   2.3. Формирование задания обработки
3. Качественный анализ
   3.1. Определить простейшие характеристики данных
   3.2. Совершить визуализацию данных
   3.3. Провести анализ структуры данных
4. Количественное описание данных
   4.1. Выбор модели данных
   4.2. Выполнение обработки
5. Итерпретация результатов
   5.1. Анализ результатов
   5.2. Выводы на основе проведённого анализа данных

Этап 1. Постановка задачи.

Является определяющим этапом от которого зависит весь ход анализа, он начинается со стадии формулировки цели всего исследования ради достижения которой и предпринимаются сбор и обработка данных.
Исходя из цели, определяется состав данных, которые необходимо собрать.
Одна из типичных ошибок исследований - это сначала сбор данных, а потом формулируют цель.
Объекты - люди, изделия, услуги и т.д. Таблицей такого типа называют **ТЭД**
На стадии выбора определяются **инструменты** для анализа.
На стадии формализации собранных данных ТЭД необходимо принять определённых вид
Этап 2. Ввод данных в обработку
Суть: объекты попадают в компьютерную память, затем попадают в архив и затем выбирается определённая часть из данныех
Этап 3. Качественный анализ
Первая задача. Попытка предстваить собранные данные в визуальной форме.
Информативное описание данных. Содержательная постановка задачи. Заключается в том, чтобы найти небольшое число наиболее важных свойств, характеристик, особенностей исследования.
Формальная постановка задачи заключается в том, чтобы устранить дублирующие признаки или построить новые признаки описывающие эти данные
Вторая задача. Группировка (классификация) объектов. Необходимо найти группы с общими свойствами.
Третья задача. Исследование зависимости одного признака от остальных
Содержательная постановка: описать взаимосвязь или зависимость избранного свойства от остальных
Формальная постановка: найти функциональную зависимость изменения целового признака при изменении других признаков.
Задачи:
Распознование образов
Содержательная постановка: найти правило, пользуясь которым можно определить пренадлежность одного объекта.
Формальная постановка: найти к какой группе точек относятся заданные объекты
Таким образом, на этапе качественного анализа, объектом исследования является структура данных. А результатом является информация о классе модели, которым можно описать являение
Этап 4. Количественное описание данных.
Ведётся поиск параметров модели созданных на предыдущем этапе.
Сопоставительный анализ позволяет отбирать лучшие варианты, которые имеют право на существование не только как формальные результаты, но и как содержательно значимая информация.
Этап 5. Итерпретация результатов
Принимается решение об анализе данных. Идёт прикращение обработки, так как поставленные результаты достигнуты, либо использовать другие обработки, если данные не удовлетворяют целям (в этом случае анализ начинается занаво).

Примеры анализа данных

Анализ тональности текста

Какой эмоциональный окрас имеет текст? (позитивный, нейтральный, негативный).

Ещё задачи:

- Какой будет спрос на товар в следующем месяце?
- Вернёт ли клиент кредит?
- Кто победит в онлайн игре?

Особенность таких задач:

- Везде - очень сложные неявные зависимости
- Нельзя выразить формулой
- Но есть некоторое число примеров
  - Тесты с известным окрасом
- Будем приближать зависимоти используя готовые примеры

Анализ данных и машинное обучение - это про то как установить сложные зависимости по конечному числу примеров.

Примеры задач

Сеть ресторанов

Хотим открыть еще один

Несколько вариантов рамещения

x – объект, sample - для чего хотим делать предсказания

X – простарнство, где мы можем расположить рестораны

y – прибыль в течение года работы

Y – все возможные ответы

$X(x_i, y_i)$ – обучающая выборка

$x\left(x^1, \ldots, x^d\right)$ – признаковое описание

Признаки:

- Про демографию:
  - Средний возраст жителей
  - Динамика количества жителей
- Про недвижимость:
  - Средняя стоимость квадратного метра жилья
  - Количество школ, банков
  - Расстояние до ближайших конкурентов

$a(x)$ – алгоритм, функция предсказывающая ответ

Линейная модель $a(x) = w_1x^1 + \ldots + w_nx^n$

$a(x) = 0$ – не принесёт никакой прибыли

Функция потерь - это мера корректности ответа алгоритма

Квадратиное отклонение: $(a(x) - y)^2$

Функционал качества - мера качества работы алгоритм (сред. квадрат ошибка)

Функционал качества должен соотвествовать бизнес требованиям.
Обучение алгоритма
Семейство алгоритмов: $\mathcal{A}\left(w_1 x^1 + \ldots + w_n x^n\right)$

Классы задач

- Прогнозирование
- Классификация
- Кластеризация
- Ассоциация
- Последовательность
- Визуализация данных
- Анализ отклонений
- ...