Введение в Data Science
Занятие 0. Знакомство

Николай Анохин Михаил Фирулик

4 марта 2014 г.

Ваши преподаватели

- Михаил Фирулик (m.firulik@corp.mail.ru / +7 916 730-97-66)
  - руководитель отдела анализа данных в Mail.Ru Group
  - многолетний опыт интеллектуального анализа данных
- Николай Анохин (n.anokhin@corp.mail.ru / +7 903 111-44-60)
  - программист-исследователь в Mail.Ru Group
  - более трех лет работы в области Data Mining

Права и обязанности

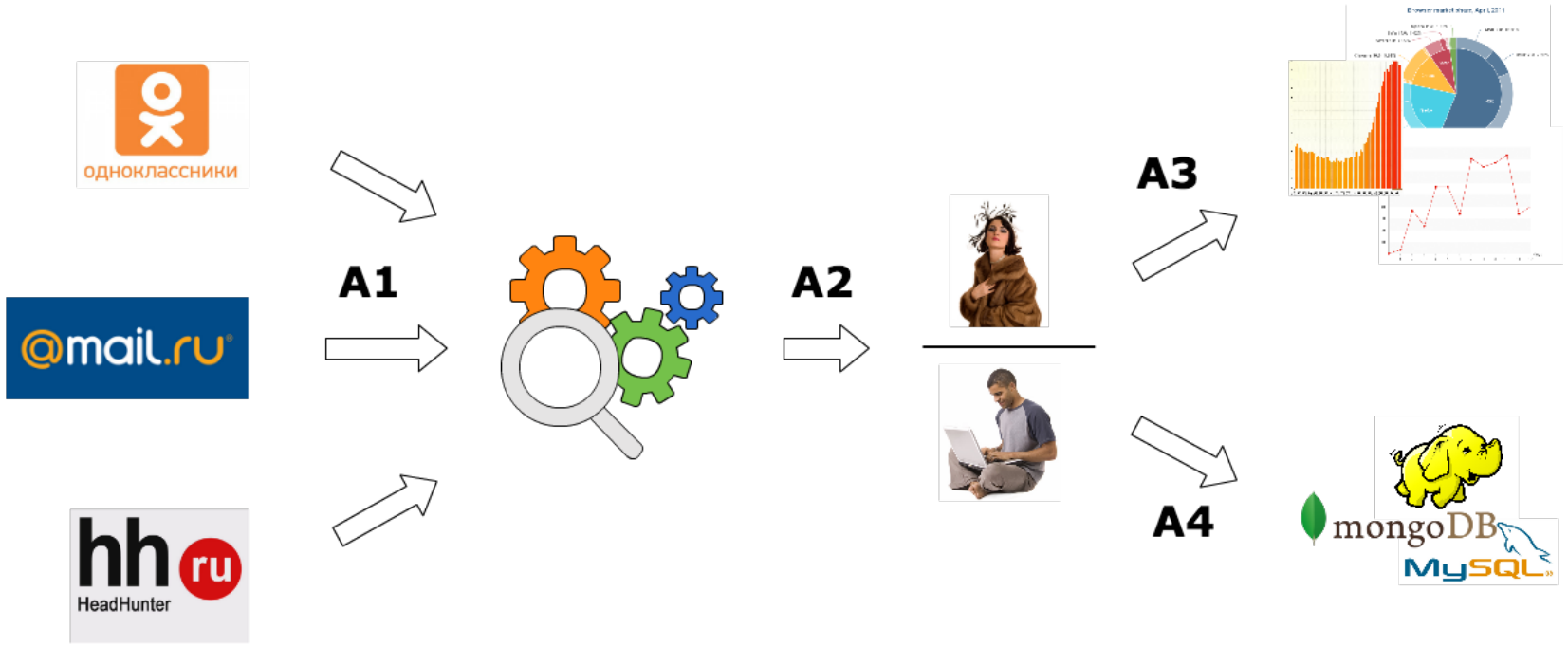
- можно
  - задавать вопросы преподавателю
  - выносить идеи на общее обсуждение
  - входить и выходить, не мешая коллегам
- не можно
  - нарушать порядок на занятии
  - разговаривать по телефону в аудитории
- общение
  - с преподавателем на “Вы”
  - с коллегами – как удобно

Ваши правила?

План занятия

Одна (типичная) задача

Рекламная компания магазина зимней одежды: определить аудиторию



A1 (data) acquisition

A2 (data) analysis

A3 (data) archiving

A4 (data) architecture

Что делать?

- Разобраться в предметной области
- Общаться с пользователями данных
- Понимать “Big Picture”
- Изучить представление данных
- Произвести подготовку и анализ данных
- Визуализировать результат
- Учитывать этические соображения

Мы бы хотели, чтобы вы

1. получили практический опыт решения задач Data Mining
2. познакомились с инструментарием
3. поиграли и получили удовольствие

Что необходимо повторить

1. Линейная алгебра
2. Теория вероятностей
3. Алгоритмы и структуры данных

Модули курса

1. Задачи классификации (6 занятий)
2. Задачи кластеризации (3 занятия)
3. Мета-алгоритмы (4 занятия)

Модуль 1. Задачи классификации

Задача Разработать алгоритм, позволяющий определить класс произвольного объекта из некоторго множества

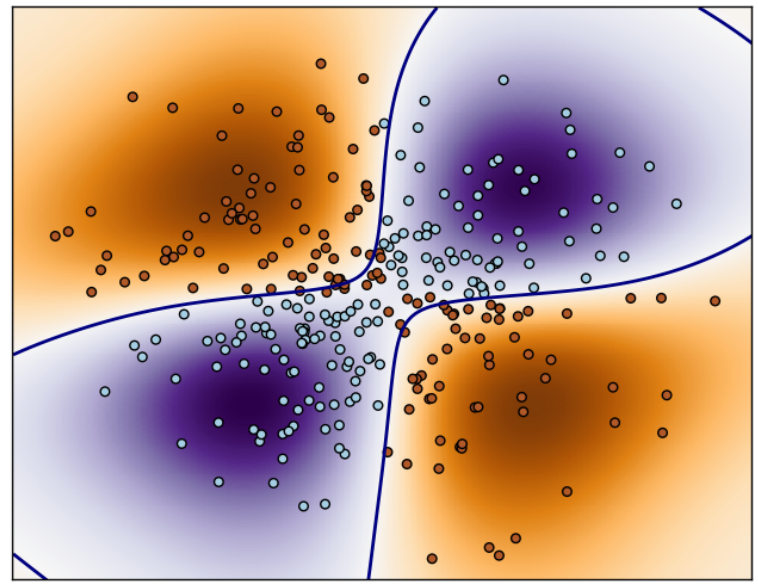
- Каждый объект заданного множества принадлежит классу из некоторого набора
- Дана *обучающая выборка*, в которой для каждого объекта известен класс

Примеры

- Определение спама
- Кредитный скоринг
- Распознавание лиц

Модуль 1. Содержание

1. Задача классификации и регрессии. Метрики ошибок
2. Линейная и логистическая регрессия
3. Решающие деревья
4. Байесовские алгоритмы
5. Метод опорных векторов



Задача модуля. Предсказание пола и возраста пользователей популярных социальных сервисов.

Модуль 2. Задачи кластеризации

Задача Разбить выборку объектов на подмножества (кластеры)

- Объекты внутри одного кластера должны быть похожи
- Объекты из разных кластеров должны существенно отличаться

Примеры

- Определение сообществ
- Сегментация изображений
- Исследование рынка

Модуль 2. Содержание

1. Задача кластеризации. Метрики качества
2. EM-алгоритм
3. Различные алгоритмы кластеризации

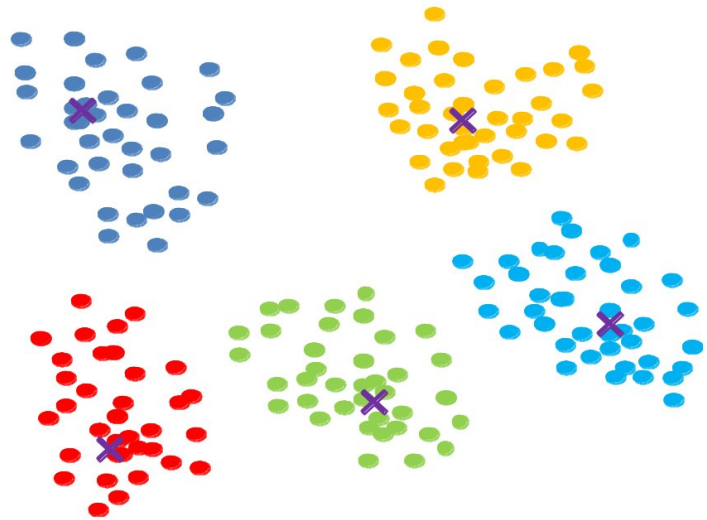


Fig. 13. Exemplary K-Means result

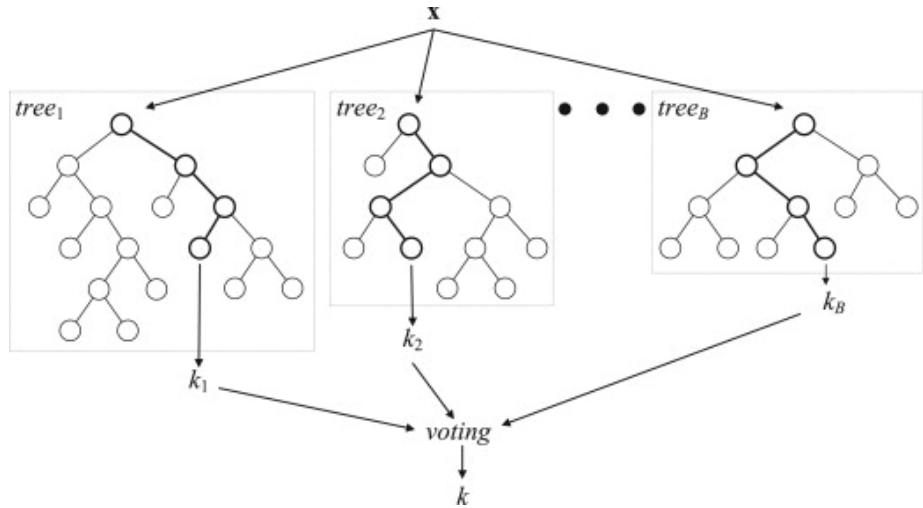
Задача модуля. Разбиение на категории товаров, предлагаемых крупными интернет-магазинами.

Модуль 3. Мета-алгоритмы

- Какие факторы выбрать для решения задачи?
- Что, если алгоритмы не дают необходимого качества?
- Что, если данные не помещаются в памяти?

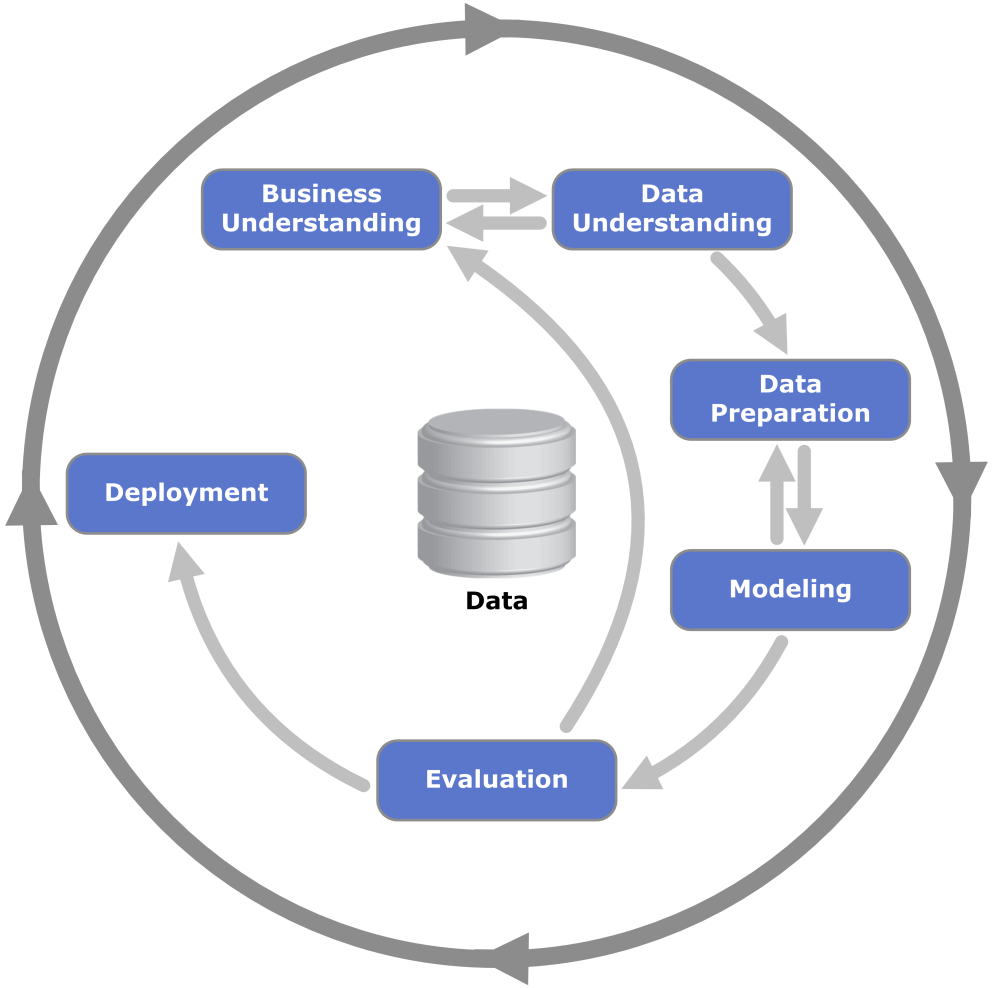
Модуль 3. Содержание

1. Метод ансамблей
2. Предобработка данных и выбор факторов
3. Вычислительная модель MapReduce



Задача модуля. Классификация пользователей интернета с использованием реальных данных сервисов Mail.Ru.

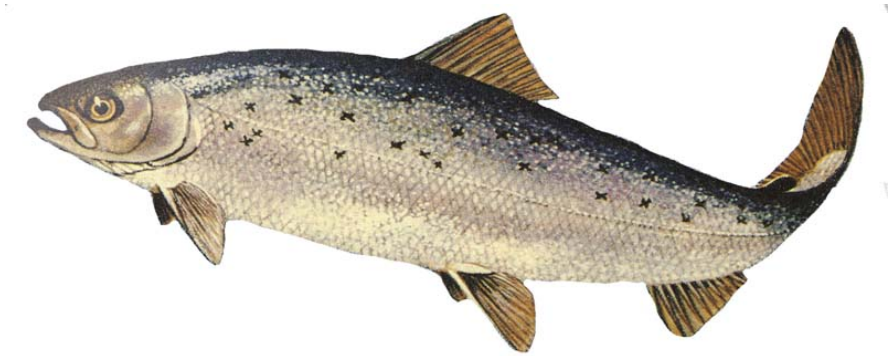
CRISP-DM



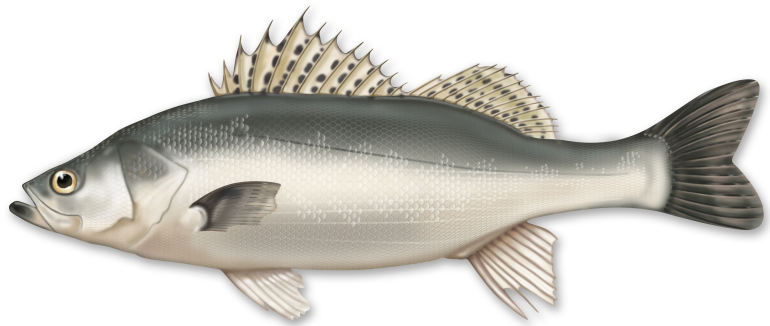
SPSS, Teradata, Daimler AG, NCR Corporation, OHRA

Business understanding

На рыболовном предприятии автоматизируем сортировку улова

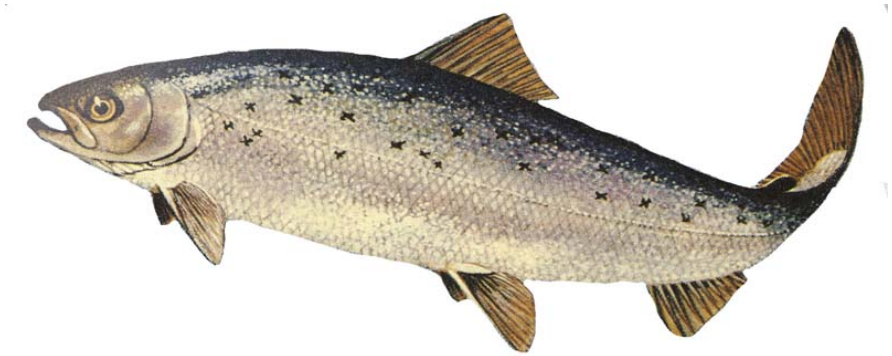


VS

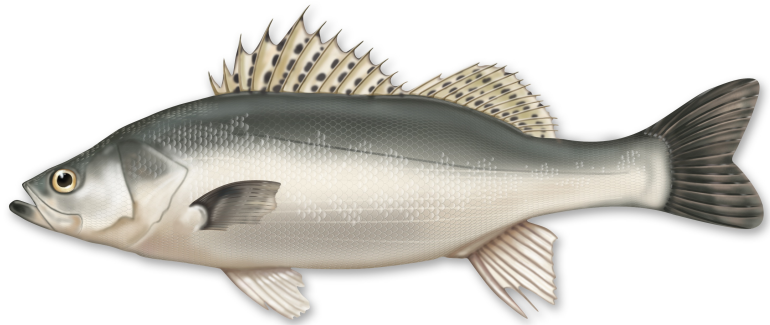


Data understanding 1

Какие факторы будем использовать?



VS



Data understanding 2

X – множество объектов. Фактор $f_i : X \to F_i$

- Бинарные/Binary: $F_i = \{true, false\}$ (есть ли пятна, двойной ли плавник)
- Номинальные/Categorical: F_i – конечно (цвет, форма чешуи)
- Порядковый/Ordinal: F_i – конечно, определен порядок (категория возраста, количество плавников)
- Количественный/Numerical: $F_i = R$ (длина, вес)

Data preparation

Эта часть проекта занимает больше всего времени

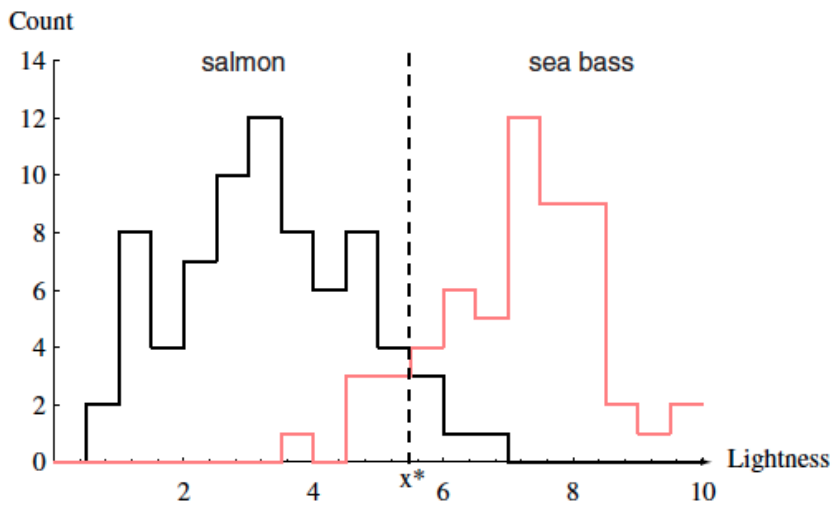
- Удаление шума
- Заполнение отсутствующих значений
- Трансформация факторов
- Выбор факторов
- Использование априорных знаний

Результат. Обучающая выборка, в формате, подходящем для моделирования

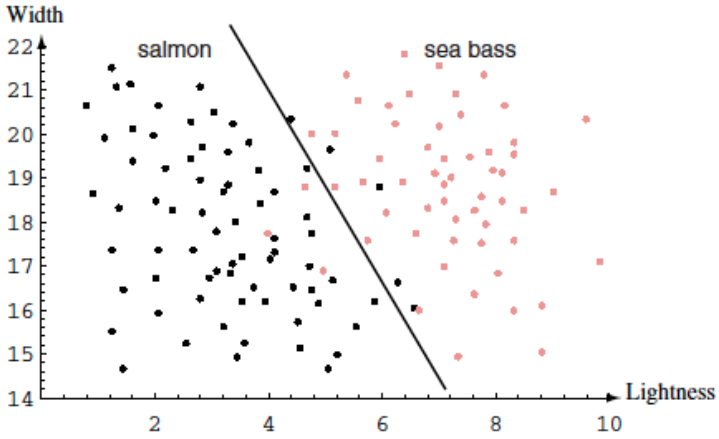
Modeling 1

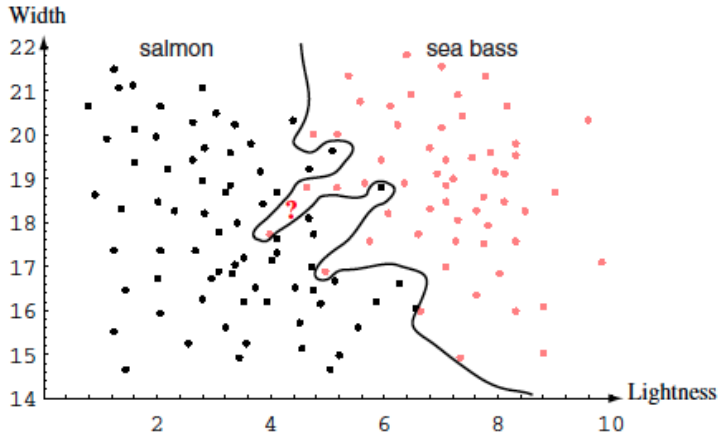
Модель – описание класса, выраженное, как правило, в математической форме. Цель – выбрать удачную модель и ее параметры так, чтобы она наилучшим образом описывала заданный класс.

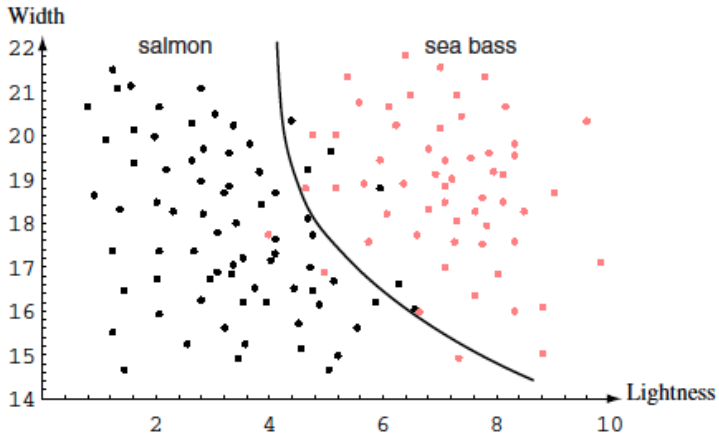
- Статистические модели
- Модели машинного обучения



Modeling 2



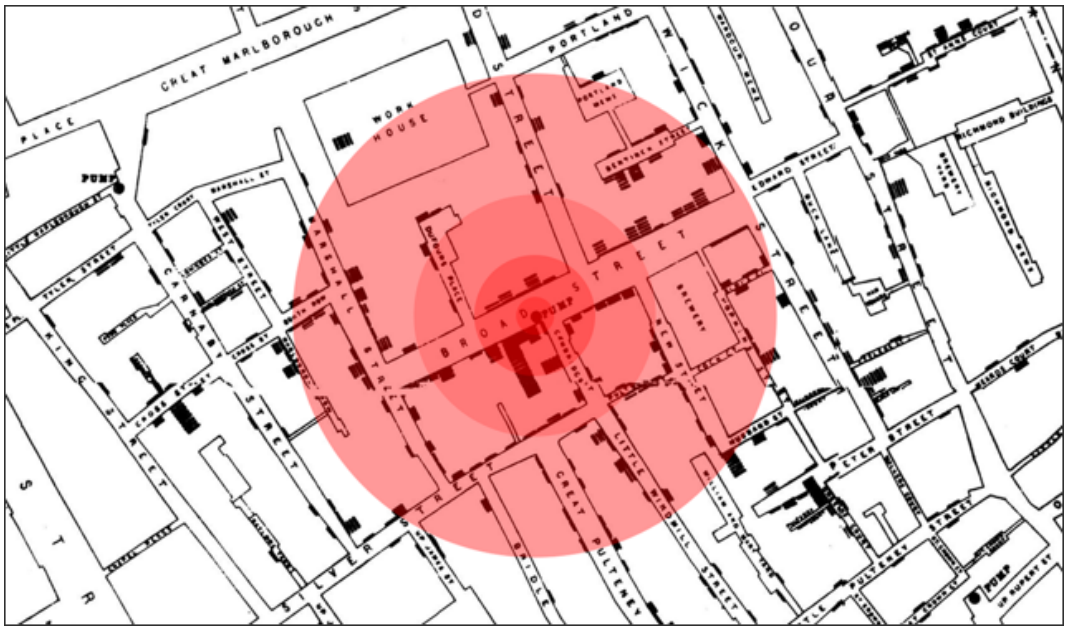




Evaluation & Deployment

- Решает ли выбранная модель задачу достаточно эффективно?
- Удовлетворяет ли модель требованиям бизнеса?
- Что вообще может пойти не так?

1854 г. Эпидемия холеры в Лондоне



Программа TIA

- Наблюдаем 10^9 человек
- Человек в среднем посещает отель раз в 100 дней
- Есть 10^5 отелей на 100 человек каждый
- Проверим посещения за 1000 дней

Вероятность для конкретной пары встретиться в отеле в конкретный день:

$$p_1 = \left(\frac{1}{100}\right)^2 \cdot 10^{-5} = 10^{-9}$$

Всего пар людей

$$n_{pp} = C_2^{10^9} \approx \frac{(10^9)^2}{2} = 5 \cdot 10^{17}$$

а пар дней

$$n_{pd} = C_2^{10^3} \approx \frac{(10^3)^2}{2} = 5 \cdot 10^5$$

Ожидаемое количество “подозрительных” встреч в отелях

$$N = p_1^2 n_{pp} n_{pd} = 250000 >> 10$$

Принцип Бонферрони

Вычислить количество рассматриваемых событий при предположении их полной случайности. Если это количество намного превосходит количество событий, о котором идет речь в задаче, полученные результаты нельзя будет считать достоверными.

Что мы обсудили на сегодняшней лекции?

- Познакомились со стандартным процессом CRISP-DM
- Вспомнили, какие бывают виды факторов
- Узнали, для чего в Data Science используется моделирование
- Разобрались с принципом Бонферрони

Задача 1

Пусть имеется простая обучающая выборка, включающая в себя 4 признака: бинарный f_1, номинальный f_2, порядковый f_3 и количественный f_4.

N	f_1	f_2	f_3	f_4
1	true	A	O1	3.14
2	false	B	O2	2.7
3	true	A	O2	11.0
4	true	C	O1	10.0

Предложенная модель работает только на бинарных признаках. Как преобразовать данную обучающую выборку в нужный формат? А если количественный? А номинальный?

Задача 2

Пусть имеется информация о покупках, совершенных 100 миллионами людей. Кажый из них идет за покупками в среднем 100 раз в год и покупает 10 из 1000 представленных товаров. Предположим, что два злоумышленника покупают одинаковые наборы товаров. Если мы ищем пары людей, купившие одинаковые наборы в течение года, сможем ли мы действительно определить террористов?

Спасибо!

Обратная связь