Какие данные бывают

Интеллектуальный анализ данных

Малютин Евгений Алексеевич

Какие бывают данные

Прежде чем приступать к решению задачи анализа данных, нужно сначала понять что у нас есть и что мы хотим

Объем данных

- От нескольких сотен записей до десятков терабайт
- Нам достаточно одного ноутбука или нужен кластер, например с **Spark**/**Hadoop**?

Хранение данных

- Реляционная база данных (MySQL, SQLite, PostgreSQL, . . .)
- Нереляционная база данных *aka* NoSQL (Cassandra, HBase)
- Может будет достаточно Pandas DataFrame?
- Как их мы будем анализировать? Нужен ли online?

Примеры

- Изображения
- Временные ряды
- Текст
- Сильно разреженные
- Есть отсутствующие значения

Кроме того, возникают вопросы приватности, этики и т. п.

Классификация изображений

- Нам нужно по фотографии определить марку машины
- Задачи классификации изображений сейчас эффективно решаются с помощью сверточных нейронных сетей
- Нужна видеокарта (GPU), на которой вычисления происходят гораздо быстрее

Определение цены квартиры

- Мы должны предсказать рыночную цену квартиры по адресу, числу комнат, общей площади, этажу
- Вероятнее всего данная задача не потребует серьезных вычислительных мощностей

Подбор рекламы для пользователя

- На сайт заходит пользователь
- Нужно подобрать ему рекламу, чтобы оптимизировать вероятность клика
- У нас есть исторические данные, информация о пользователе (ник на форуме, возможно демографические характеристики)
- Подобные задачи требуют существенных вычислительных ресурсов

О данных

А если не примеры?

- понятие данных
- шкалы
- типы наборов данных
- проблемы качества
- этапы предобработки

Что такое данные?

Представление фактов в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes

Шкалы

Теория измерений (С. С. Стивенсон)

- номинальная шкала (**КДП**: биективные преобразования)
- порядковая шкала (**КДП**: строго монотонные преобразования)
- шкала интервалов (**КДП**: преобразования вида $x' = kx + b$)
- шкала отношений (**КДП**: преобразования вида $x' = kx + b$)
- шкала разностей (**КДП**: преобразования вида $x' = x + b$)
- абсолютная шкала (**КДП**: преобразования вида $x' = x$)

Шкалы

И что?

- алгоритм анализа данных должен быть инвариантен относительно КДП исследуемой величины
- алгоритм, применимый к более слабой шкале, применим и к более сильной.

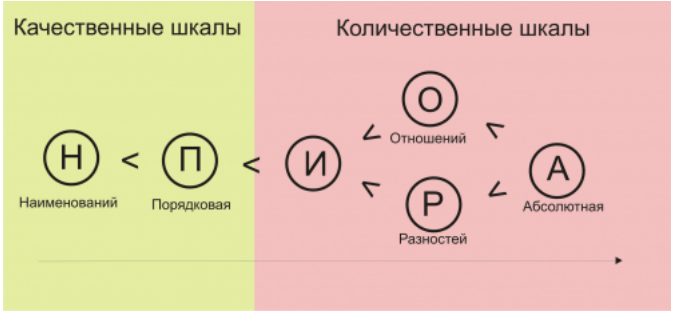


Рис.: Иерархия шкал измерений. Слева - самая слабая шкала, справа - самая сильная

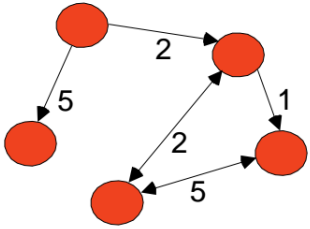
Типы данных

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Табличные

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Транзакционные



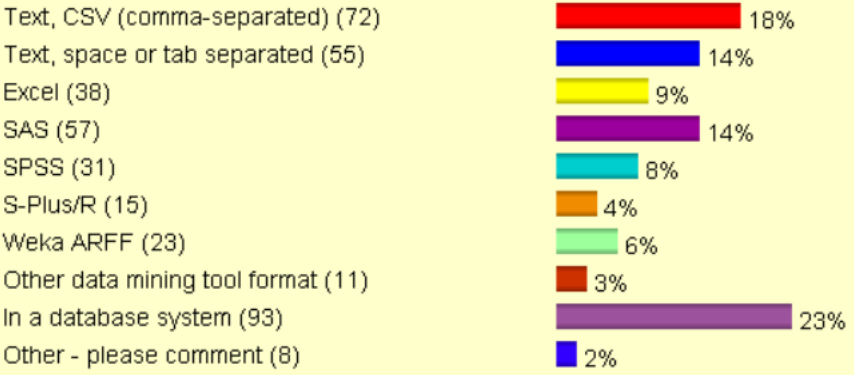
Графовые

Хранилища

KDnuggets : **Polls** : Data Storage Formats (June 2005)

Poll

What are your preferred methods for storing data for data mining? [403 votes total]



И ещё:

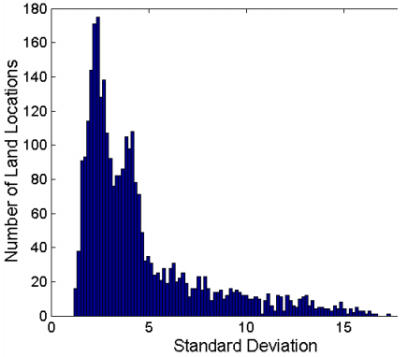
Проблемы в данных:

- шумы и выбросы
- пропущенные значения
- дубликаты

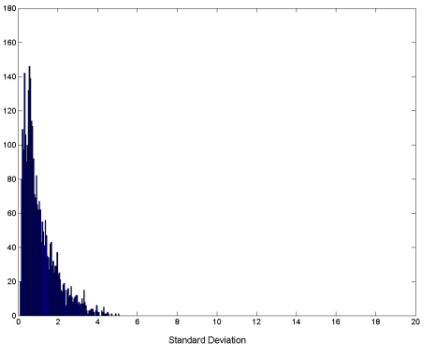
Этапы обработки данных:

- объединение
- формирование выборки
- снижение размерности
- выбор основных характеристик
- создание характеристик
- изменение атрибутов
- дискретизация

Объединение выборок



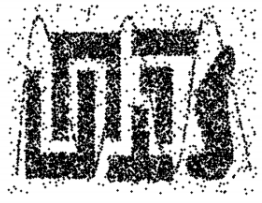
Standard Deviation of Average Monthly Precipitation



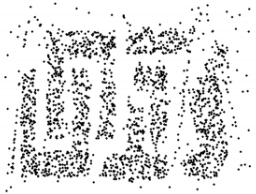
Standard Deviation of Average Yearly Precipitation

Рис.: Зачем объединять и аггрегировать данные

Размер выборки



8000 points



2000 Points

500 Points

Проклятие размерности

Что происходит:

- разреженные данные
- метрики не помогают

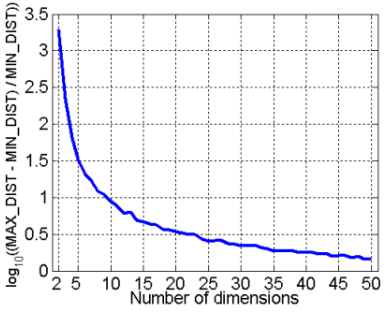


Рис.: 500 случайно-сгенерированных точек и расстояние между ними

Понижение размерности

Цель:

- избежать «проклятия»
- уменьшить временные затраты
- уменьшить объем хранимой информации
- повысить легкость визуализации
- снизить шум и убрать «неважные» характеристики

Способы:

- Principle Component Analysis
- Singular Value Decomposition
- Autoencoders

PCA Analysis

Рис.: Одно и то же изображение при PCA

Отбор признаков

Выбор основных характеристик или их создание:

- Избыточные характеристики
- «Неважные» характеристики
- Перебор всех подмножеств
- Заложено в самом алгоритме
- Здравый смысл
- Использование <другого> алгоритма DM
- Проекция данных в другое подпространство

Изменение аттрибутов

Способы:

- Стандартизация $x = x - mean/std$
- Нормализация
- Сглаживание
- Конструирование признаков (случайно, генетика, etc.)

ЧОКАК?