



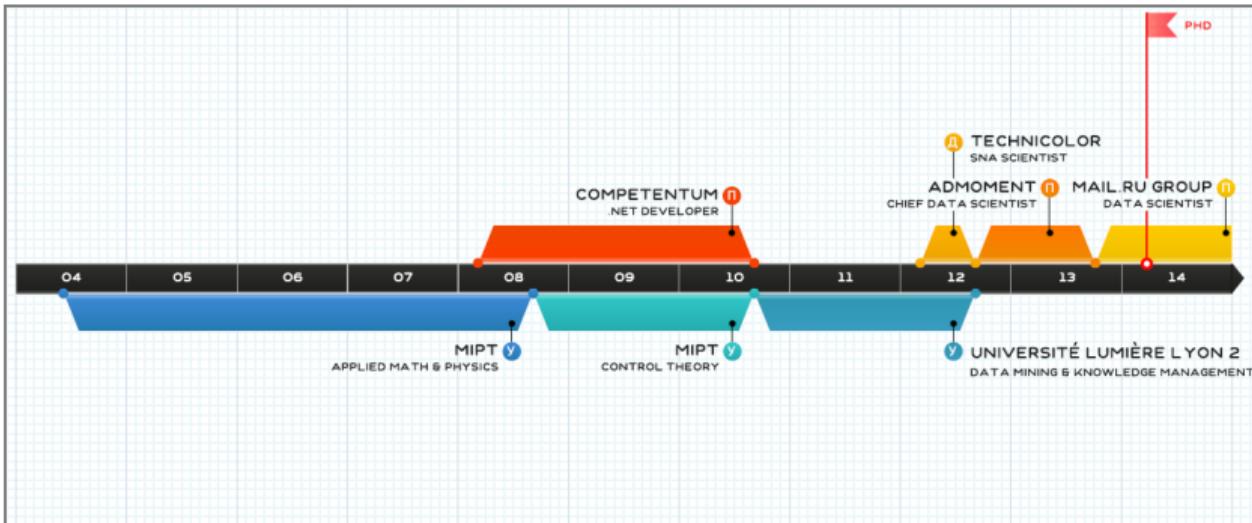
ТЕХНОСФЕРА

Лекция 1 Задачи Data Mining

Николай Анохин

21 сентября 2015 г.

Николай Анохин



e-mail: n.anokhin@corp.mail.ru

тел.: +7 (903) 111-44-60

План лекции

Структура курса

Что такое Data Mining

Унификация процесса Data Mining

Exploratory data analysis

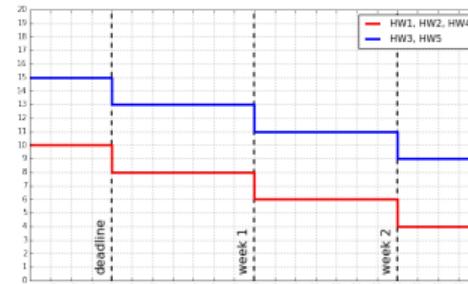
Структура курса

1. Задачи Data Mining^{HW1}
2. Задача кластеризации и EM-алгоритм^{HW2}
3. Алгоритмы кластеризации^{HW3}
4. Визуализация результатов кластеризации
5. Кластеризация: итоговое занятие^{K1}
6. Задача классификации
7. Naive Bayes и работа с текстом^{HW4}
8. Решающие деревья
9. Линейные модели^{HW5}
10. Метод опорных векторов
11. Data Mining в реальных системах
12. Классификация: итоговое занятие^{K2}
13. Защита семестрового проекта

Контроль знаний

дз

- ▶ дз-1, дз-2, дз-4 : максимум 10 баллов за каждое
- ▶ дз-3, дз-5: максимум 15 баллов за каждое



Теория

- ▶ К-1, К-2 : максимум 10 баллов за каждый
- ▶ Защита проекта: максимум 20 баллов

Шкала оценок



(a) 0 — 49



(c) 80 — 94



(b) 50 — 79



(d) 95 — 100

Правила

- + Можно задавать вопросы по ходу лекции
- + Можно входить и выходить, не мешая коллегам
- Нельзя нарушать порядок в аудитории
- Нельзя разговаривать по телефону
- ▶ Общение с преподавателем на “Вы”

Ваши правила?

Что такое Data Mining?



Data Mining как KDD

*Knowledge Discovery in Databases (KDD) – это процесс получения точных, неизвестных, потенциально полезных и интерпретируемых закономерностей из данных.*¹

¹U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery: an overview. 1996

Data Mining как моделирование

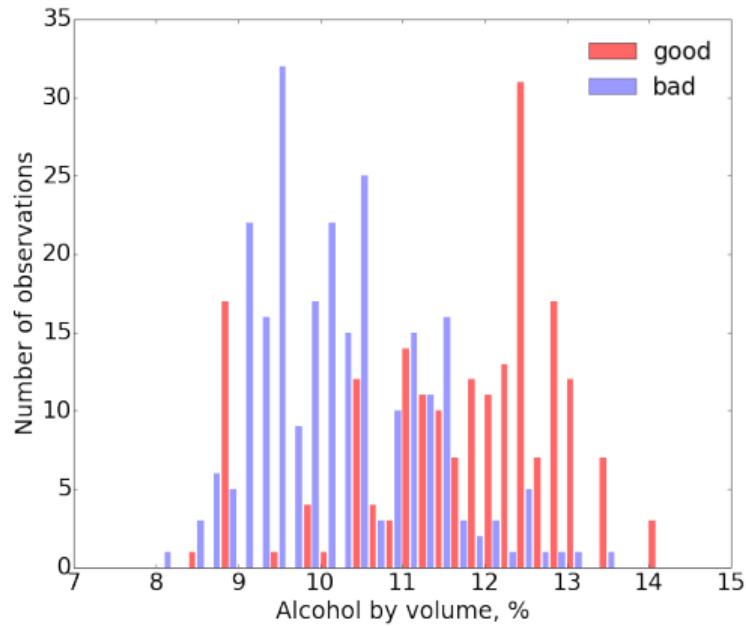
Data Mining – процесс построения модели, хорошо описывающей закономерности, которые порождают данные.

Подходы к построению моделей

- ▶ статистический
- ▶ машинное обучение
- ▶ вычислительный

Качество вина²

	ABV, %	Quality
1	12.8	good
2	12.8	good
3	10.5	good
4	10.7	good
5	10.7	good
...
198	11.4	good
199	10.10	bad
200	10.30	bad
201	10.90	bad
202	9.95	bad
...
444	9.05	bad



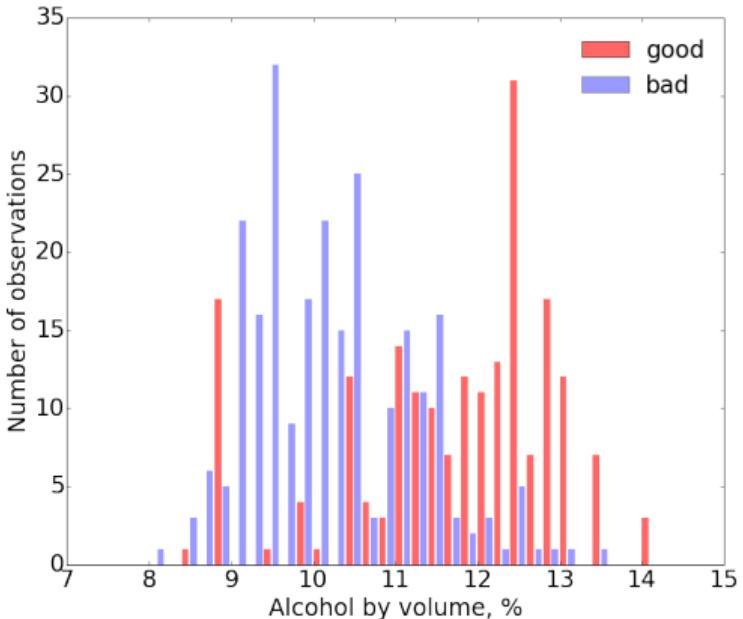
²Wine Quality Data Set. UCI Machine Learning Repository

Качество вина: статистический подход

$$\begin{cases} p(\text{alcohol} \mid \text{good}) \sim \mathcal{N}(\text{alcohol} \mid \mu_g, \sigma_g) \\ p(\text{alcohol} \mid \text{bad}) \sim \mathcal{N}(\text{alcohol} \mid \mu_b, \sigma_b) \end{cases}$$

↓ (ML-принцип)

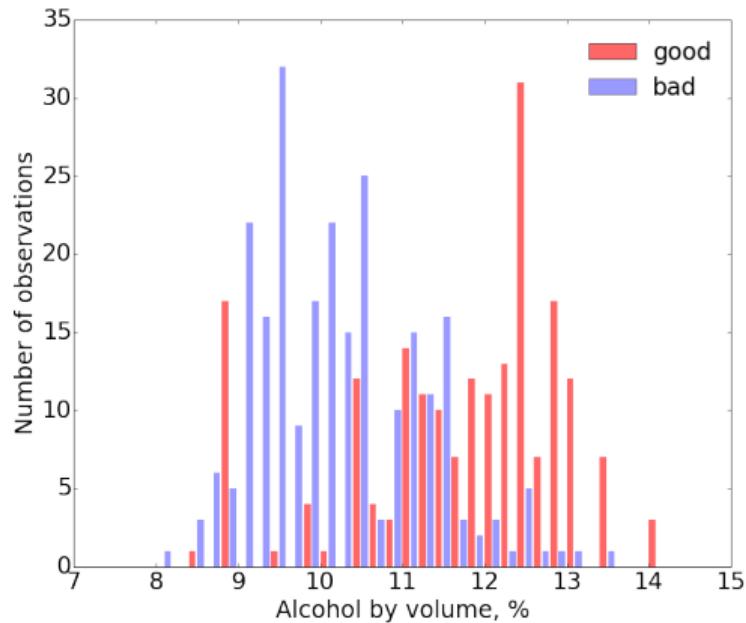
$$\begin{cases} \mu_g = 11.4, \sigma_g = 1.3 \\ \mu_b = 10.2, \sigma_b = 1.0 \end{cases}$$



Качество вина: машинное обучение

Обучаем линейный SVM:

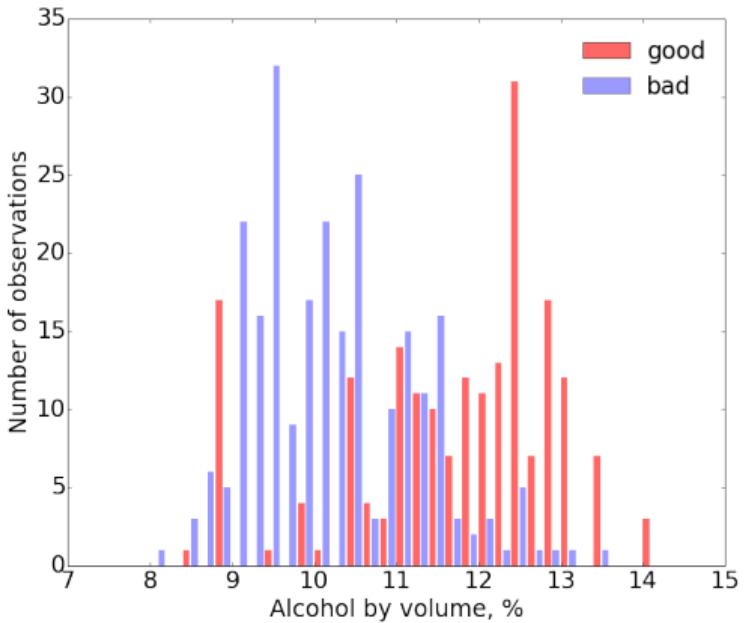
$$\text{alcohol} > 11.2 \Rightarrow \text{good}$$



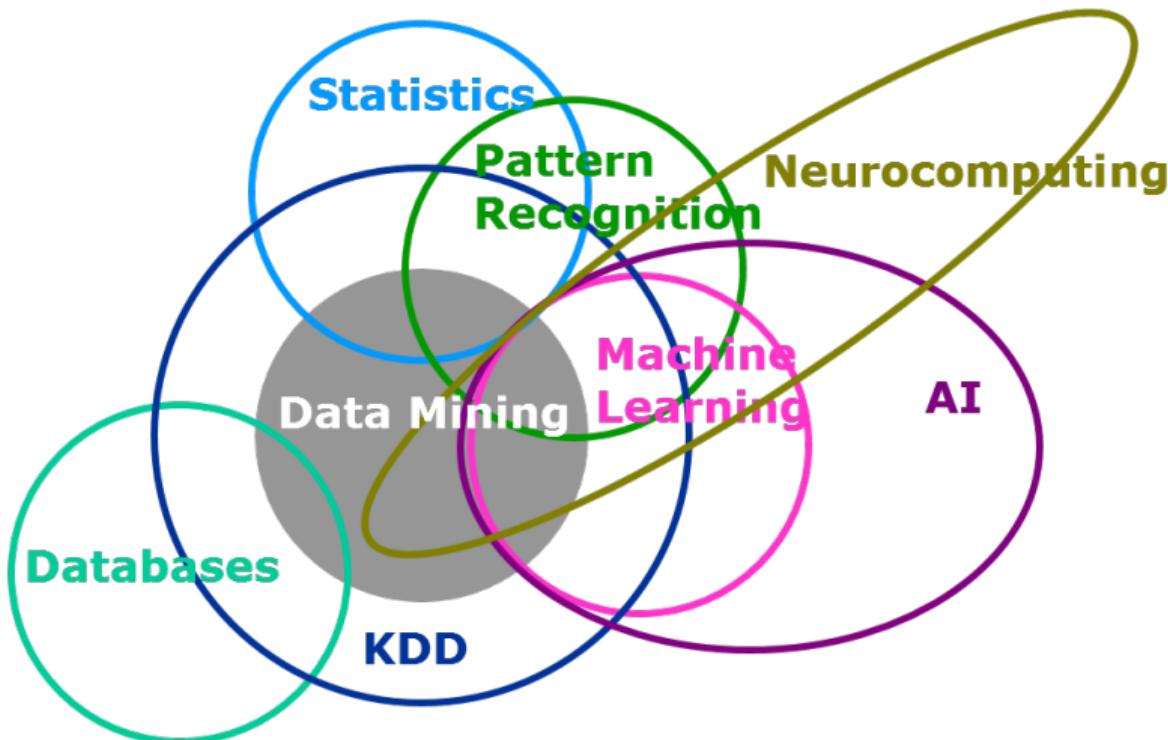
Качество вина: вычислительный подход

Подсчитываем параметры данных:

$$\langle \text{alcohol} \rangle_g = 11.4, \langle \text{alcohol} \rangle_b = 10.2$$



Data Mining – область на пересечении дисциплин²



²Looking backwards, looking forwards: SAS, data mining, and machine learning

Data Mining – область тысячи имен

1960-е Data Fishing, Data Dredging

1980-е Knowledge Discovery in Databases

1990-е Data Mining, Database miningTM

2000-е Data Analytics, Data Science³⁴

³Data Scientist is a Data Analyst who lives in California

⁴A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

Некоторые важные события в истории Data Mining

- 1989 IJCAI-89 Workshop on Knowledge Discovery in Databases
- 1995 ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- 2001 Leo Breiman's "Statistical Modeling: The Two Cultures"
- 2003 Программа Total Information Awareness
- 2005 Doug Cutting и Mike Cafarella разработали пакет обработки данных Hadoop
- 2007 Первый релиз библиотки scikit-learn
- 2010 Заработал сайт Kaggle – платформа для проведения соревнований по Data Science
- 2012 Harvard Business Review публикует статью Data Scientist: The Sexiest Job of the 21st Century
- 2013 Первая встреча сообщества Moscow Data Science⁵ в московском офисе Mail.Ru Group

⁵<http://www.meetup.com/Moscow-Data-Science/>

Data Scientist

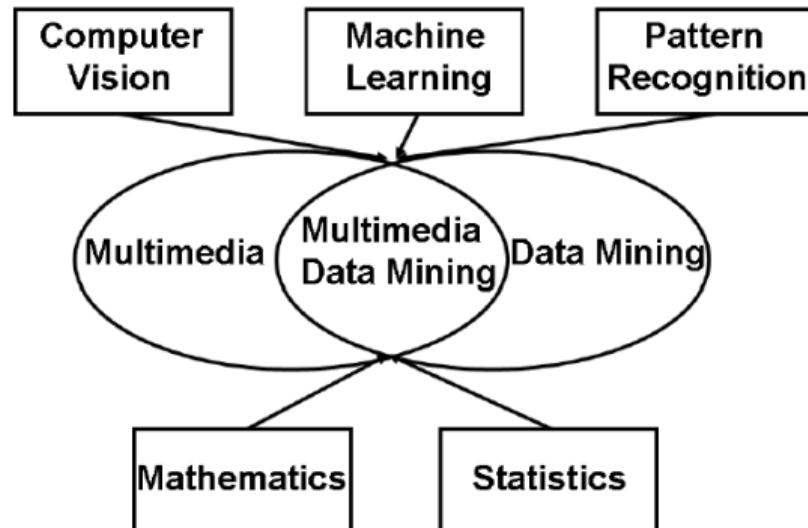
Person who is better at statistics than any software engineer and better at software engineering than any statistician

(J. Wills, Data Scientist at Cloudera Inc.)

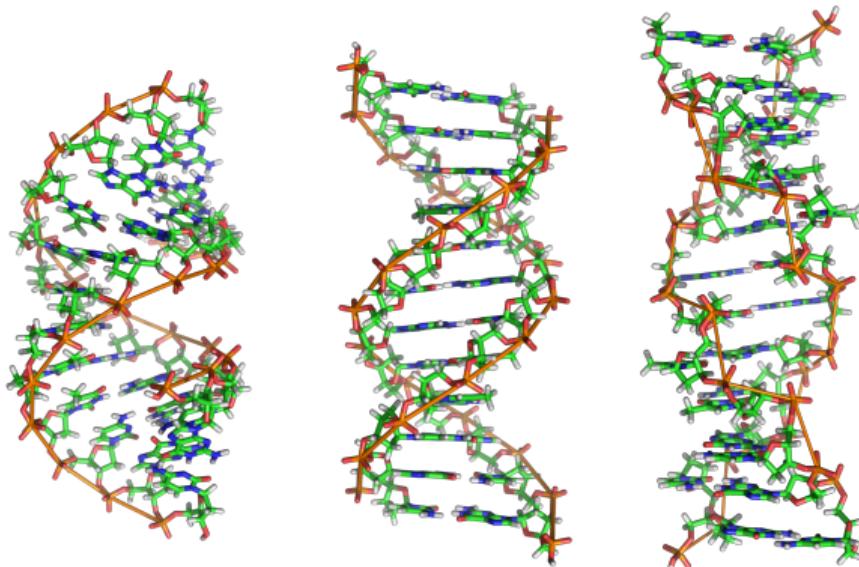
Data-

- ✖ -architecture
- ✖ -acquisition
- ✓ -analysis
- ✖ -archiving

Multimedia Data Mining



Data Mining в медицине и биологии



Data Mining в финансовой сфере

- ▶ Предсказание биржевых цен
- ▶ Предсказание курса валют
- ▶ Работа с рисками и банкротством
- ▶ Кредитный scoring
- ▶ Выявление мошенников

Data Mining для CRM и целевого маркетинга

CRM = Customer Relationship Management

- ▶ Как определить, кто собирается уйти?
- ▶ Какие продукты предложить клиенту?
- ▶ Как найти новых клиентов?
- ▶ Реклама!

Data Mining для Высшего блага (на самом деле fail)

- ▶ 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

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

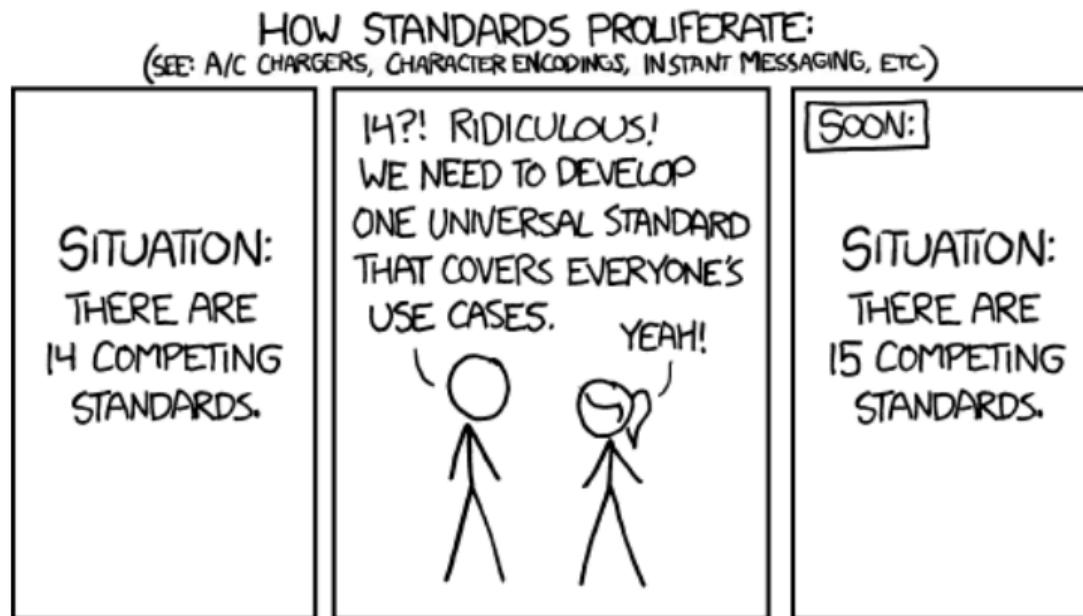
Ожидаемое количество “подозрительных” встреч в отелях

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

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.

Унификация процесса Data Mining



CRISP-DM

Cross Industry Standard Process for Data Mining



Игра в гольф⁶

Business understanding

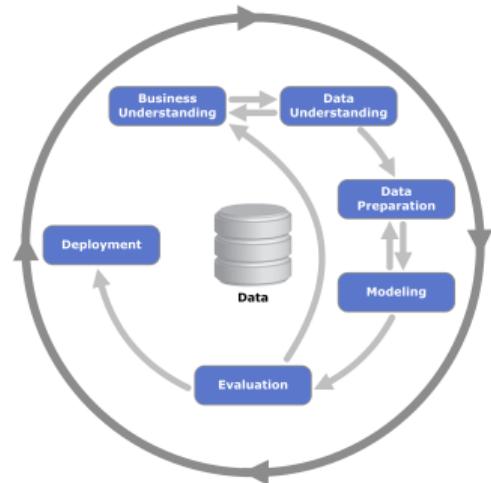
- ▶ понимание задачи с точки зрения бизнеса
- ▶ сбор требований и ограничений
- ▶ постановка задачи в терминах Data Mining

\mathcal{D} – множество, содержащее все рассматриваемые в задаче объекты

$f : \mathcal{D} \rightarrow \mathcal{Y}$ – целевая функция

Цель – с использованием данных о конечном множестве объектов из \mathcal{D} (data set) научиться предсказывать значения целевой функции для любых объектов из \mathcal{D}

Задача **с учителем** – для “известных” объектов дано значение целевой функции, иначе – задача **без учителя**.



⁶Induction of Decision Trees / R. Quinlan

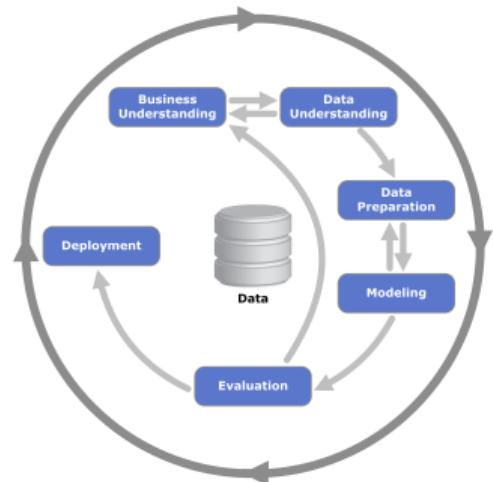
Игра в гольф

Data understanding

- ▶ первичный сбор данных
- ▶ ознакомление с данными и понимание их специфики

Data preparation

- ▶ формирование финального набора данных



Признаки

\mathcal{D} – множество, содержащее все рассматриваемые в задаче объекты

$d \in \mathcal{D}$ – объект, $\phi_j : \mathcal{D} \rightarrow F_j$ – признак

Виды признаков

- ▶ Бинарные/Binary

$$F_j = \{\text{true}, \text{false}\}$$

- ▶ Номинальные/Categorical

F_j – конечное

- ▶ Порядковые/Ordinal

F_j – конечное, упорядоченное

- ▶ Количественные/Numerical

$$F_j = \mathbb{R}$$

Признаковое представление объекта d

$$\mathbf{x} = (\phi_1(d), \dots, \phi_m(d)) \in \mathcal{X}$$

Игра в гольф: признаки

Outlook	Temperature	Humidity	Wind	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Моделирование

- ▶ перебор различных моделей
- ▶ настройка параметров моделей

Модель

признаковое описание объекта d :

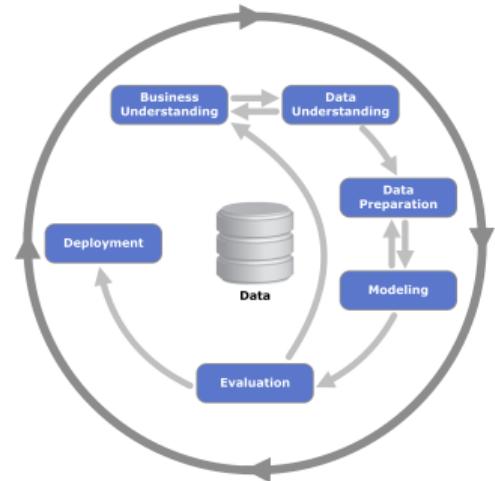
$$\mathbf{x} = (\phi_1(d), \dots, \phi_m(d)) \in \mathcal{X}$$

значение целевой функции для объекта d : $f(d) = y \in \mathcal{Y}$

модель – семейство функций вида

$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}\},$$

где $\theta \in \Theta$ – неизвестный вектор параметров



Виды моделей

Генеративные модели. Смоделировать $p(\mathbf{x}|y_k)$ и $p(y_k)$, применить теорему Байеса

$$p(y_k|\mathbf{x}) = \frac{p(\mathbf{x}|y_k)p(y_k)}{p(\mathbf{x})}$$

и использовать $p(y_k|\mathbf{x})$ для принятия решения
(NB, Bayes Networks, MRF)

Дискриминативные модели. Смоделировать $p(y_k|\mathbf{x})$ и использовать ее для принятия решения
(Logistic Regression, Decision Trees)

Функции решения. Смоделировать напрямую $h^*(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{Y}$
(SVM, Neural Networks)

Вероятностные модели VS Функции решения

- Отказ от классификации (reject option)
- Дисбаланс в выборке
- Ансамбли моделей
- Сильные предположения о природе данных
- Излишняя (вычислительная) сложность

Качество вина

признаковое описание: $\mathbf{x} \in \mathbb{R}^1$

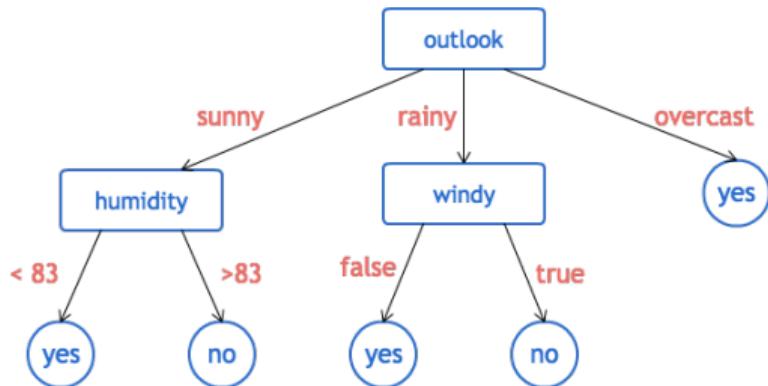
целевая переменная: $y = 1$, если вино хорошее, $y = 0$ иначе

модель:

$$\begin{cases} p(\mathbf{x}|\text{good}) \sim \mathcal{N}(\mathbf{x}|\mu_g, \sigma_g), & p(\text{good}) = \frac{1}{2} \\ p(\mathbf{x}|\text{bad}) \sim \mathcal{N}(\mathbf{x}|\mu_b, \sigma_b), & p(\text{bad}) = \frac{1}{2} \end{cases} + y = \mathcal{I}(p(\text{good}|\mathbf{x}) > p(\text{bad}|\mathbf{x}))$$

параметры: $\theta = (\mu_g, \sigma_g, \mu_b, \sigma_b)$

Дерево решений



Outlook	Temperature	Humidity	Wind	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Обучение модели

- ▶ дана обучающая выборка (data set) $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$
- ▶ для каждого из объектов обучающей выборки дано значение целевой функции $Y = \{y_1, \dots, y_N\}$ (если задача с учителем)

Алгоритм обучения

Выбор наилучших параметров θ^* с использованием обучающей выборки

$$A(X, Y) : (\mathcal{X} \times \mathcal{Y})^N \rightarrow \Theta$$

В итоге:

$$h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$$

Learning = Representation + Evaluation + Optimization⁷

⁷A Few Useful Things to Know about Machine Learning // Pedro Domingos

Evaluation

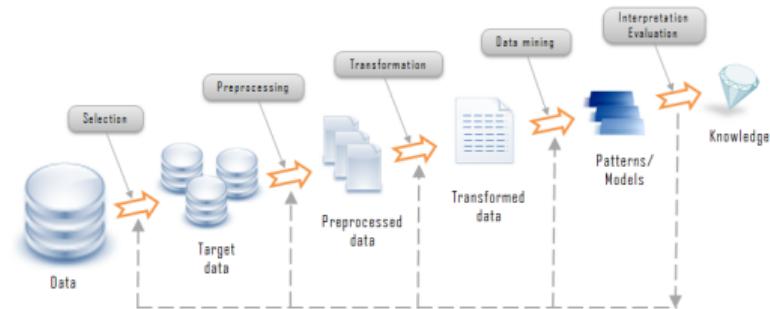
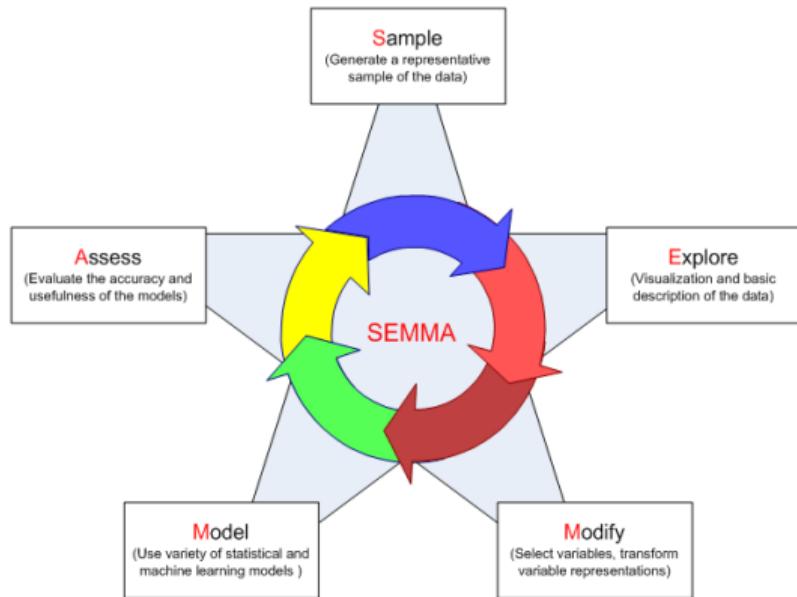
- ▶ тщательная проверка качества модели
- ▶ подробное рассмотрение шагов, предпринятых при построении
- ▶ поиск бизнес-требований, которые не удовлетворены

Deployment

- ▶ презентация модели клиенту
- ▶ развертывание и использование модели



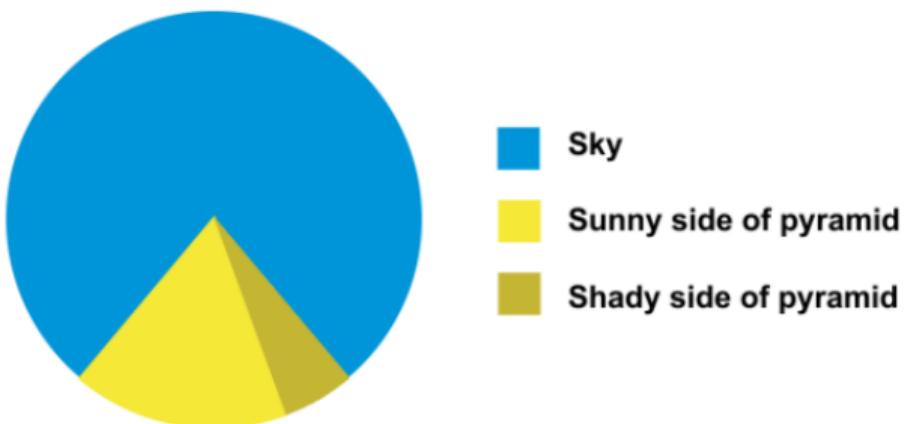
Другие процессы: SEMMA⁸, KDD⁹



⁸<http://timkienthuc.blogspot.ru/2012/04/crm-and-data-mining-day-08.html>

⁹<http://www.rithme.eu/>

Exploratory data analysis



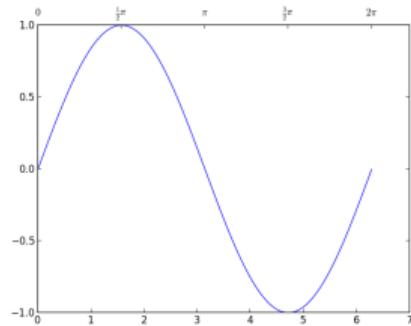
Exploratory data analysis

EDA направлен на предварительное изучение данных

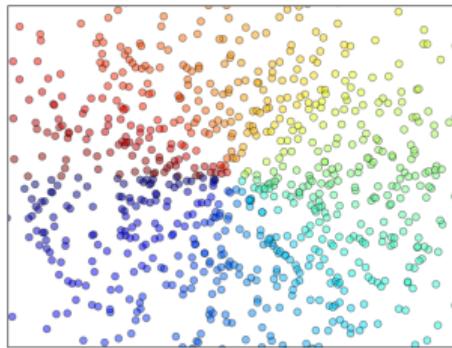
- ▶ формирование гипотез относительно структуры данных
- ▶ выбор необходимых инструментов анализа

Особенность метода состоит в визуализации и поиске важных характеристик и тенденций

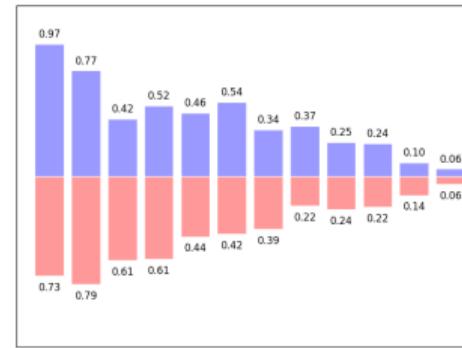
Примеры



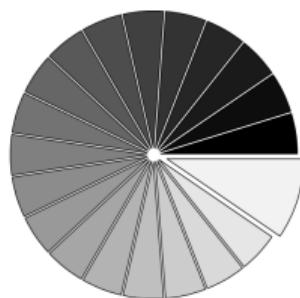
(a) Plot



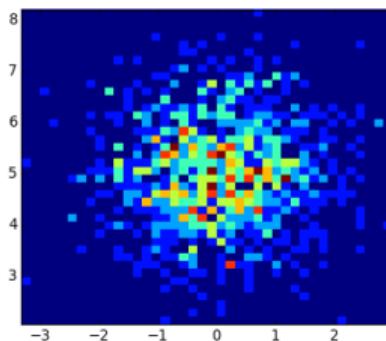
(b) Scatter



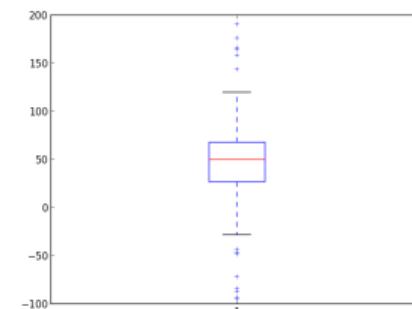
(c) Barplot



(d) Piechart



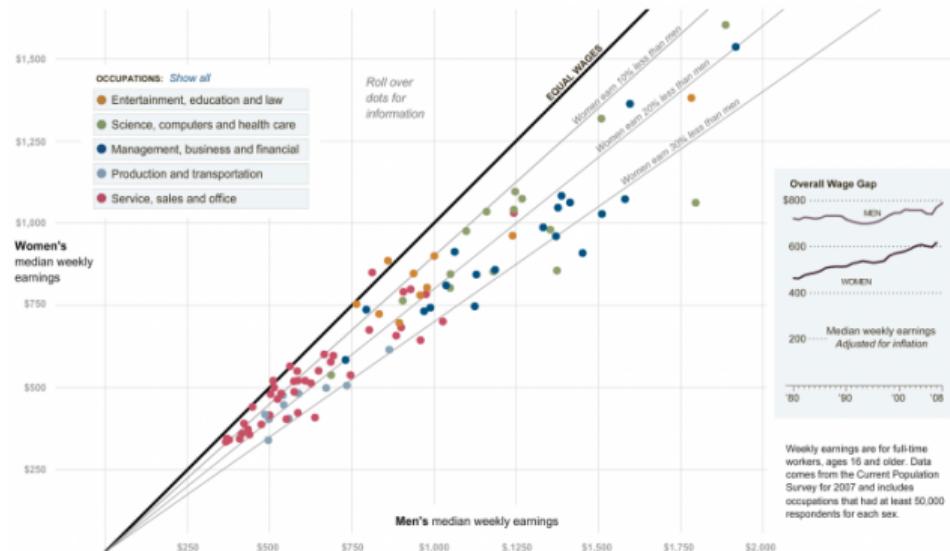
(e) Heatmap



(f) Boxplot

Полезные советы

- ▶ Все познается в сравнении
- ▶ Причинно-следственные связи
- ▶ Размерность имеет значение
(больше-лучше)
- ▶ Не избегать пояснений
- ▶ Content is king



Вопросы

