Краткое введение в data mining

Николай Анохин

Data Mining как KDD

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

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

Data Mining как моделирование

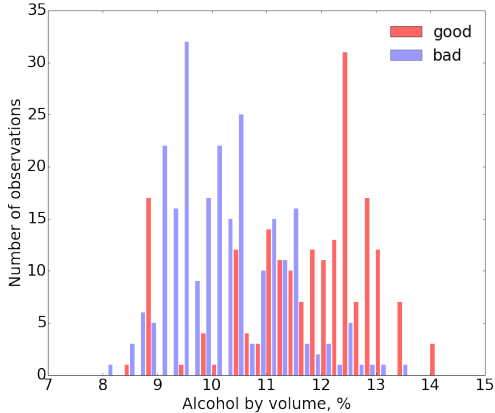
Data Mining – процесс построения модели, хорошо описывающей закономерности, которые порождают данные.

Подходы к построению моделей

- статистический
- машинное обучение
- вычислительный

Качество вина[2]

	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



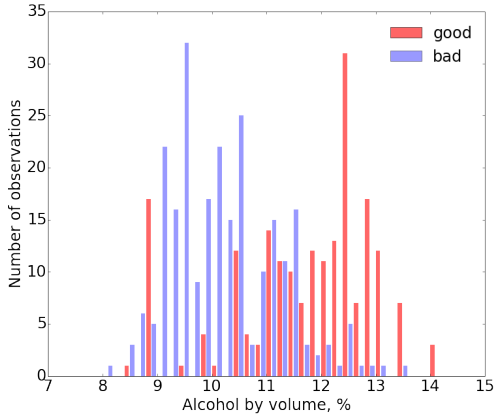
[2]Wine Quality Data Set. UCI Machine Learning Repository

Качество вина: статистический подход

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

$$\Downarrow \text{(ML-принцип)}$$

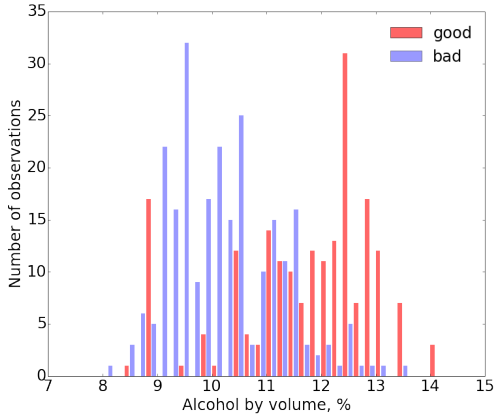
$$\begin{cases} \mu_g = 11.4, \sigma_g = 1.3 \\ \mu_b = 10.2, \sigma_b = 1.0 \end{cases}$$



Качество вина: машинное обучение

Обучаем линейный SVM:

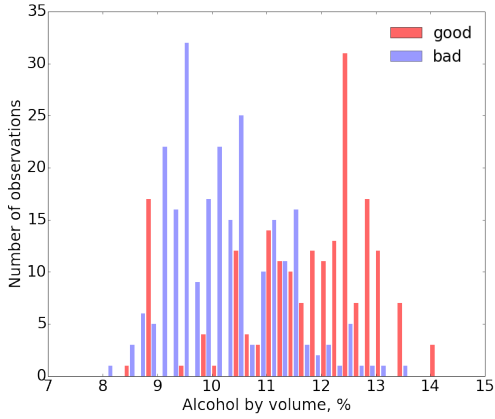
$$\text{alcohol} > 11.2 \Rightarrow \text{good}$$



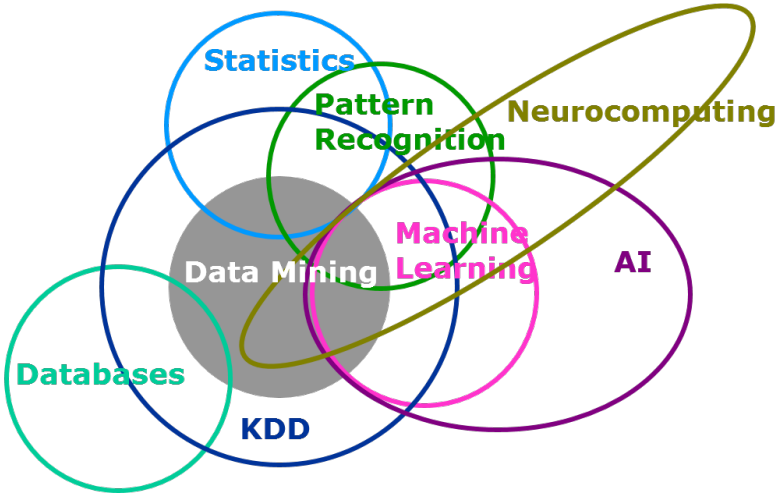
Качество вина: вычислительный подход

Подсчитываем параметры данных:

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



Data Mining – область на пересечении дисциплин[2]



[2]Looking 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 mining™

2000-е Data Analytics, Data Science[3][4]

[3] Data Scientist is a Data Analyst who lives in California

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

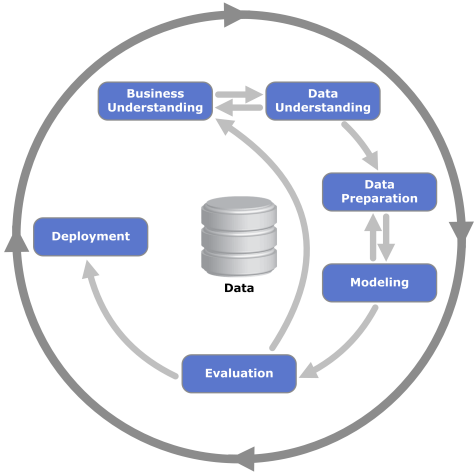
Некоторые важные события в истории Data Mining

- 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 Science[5] в московском офисе Mail.Ru Group

[5] http://www.meetup.com/Moscow-Data-Science/

CRISP-DM

(Cross Industry Standard Process for Data Mining)



Игра в гольф[6]

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

$f : \mathcal{D} \to \mathcal{Y}$ – 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**.

[6] Induction of Decision Trees / R. Quinlan

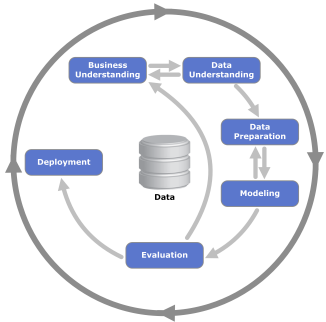
Игра в гольф

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, $\phi_j : \mathcal{D} \to F_j$ – признак

Виды признаков

- Бинарные/Binary
  $F_j = \{true, false\}$
- Номинальные/Categorical
  F_j – конечное
- Порядковые/Ordinal
  F_j – конечное, упорядоченное
- Количественные/Numerical
  $F_j = \mathbb{R}$

Признаковое представление объекта d

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

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:

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

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

модель – семейство функций вида

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

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

Качество вина

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

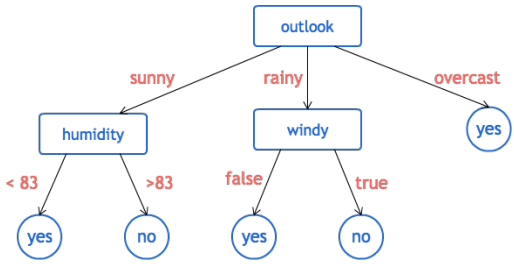
целевая переменная: $y = 1$, если вино хорошее, $y = 0$ иначе

модель:

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

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) $X = \{\mathbf{x}_1, \ldots, \mathbf{x_N}\}$
- для каждого из объектов обучающей выборки дано значение целевой функции $Y = \{y_1, \ldots, y_N\}$ (если задача с учителем)

Алгоритм обучения

Выбор наилучших параметров θ^* с использованием обучающей выборки

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

В итоге:

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

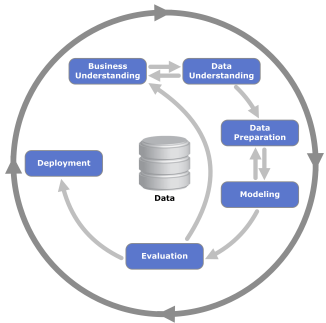
Пример 2. Игра в гольф

Evaluation

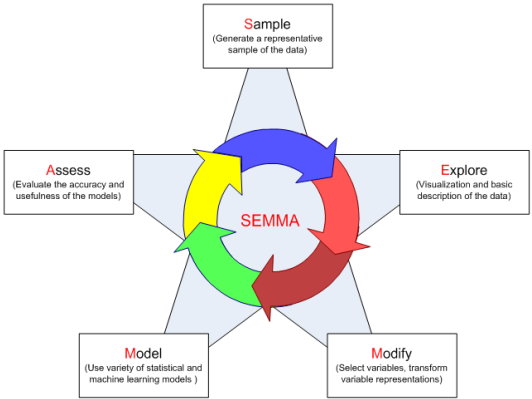
- тщательная проверка качества модели
- подробное рассмотрение шагов, предпринятых при построении
- поиск бизнес-требований, которые не удоволетворены

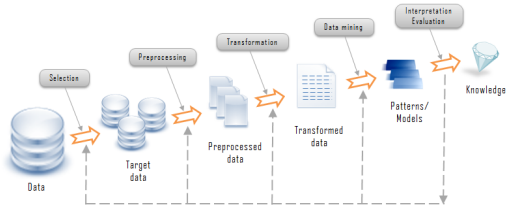
Deployment

- презентация модели клиенту
- развертывание и использование модели



Другие процессы: SEMMA[7], KDD[8]





[7] http://timkienthuc.blogspot.ru/2012/04/crm-and-data-mining-day-08.html
[8] http://www.rithme.eu/

Задача кластеризации

В задачах кластеризации целевая переменная не задана. Цель – отыскать “скрытую структуру” данных.

Дано. Признаковые описания N объектов $\mathbf{x} \in \mathcal{X}$, образующие тренировочный набор данных X

Найти. Модель из семейства параметрических функций

$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathcal{Y} \mid \mathcal{Y} = \{1, \ldots, K\}\},$$

ставящую в соответствие произвольному $\mathbf{x} \in \mathcal{X}$ один из K кластеров так, чтобы объекты внутри одного кластера были похожи, а объекты из разных кластеров различались