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

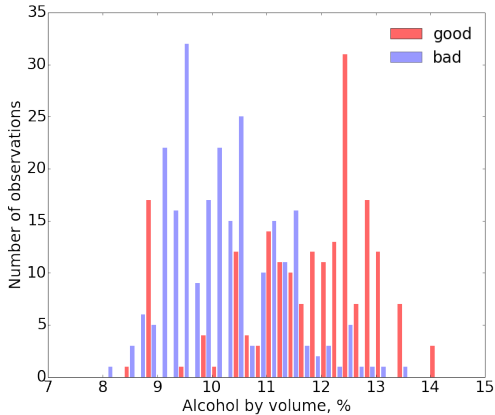
Data Mining – процесс построения модели, хорошо описывающей закономерности, которые порождают данные.

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



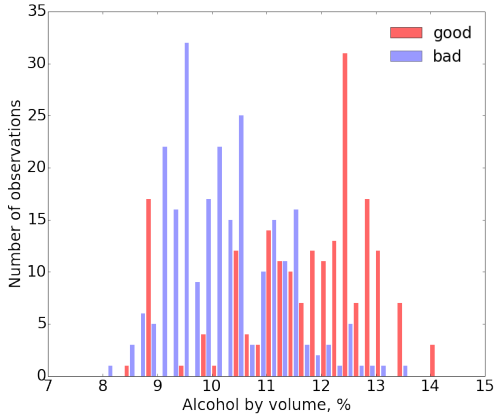
[2]Wine Quality Data Set. UCI Machine Learning Repository

Качество вина: статистический подход

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

$$\Downarrow \text{(ML-принцип)}$$

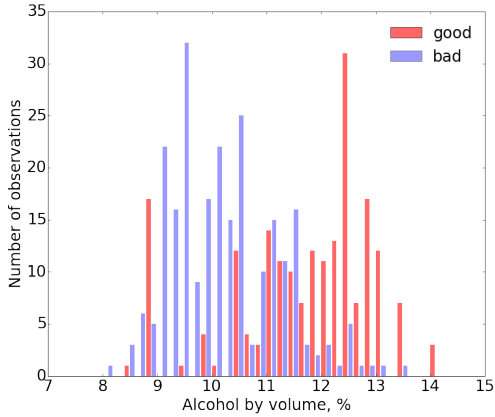
$$\begin{cases} \mu_g = 11.4, \sigma_g = 1.3 \\ \mu_b = 10.2, \sigma_b = 1.0 \end{cases}$$



Качество вина: машинное обучение

Обучаем линейный SVM:

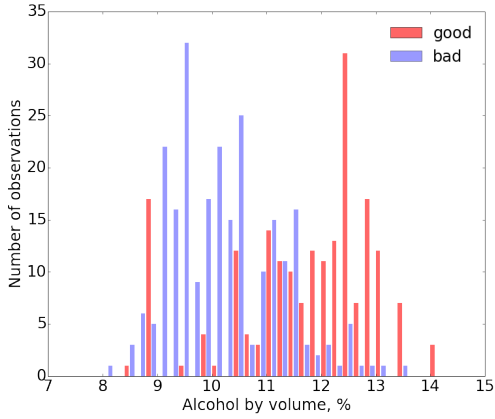
$$\text{alcohol} > 11.2 \Rightarrow \text{good}$$



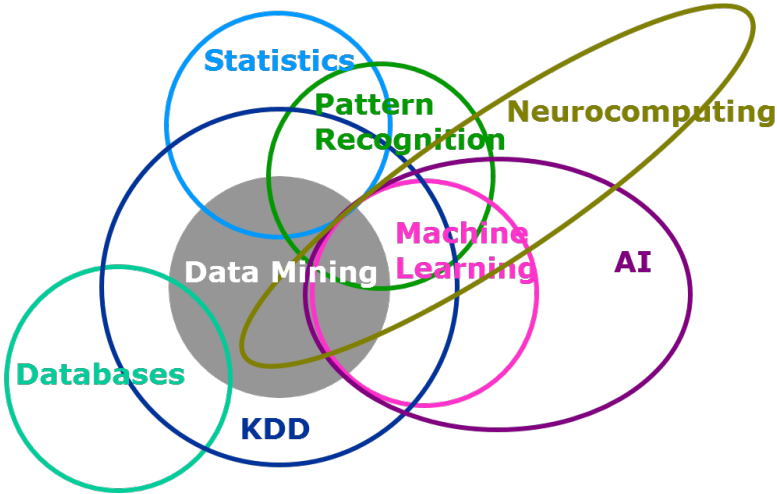
Качество вина: вычислительный подход

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/