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

Пример 1. Красная икра на новогодний стол

настоящая	446	521	550	315	613	292	469	658	255	310
искусственная	372	351	361	398	348	457	370	473	475	435

Пример 1. Красная икра на новогодний стол

настоящая	446	521	550	315	613	292	469	658	255	310
искусственная	372	351	361	398	348	457	370	473	475	435

Статистический подход

$$\begin{cases} p(\text{цена}|\text{настоящая}) \sim \mathcal{N}(\text{цена}|\mu_r, \sigma_r) \\ p(\text{цена}|\text{искусственная}) \sim \mathcal{N}(\text{цена}|\mu_a, \sigma_a) \end{cases} \xRightarrow{MLE} \begin{cases} \mu_r = 443, \sigma_r = 136 \\ \mu_a = 404, \sigma_a = 49 \end{cases}$$

Машинное обучение

Обучаем линейный SVM: цена $> 482 \Rightarrow$ настоящая

Вычислительный подход

Посчитываем параметры данных: $\langle \text{цена}_r \rangle = 443, \langle \text{цена}_a \rangle = 404$

Некоторые важные события в истории Data Mining

1989 IJCAI-89 Workshop on Knowledge Discovery in Databases

Data Mining – дисциплина тысячи имен

1960-е Data Fishing, Data Dredging

1980-е Knowledge Discovery in Databases

1990-е Data Mining, Database mining^{TM}

2000-е Data Analytics, Data Science[2][3]

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

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