Overview: Data Mining Pipeline
Contents

● Data Mining Pipeline Models
  ○ Fayyad Pipeline
  ○ Cross Industry Standard Process of Data Mining
● Steps in Data Mining Pipeline
● Tasks of Modelling
● Data Mining use case in a bank: Credit or no credit?
● Summary
Knowledge Discovery in Databases - Background

- First workshop in 1989 by Gregory Piatetsky-Shapiro
- 1995 annual ACM Special Interest Group on Knowledge Discovery and Data Mining
- 1996 first formal Data Mining Pipeline Steps
Example

- In Science: Knowledge extraction as basis to theory creation
- Medicine: Patient Health history and diagnosis.
- Biology: DNA and Protein Sequencing
- In Business to forecast the market decision help.
  - Marketing: behaviour of different user groups
  - Banking: Frauds, credit card frauds, money laundering
  - Production: to diagnose and foresee the product defects.
Challenges of Knowledge Discovery in Databases

- Gigantic Databases
- Complex Relationships between attributes
- Missing data, fake data and noise
- Understanding the won patterns
- Integration in other Systems.
Fayyad’s Data Mining Pipeline

Fayyad Model [1]
Cross Industry Standard Process for Data Mining

Source: [2]
Business Understanding

- Determine business objectives and business success criteria.
- Inventory of resources and determining the requirements, assumptions, constraints, risks, costs and benefits.
- Define data mining goals and data mining success.

Our example: A bank is a financial institution that accepts deposits from the public and creates credit. We want to decide whether a person is eligible to receive a credit.
Data Understanding

- Begins with collecting the data
- Describe the data:
  - Data type: static, nominal, ordinal
  - Data Structure: Object, Dynamic, Stream-Data, Time queues
  - Data sources: Databases, Data Warehouse, WWW, Sensors
  - Size: Number of entries, number of data sentences, number of attributes, number of allowed attribute values.
- Data Quality
  - Consistency
  - Exactness

*Our example: Data on client’s credibility.*
Data Preparation

- Data Selection
  - Relevant table, attributes, records

- Data Cleaning
  - Missing values, extreme values
  - Two attributes for the same variable
  - Discretisation
  - Normalisation

Our example: The data has been cleaned
Modeling

Data Mining Paradigms

- Verification
  - Goodness of fit
  - Hypothesis testing
  - Analysis of variance

- Discovery
  - Prediction
  - Classification
  - Regression
  - Description
    - Clustering
    - Summarization
    - Linguistic summary
    - Visualization

Source: [5]
Evaluation

- Resampling methods: simple partitioning, cross validation, bootstrap for monitored methods.
- Criteria for interest and distance measurements.
- Select the simplest models.

Our example: Which candidates receive the credit
Basic Set

Source: [1]
Linear Classification

Source: [1]
Linear Regression

Source: [1]
Clustering

Source: [1]
Single Threshold

Source: [1]
Nearest Neighbour

Source: [1]
Deployment

- Knowledge gained will need to be organized and presented in a way that the customer can use it.

*Our example: we now have a mechanism to automatically decide whether a customer is creditable*
CRISP - DM
Iteratively

Source: [2]
Source: [1]
Summary

- Knowledge Discovery in Databases (KDD) is a non-trivial process of knowledge extraction from data.
- The steps of KDD are described in the models of Fayyad or CRISP-DM.
- Selection of suitable data mining methods is dependant on the traits of data and goals.
- The process is iterative.
Literature

1. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From Data-Mining to Knowledge Discovery in Databases. AI Magazine Volume 1, Number 3 (1996).
3. Jiawei Han, M. Kamber. Data-Mining, Concepts and Techniques. Elsevier Inc. 2006.
7. Dokumentationen von Knime, Rapid Miner, SPSS Clementine.
8. Data used or the demo comes from the Pennsylvania State University Applied Data Mining and Statistical Course