Advanced Data Mining

Piotr Lipiński

Course Organization and Structure

- See my home page for
  - lecture slides
  - assignments
  - proposals of mini-speeches
  - additional notes
  - announcements

Course Organization and Structure

- Labs:
  - points for:
    - assigments (80 points in total)
    - project (40 points)
    - bonus points for additional activities (bonus assignments, mini-speeches, itp.)
  - 120 points in total (except bonus points)
  - For passing the classes, 60 points are required. For other grades:

3.0	60 points
3.5	72 points
4.0	84 points
4.5	96 points
5.0	108 points

  - For the very good grade (5.0), in addition, it is required to prepare and present a mini-speech.
- Lecture: exam

Course Organization and Structure

- Scope:
  - selected topics in:
    - Temporal Data Mining
    - Recommender Systems
    - Dimensionality Reduction
    - other
- Requirements:
  - basic knowledge on:
    - computational intelligence
    - data mining
    - machine learning
    - probability and statistics
  - basic experiences with:
    - python for scientific computing

Data Mining

- Data mining concerns studying large datasets in order to extract non-trivial and useful knowledge.

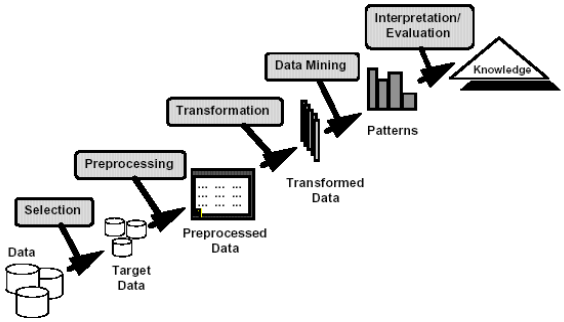
- Difference between information and knowledge:
  - information = data stored in a database or data warehouse
    - usually of large volumes
    - usually describes recorded observations of a certain phenomenon
    - usually biased by some measurement error or by a different noise
    - sometimes hard to understand by a man
      (a man cannot notice certain relations in the data)
  - knowledge
    - a model of the phenomenon or a part of it
    - usually contains a description of relations between the data
    - usually explain and helps to understand the phenomenon

Data Mining

- It is easy to extract some useless knowledge from the information:
  - we can always compute an average (from numerical attributes) or a median (from numerical or categorical attiributes)
  - we can always draw some figures
  - we can always develop an overfitted classifier system

Data Mining

- Typical process of data analysis



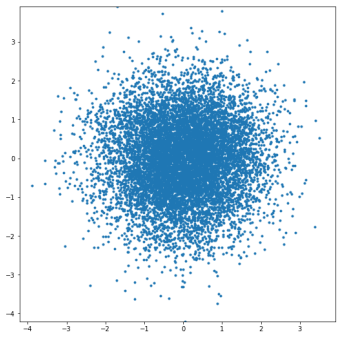
- Popular tools for data-mining:
  - Oracle Data-Mining, IBM SPSS
  - Matlab, Octave, R, Statistica
  - WEKA
  - own algorithms and their implementations

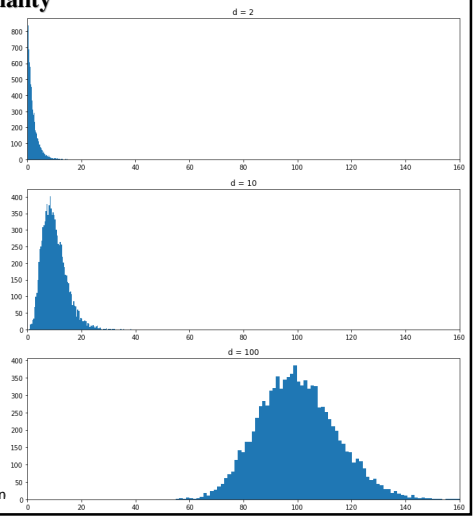
Challenges in Data Mining

- **high-dimensional data**
- **different types of noise in the data**
  - measurement errors / transmission errors / other technical errors
  - „natural noise”
  - uncertainty of the data
- **complex structure of the data / unstructured data**
- **volatility over time of the data**
  - not only in time series, but also in most of practical applications
    - recommender systems
    - switching dynamical systems
- **long-tailed data / heavy-tailed data**
- **geospatial data**
  - resolution, frequency, inaccuracy, noise

High-dimensional Data

- **High-dimensional data**
  - not only long computation time and large memory requirements
  - but above all – **curse of dimensionality**





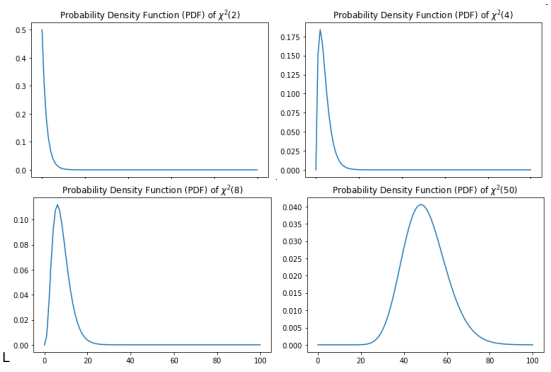
Piotr Lipiński, Advan

High-dimensional Data

- **High-dimensional data**
  - not only long computation time and large memory requirements
  - but above all – **curse of dimensionality**

REMARK: If $X = (X_1, X_2, \ldots, X_d) \in R^d \sim N(\mathbf{0}, \mathbf{I})$, then

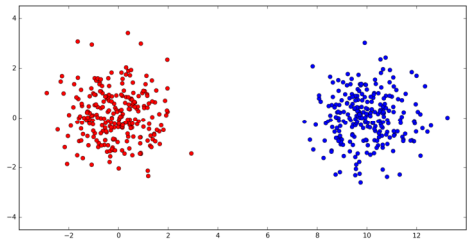
$$X_1^2 + X_2^2 + \ldots + X_d^2 \sim \chi^2(d)$$



Piotr L

High-dimensional Data

- **High-dimensional data**
  - not only long computation time and large memory requirements
  - but above all – **curse of dimensionality**

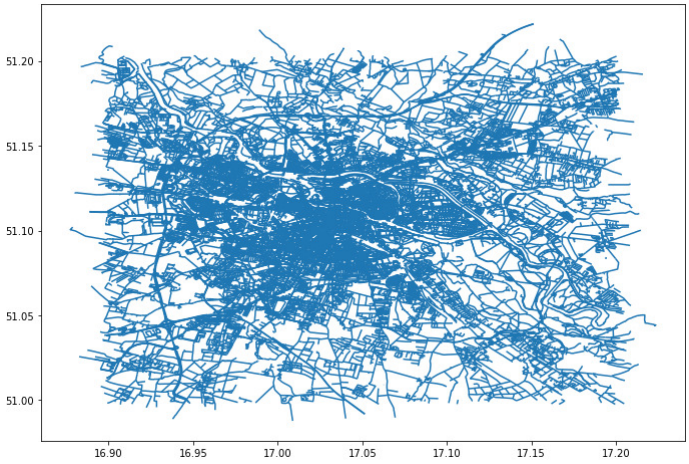


Challenges in Data Mining

- **high-dimensional data**
- **different types of noise in the data**
  - measurement errors / transmission errors / other technical errors
  - „natural noise”
  - uncertainty of the data
- **complex structure of the data / unstructured data**
- **volatility over time of the data**
  - not only in time series, but also in most of practical applications
    - recommender systems
    - switching dynamical systems
- **long-tailed data / heavy-tailed data**
- **geospatial data**
  - resolution, frequency, inaccuracy, noise

Complex Structure / Unstructured Data

- Geospatial information



Complex Structure / Unstructured Data

- London Stock Exchange Rebuilt Order Book

Complex Structure / Unstructured Data

- London Stock Exchange Rebuilt Order Book
  - DeepLOB
    - https://arxiv.org/pdf/1808.03668.pdf

Complex Structure / Unstructured Data

- unstructured data

```
<Product>
        <Product_ID>8100</Product_ID>
        <SKU>WKS-6016</SKU>
        <Name>Uptown Girl Blouse</Name>
        <Product_URL>https://www.domain.com/product/wks-6016</Product_URL>
        <Price>58</Price>
        <Retail_Price>89.95</Retail_Price>
        <Thumbnail_URL>https://www.domain.com/images/wks-6016_600x600.png</Thumbnail_URL>
        <Description>Sociosqu facilisis duis ...</Description>
        <Category>Clothing>Tops>Blouses</Category>
        <Category_ID>285</Category_ID>
        <Brand>Entity Apparel</Brand>
        <Child_SKU>WKS-6016-CORD-MD|WKS-6016-KEGR-MD</Child_SKU>
        <Child_Price></Child_Price>
        <Color>Coral Red|Kelly Green</Color>
        <Color_Family>Red|Green</Color_Family>
        <Size>Medium</Size>
        <Shoe_Size></Shoe_Size>
        <Pants_Size></Pants_Size>
        <Season>Summer|Spring</Season>
        <Badges>Exclusive</Badges>
</Product>
```

- unstructured databases (MongoDB, Redis)

Challenges in Data Mining

- **high-dimensional data**
- **different types of noise in the data**
  - measurement errors / transmission errors / other technical errors
  - „natural noise”
  - uncertainty of the data
- **complex structure of the data / unstructured data**
- **volatility over time of the data**
  - not only in time series, but also in most of practical applications
    - recommender systems
    - switching dynamical systems
- **long-tailed data / heavy-tailed data**
- **geospatial data**
  - resolution, frequency, inaccuracy, noise