Data Mining II
Organization

Heiko Paulheim, Oliver Lehmberg

Hello

- Heiko Paulheim
- Professor (interim) for Data Science
- Research Interests:
  - Semantic Web and Linked Open Data
  - Data Mining with Linked Open Data
  - Ontology Matching
  - Data Quality and Data Cleaning
- Consultation: by appointment
- Heiko will teach the lectures

Hello

- M.Sc. Wi.-Inf. Oliver Lehmberg
- Graduate Research Associate
- Research Interests:
  - Data and Web Mining
  - Network Analysis
  - Web Data Integration
- eMail: oli@informatik.uni-mannheim.de
- Oliver will teach the exercises

Course Organisation

- Poll on the exercise date
  - Monday, 10.15
  - Friday, 13.45

Course Organization

- Lecture
  - addresses advanced data mining topics
  - builds on Data Mining I lecture contents!
- Project Work
  - we will take part in the Data Mining Cup 2018
  - with four teams
    - the two best performing teams submit their solutions
  - regular presentations of your approaches
  - paper and final presentation
- Exercise
  - weekly with warm up on DMC tasks from previous years

Course Organization

- Registration
  - if not yet done, please register online at ILIAS
- Policy: two strikes out
  - we have a waiting list
  - you have to attend at least one of the first **two** lectures (today and next Tuesday)
  - otherwise, we will give your place away
- If you are on the waiting list
  - you may be assigned a place after next week's lecture
  - waiting list is cleared after this semester (i.e., no priority next year!)

Requirements

- Final exam
  - 60 % written exam
  - 40 % project work
- Project work
  - work on DMC tasks
- Presentations
  - four intermediate presentations
    - open questions, problems, current results (numbers!)
  - one final presentation
  - everybody has to present once during those four presentations
- Final report
  - 10 pages
  - solutions, results, lessons learned

i.e., grades are added and weighted,
no individual pass/fail
of exam and project

The Data Mining Cup

- An annual competition
  - for students
  - run since 2002
  - participation from all over the world
  - max. two teams per institution (i.e., university)
  - 2017: 202 participating teams from 48 countries
- Timeline
  - DMC registration on March 1st
  - tasks are published on April 5th
  - submissions are due on May 17th (internal submission: May 15th)
- Further information: http://www.data-mining-cup.de/en

The Data Mining Cup

- 2017: both Uni Mannheim teams among top 10 (out of 202)
- Prices are awarded at a conference in Berlin in June
  - Top 10 teams are invited to present their solutions

Schedule

- 13.02.18 Lecture: Preprocessing
- 20.02.18 Lecture: Regression
- 27.02.18 Lecture: Anomaly Detection
- 06.03.18 Lecture: Ensembles
- 13.03.18 Lecture: Time Series
- 20.03.18 Lecture: Neural Networks
  - 26.03. - 06.04. Easter Break
- 10.04.18 Lecture: Parameter Tuning
- 17.04.18 DMC intermediate presentation
- 24.04.18 DMC intermediate presentation
  - 01.05.18 Holiday
- 08.05.18 DMC intermediate presentation
- 15.05.18 DMC intermediate presentation
- 22.05.18 DMC final presentation

DMC task published on 05.04.

includes discussion of DMC task

final DMC submission 17.05.

Deadlines at a Glance

- March 1st: DMC registration
- April 5th: you know the DMC tasks and your team
- May 15th: submission of your DMC solution to Oli and Heiko
- May 17th: official submission of your DMC solution
- May 21st: submission of your final report
- May 22nd: final presentations

RapidMiner Analyst Certification

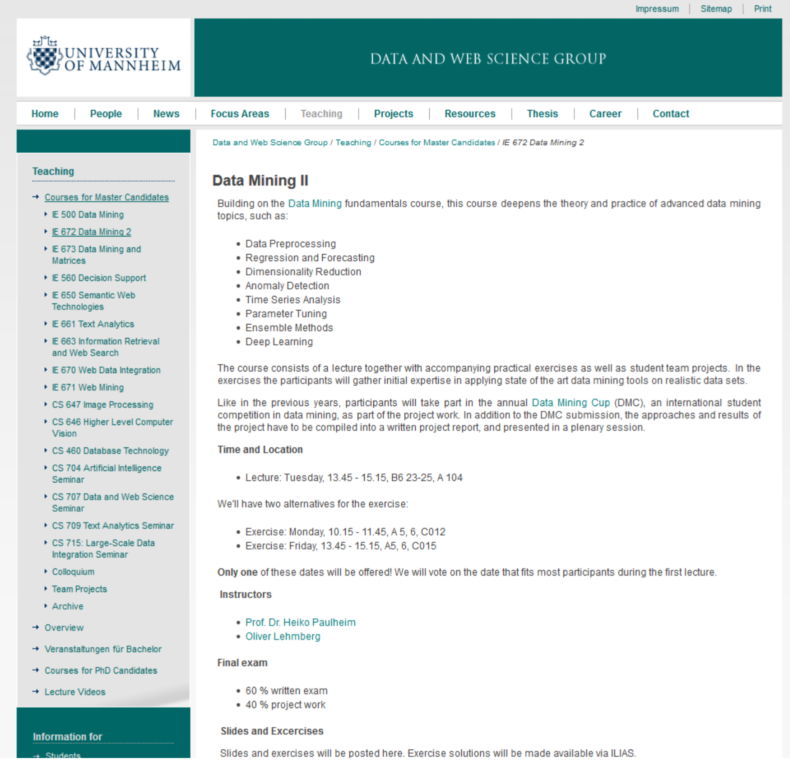
- Offered for the third time this semester
- Online exam run by RapidMiner
  - *voluntary* part of this lecture
  - does *not* replace the DM2 exam
  - last week of lecture period
  - free of charge

Lecture Contents

- Data Preprocessing (today!)
- Regression
- Anomaly Detection
- Ensemble Learning
- Time Series Analysis
- Neural Networks and Deep Learning
- Parameter Tuning

Course Organization

- Lecture Webpage: Slides, Announcements
  - http://dws.informatik.uni-mannheim.de/en/teaching/courses-for-master-candidates/ie-672-data-mining-2/
  - hint: look at version tags!
- Additional Material
  - ILIAS eLearning System, https://ilias.uni-mannheim.de/


Impressum | Sitemap | Print
UNIVERSITY OF MANNHEIM
DATA AND WEB SCIENCE GROUP
Home | People | News | Focus Areas | Teaching | Projects | Resources | Thesis | Career | Contact
Data and Web Science Group / Teaching / Courses for Master Candidates / IE 672 Data Mining 2
Teaching
Courses for Master Candidates
IE 500 Data Mining
IE 672 Data Mining 2
IE 673 Data Mining and Matrices
IE 560 Decision Support
IE 650 Semantic Web Technologies
IE 661 Text Analytics
IE 663 Information Retrieval and Web Search
IE 670 Web Data Integration
IE 671 Web Mining
CS 647 Image Processing
CS 646 Higher Level Computer Vision
CS 460 Database Technology
CS 704 Artificial Intelligence Seminar
CS 707 Data and Web Science Seminar
CS 709 Text Analytics Seminar
CS 715: Large-Scale Data Integration Seminar
Colloquium
Team Projects
Archive
Overview
Veranstaltungen für Bachelor
Courses for PhD Candidates
Lecture Videos
Information for
Students

Data Mining II

Building on the Data Mining fundamentals course, this course deepens the theory and practice of advanced data mining topics, such as:

- Data Preprocessing
- Regression and Forecasting
- Dimensionality Reduction
- Anomaly Detection
- Time Series Analysis
- Parameter Tuning
- Ensemble Methods
- Deep Learning

The course consists of a lecture together with accompanying practical exercises as well as student team projects. In the exercises the participants will gather initial expertise in applying state of the art data mining tools on realistic data sets.

Like in the previous years, participants will take part in the annual Data Mining Cup (DMC), an international student competition in data mining, as part of the project work. In addition to the DMC submission, the approaches and results of the project have to be compiled into a written project report, and presented in a plenary session.

Time and Location

- Lecture: Tuesday, 13.45 - 15.15, B6 23-25, A 104

We'll have two alternatives for the exercise:

- Exercise: Monday, 10.15 - 11.45, A 5, 6, C012
- Exercise: Friday, 13.45 - 15.15, A5, 6, C015

Only one of these dates will be offered! We will vote on the date that fits most participants during the first lecture.

Instructors

- Prof. Dr. Heiko Paulheim
- Oliver Lehmberg

Final exam

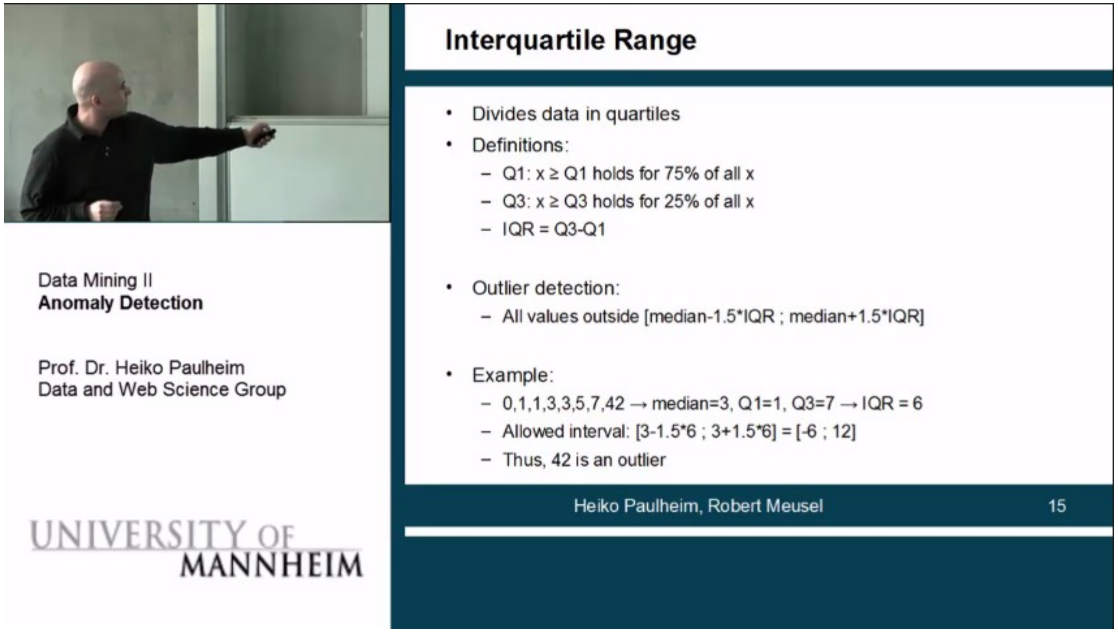
- 60 % written exam
- 40 % project work

Slides and Excercises

Slides and exercises will be posted here. Exercise solutions will be made available via ILIAS.

Video Recordings of Last Year's Lecture

- http://dws.informatik.uni-mannheim.de/en/teaching/lecture-videos/
  - Accessible from within university network and VPN



Data Mining II
Anomaly Detection

Prof. Dr. Heiko Paulheim
Data and Web Science Group

UNIVERSITY OF MANNHEIM

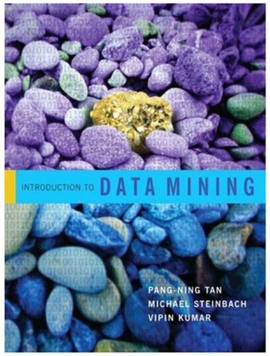
Interquartile Range

- Divides data in quartiles
- Definitions:
  - Q1: x ≥ Q1 holds for 75% of all x
  - Q3: x ≥ Q3 holds for 25% of all x
  - IQR = Q3-Q1
- Outlier detection:
  - All values outside [median-1.5*IQR ; median+1.5*IQR]
- Example:
  - 0,1,1,3,3,5,7,42 → median=3, Q1=1, Q3=7 → IQR = 6
  - Allowed interval: [3-1.5*6 ; 3+1.5*6] = [-6 ; 12]
  - Thus, 42 is an outlier

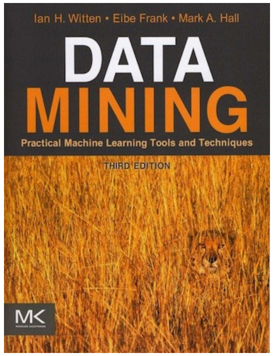
Heiko Paulheim, Robert Meusel 15

Literature & Slide Sources

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
  - 10 copies in university library.
  - we provide scans of important chapters via ILIAS



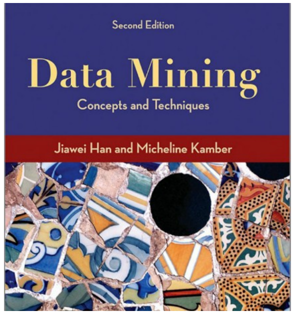
- Ian H. Witten, Eibe Frank, Mark A. Hall: Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann.
  - several copies in university library
  - we provide scans of important chapters via ILIAS



Literature & Slide Sources

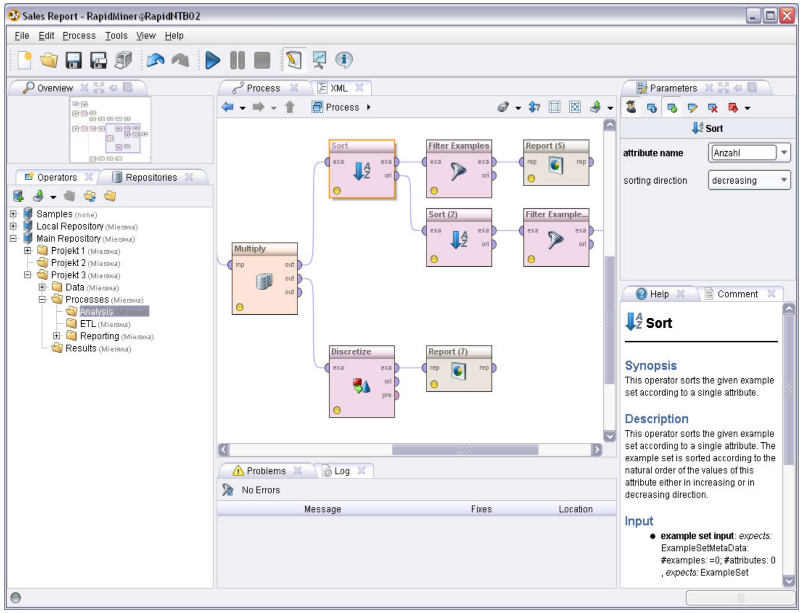
- Gregory Piatetsky-Shapiro, Gary Parker:
  KDNuggets Data Mining course:
  http://www.kdnuggets.com/data_mining_course/

- Jiawei Han and Micheline Kamber:
  Data Mining – Concepts and Techniques
  - free e-book access via university library



Software

- Powerful open-source data mining suite
- Download: http://www.rapidminer.com
- We use the free version of RapidMiner Studio
- You are invited to use other tools as well (e.g., Python, R, ...)



Sort

attribute name: Anzahl

sorting direction: decreasing

Sort

Synopsis

This operator sorts the given example set according to a single attribute.

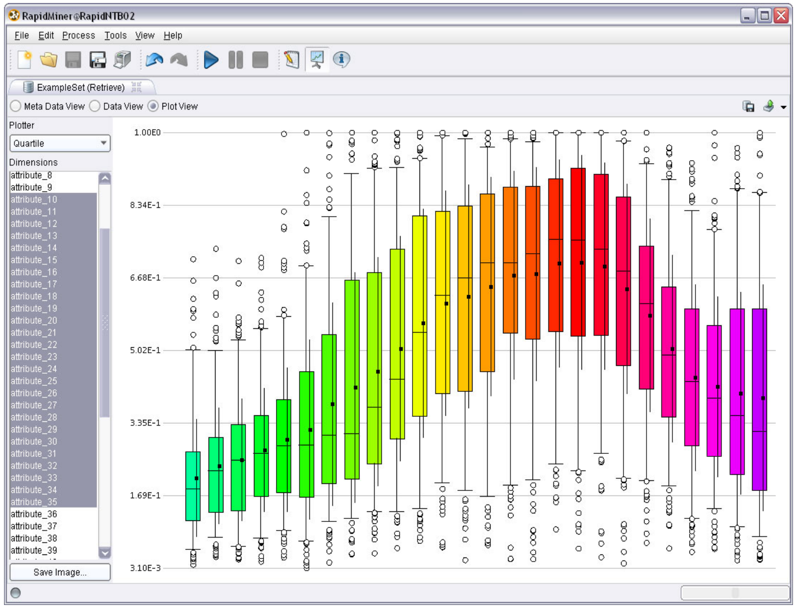
Description

This operator sorts the given example set according to a single attribute. The example set is sorted according to the natural order of the values of this attribute either in increasing or in decreasing direction.

Input

- **example set input**: *expects:* ExampleSetMetaData: #examples: =0; #attributes: 0 , *expects:* ExampleSet

No Errors



Plotter: Quartile

Questions?