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 1\textsuperscript{st}
  – tasks are published on April 5\textsuperscript{th}
  – submissions are due on May 17\textsuperscript{th} (internal submission: May 15\textsuperscript{th})

• 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 1\textsuperscript{st}: DMC registration
- April 5\textsuperscript{th}: you know the DMC tasks and your team
- May 15\textsuperscript{th}: submission of your DMC solution to Oli and Heiko
- May 17\textsuperscript{th}: official submission of your DMC solution
- May 21\textsuperscript{st}: submission of your final report
- May 22\textsuperscript{nd}: 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
  – hint: look at version tags!

• Additional Material
Video Recordings of Last Year's Lecture

  - Accessible from within university network and VPN

Interquartile Range

- Divides data in quartiles
- Definitions:
  - $Q_1 : x \geq Q_1$ holds for 75% of all $x$
  - $Q_3 : x \geq Q_3$ holds for 25% of all $x$
  - $IQR = Q_3 - Q_1$

- Outlier detection:
  - All values outside $[\text{median} - 1.5\times IQR ; \text{median} + 1.5\times IQR]$

- Example:
  - $0, 1, 1, 3, 3, 5, 7, 42 \rightarrow \text{median} = 3, Q_1 = 1, Q_3 = 7 \rightarrow \text{IQR} = 6$
  - Allowed interval: $[3 - 1.5\times 6 ; 3 + 1.5\times 6] = [-6 ; 12]$
  - Thus, 42 is an outlier

Data Mining II
Anomaly Detection
Prof. Dr. Heiko Paulheim
Data and Web Science Group

UNIVERSITY OF MANNHEIM
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

  – several copies in university library
  – we provide scans of important chapters via ILIAS
Literature & Slide Sources

• Gregory Piatetsky-Shapiro, Gary Parker: KD Nuggets 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, ...)

![RapidMiner Studio](http://www.rapidminer.com)