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 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 2017
  – 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
  - three 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
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)
  – 2016: 120 participating teams from 30 countries

• Timeline
  – DMC registration on March 8th
  – 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
<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Tuesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.02.17</td>
<td>–</td>
<td>Lecture: Preprocessing</td>
</tr>
<tr>
<td>20.02.17</td>
<td>Exercise: Preprocessing</td>
<td>Lecture: Regression</td>
</tr>
<tr>
<td>27.02.17</td>
<td>Exercise: Regression</td>
<td>Lecture: Anomaly Detection</td>
</tr>
<tr>
<td>06.03.17</td>
<td>Exercise: Anomaly Detection</td>
<td>Lecture: Ensembles</td>
</tr>
<tr>
<td>13.03.17</td>
<td>Exercise: Ensembles</td>
<td>Lecture: Time Series</td>
</tr>
<tr>
<td>20.03.17</td>
<td>Exercise: Time Series</td>
<td>Lecture: Neural Networks</td>
</tr>
<tr>
<td>27.03.17</td>
<td>Exercise: Neural Networks</td>
<td>Lecture: Parameter Tuning</td>
</tr>
<tr>
<td>03.04.17</td>
<td>Exercise: Parameter Tuning</td>
<td>–</td>
</tr>
<tr>
<td>10.04.17</td>
<td><em>Easter Break</em></td>
<td></td>
</tr>
<tr>
<td>17.04.17</td>
<td><em>Easter Break</em></td>
<td></td>
</tr>
<tr>
<td>24.04.17</td>
<td>Work on DMC tasks</td>
<td>Intermediate result presentation</td>
</tr>
<tr>
<td>01.05.17</td>
<td>Holiday</td>
<td>Intermediate result presentation</td>
</tr>
<tr>
<td>08.05.17</td>
<td>Work on DMC tasks</td>
<td>Intermediate result presentation</td>
</tr>
<tr>
<td>15.05.17</td>
<td>Work on DMC tasks</td>
<td>Consultation for final submission</td>
</tr>
<tr>
<td>22.05.17</td>
<td>–</td>
<td>Final result presentation</td>
</tr>
</tbody>
</table>
Deadlines at a Glance

• March 8th: 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 23rd: final presentations
RapidMiner Analyst Certification

• Offered for the second time this semester
• Online exam run by RapidMiner
  – voluntary part of this lecture
  – does not replace the DM2 exam
  – likely date: May 23\textsuperscript{rd}, during exercise slot
  – 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

• Time and Location
  – Monday, 10.15 – 11.45
    Room A104 in B6
  – Tuesday, 13.45 – 15.15
    Room A104 in B6

• Bring a laptop each time!
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:
  - $0, 1, 1, 3, 3, 5, 7, 42 \rightarrow \text{median} = 3, Q_1 = 1, Q_3 = 7 \rightarrow IQR = 6$
  - Allowed interval: $[3 - 1.5 \times 6 ; 3 + 1.5 \times 6] = [-6 ; 12]$
  - Thus, 42 is an outlier
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: 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
Questions?