Data Mining II
Organization

Heiko Paulheim, Robert Meusel
Hello

• Heiko Paulheim
• Assistant Professor
• Research Interests:
  – Semantic Web and Linked Open Data
  – Data Mining with Linked Open Data
  – Ontology Matching
  – Data Quality and Data Cleaning
• Room: B6 – C1.09
• Consultation: by appointment
• Heiko will teach the lectures
Hello

• Dipl.-Wi.-Inf. Robert Meusel
• Graduate Research Associate
• Research Interests:
  – Data and Web Mining
  – Social Network Analysis
  – Linked Data Technologies
• Robert will teach the exercise blocks and co-supervise the team projects
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 2016
  – 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
Requirements

• Final exam
  – 50 % written exam
  – 50 % 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)
  – 2015: 188 participating teams from 48 countries

• Timeline
  – DMC registration on March 9th
  – tasks are published on April 6th
  – submissions are due on May 18th (internal submission: May 11th)

• 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>05.02.16</td>
<td>Lecture: Preprocessing</td>
<td>Exercise: Preprocessing</td>
</tr>
<tr>
<td>22.02.16</td>
<td>Lecture: Regression</td>
<td>Exercise: Regression</td>
</tr>
<tr>
<td>29.02.16</td>
<td>Lecture: Anomaly Detection</td>
<td>Exercise: Anomaly Detection</td>
</tr>
<tr>
<td>07.03.16</td>
<td>Lecture: Ensembles</td>
<td>Exercise: Ensembles</td>
</tr>
<tr>
<td>14.03.16</td>
<td>Lecture: Time Series</td>
<td>Exercise: Time Series</td>
</tr>
<tr>
<td>21.03.16</td>
<td>Easter Break</td>
<td>Easter Break</td>
</tr>
<tr>
<td>28.03.16</td>
<td>Easter Break</td>
<td>Easter Break</td>
</tr>
<tr>
<td>04.04.16</td>
<td>Lecture: Online Learning</td>
<td>Exercise: Online Learning</td>
</tr>
<tr>
<td>11.04.16</td>
<td>DMC Task discussion</td>
<td>Team building and brainstorming</td>
</tr>
<tr>
<td>18.04.16</td>
<td>Lecture: Parameter Tuning</td>
<td>Exercise: Parameter Tuning</td>
</tr>
<tr>
<td>25.04.16</td>
<td>Intermediate result presentation</td>
<td>Work on DMC tasks</td>
</tr>
<tr>
<td>02.05.16</td>
<td>Intermediate result presentation</td>
<td>Work on DMC tasks</td>
</tr>
<tr>
<td>09.05.16</td>
<td>Intermediate result presentation</td>
<td>Work on DMC tasks</td>
</tr>
<tr>
<td>16.05.16</td>
<td>Work on Final Submission and Presentation</td>
<td></td>
</tr>
<tr>
<td>22.05.16</td>
<td>Final presentation</td>
<td></td>
</tr>
</tbody>
</table>
Deadlines at a Glance

- April 6\textsuperscript{th}: you know the DMC tasks and your team
- May 11\textsuperscript{th}: submission of your DMC solution to Robert and Heiko
- May 18\textsuperscript{th}: official submission of your DMC solution
- May 19\textsuperscript{th}: submission of your final report
- May 22\textsuperscript{nd}: final presentations
RapidMiner Analyst Certification

• Offered for the first 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
• Online 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:
  - $Q1: x \geq Q1$ holds for 75% of all $x$
  - $Q3: x \geq Q3$ holds for 25% of all $x$
  - $IQR = Q3 - Q1$
- Outlier detection:
  - All values outside $[\text{median} - 1.5\times IQR; \text{median} + 1.5\times IQR]$.
- Example:
  - $0, 1, 1, 3, 3, 5, 7, 42 \rightarrow \text{median}=3, Q1=1, Q3=7 \rightarrow IQR = 6$
  - 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
Literature & Slide Sources

• Albert Bifet: Adaptive Stream Mining
  – we will provide scans of important chapters in ILIAS

• Joao Gama: Knowledge Discovery from Data Streams
  – we will provide scans of important chapters in ILIAS
Software

• Powerful open-source data mining suite
• Download: http://rapid-i.com/
• We use the most recent version of RapidMiner Studio
  - You will obtain a license key
Questions?