Hello

• Heiko Paulheim
• Professor 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: Tuesdays, 9-10am
  – Please make an appointment via e-mail to Ms. Lermer
• Heiko will teach the lectures
Hello

• M.Sc. Nicolas Heist
• Graduate Research Associate
• Research Interests:
  – Semantic Web Technologies
  – Knowledge Graphs and Linked Data
• eMail: nico@informatik.uni-mannheim.de
• Nico will teach the exercises and co-supervise the 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 2019
  – 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
  - 100 % written exam
  - project is not graded, but mandatory!

- Project work
  - work on DMC tasks

- Presentations
  - up to three intermediate presentations
    - open questions, problems, current results (numbers!)
    - everybody has to present once during those presentations

- Final report
  - 10 pages
  - solutions, results, lessons learned

**different to last years!**
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)
  – 2018: 197 participating teams from 47 countries

• Timeline
  – DMC registration on March 5th
  – tasks are published on April 4th
  – submissions are due on May 16th (internal submission: May 13th)

• Further information: http://www.data-mining-cup.de/en
The Data Mining Cup

- 2017: both Uni Mannheim teams among top 10 (out of 202)
- 2018: team from Uni Mannheim scores 2nd place (out of 197)
- Prices are awarded at a conference in Berlin in June
  - Top 10 teams are invited to present their solutions
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.02.18</td>
<td>Lecture: Preprocessing</td>
</tr>
<tr>
<td>26.02.18</td>
<td>Lecture: Regression</td>
</tr>
<tr>
<td>05.03.18</td>
<td>Lecture: Anomaly Detection</td>
</tr>
<tr>
<td>12.03.18</td>
<td>Lecture: Ensembles</td>
</tr>
<tr>
<td>19.03.18</td>
<td>Lecture: Time Series</td>
</tr>
<tr>
<td>26.03.18</td>
<td>Lecture: Neural Networks</td>
</tr>
<tr>
<td>02.04.18</td>
<td>Lecture: Parameter Tuning</td>
</tr>
<tr>
<td>09.04.18</td>
<td>DMC intermediate presentation</td>
</tr>
<tr>
<td></td>
<td>- Easter Break</td>
</tr>
<tr>
<td>29.04.18</td>
<td>DMC intermediate presentation</td>
</tr>
<tr>
<td>06.05.18</td>
<td>DMC intermediate presentation</td>
</tr>
</tbody>
</table>

DMC task published on 04.04.

final DMC submission 16.05.
Deadlines at a Glance

- March 5\textsuperscript{th}: DMC registration
- April 4\textsuperscript{th}: you know the DMC tasks and your team
- May 13\textsuperscript{th}: submission of your DMC solution to Nico and Heiko
- May 16\textsuperscript{th}: official submission of your DMC solution
- May 20\textsuperscript{th}: submission of your final report
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:
  - 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 \([\text{median}-1.5\times\text{IQR} ; \text{median}+1.5\times\text{IQR}]\)

- Example:
  - 0, 1, 1, 3, 3, 5, 7, 42 → median=3, Q1=1, Q3=7 → IQR = 6
  - Allowed interval: \([3-1.5\times6 ; 3+1.5\times6] = [-6 ; 12]\)
  - 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