Data Mining Lab Course WS 2019/20

Introduction

L. Richter  D. Nechaev

Department of Computer Science
Technische Universität München

Wednesday, Oct 23th
Outline

Timeline

Organization

Grading

Data Set Presentation – Task1
Organization

- schedule
- meeting structure
- task for next week
- remaining for today
# Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct 23th</td>
<td>Intro</td>
<td>Dec 11th</td>
<td>Descriptive Mining V</td>
</tr>
<tr>
<td>Oct 30st</td>
<td>Data Set Presentation</td>
<td>Dec 18th</td>
<td>Predictive Mining I</td>
</tr>
<tr>
<td>Nov 6th</td>
<td>Data Set Selection</td>
<td>Jan 8th</td>
<td>Predictive Mining II</td>
</tr>
<tr>
<td>Nov 13th</td>
<td>Descriptive Mining I</td>
<td>Jan 15th</td>
<td>Predictive Mining III</td>
</tr>
<tr>
<td>Nov 20th</td>
<td>Descriptive Mining II</td>
<td>Jan 22th</td>
<td>Predictive Mining IV</td>
</tr>
<tr>
<td>Nov 27th</td>
<td>Descriptive Mining III</td>
<td>Jan 29th</td>
<td>Final Presentation</td>
</tr>
<tr>
<td>Dec 4th</td>
<td>Descriptive Mining IV</td>
<td>Feb 5th</td>
<td>Final Presentation</td>
</tr>
</tbody>
</table>
Slot Layout

- each lab meeting is ~90 minutes
- we plan to give mini-talks for selected topics
- group formation depending on the number of participants
- each slot: 5-10 min presentation, 10 min discussion
Documentation and Communication

- your work is accompanied by a wiki-based lab journal
- https://i12r-studfilesrv.informatik.tu-muenchen.de/wise19/dmlab
- the wiki is the place to document experimental set-ups, results, figures and decisions
- keep your entries up-to-date
- we can setup a slack channel if there is demand
Weekly Preparations

▶ prepare your group’s presentation with a clear and elaborate report in the wiki
▶ for the meeting prepare a compressed digest for a 5-10 min presentation
▶ have your wiki report ready until Tuesday morning to give us time to read
Grading Criteria I

- We award grades for groups based on: The whole semester performance: the complete wiki entries and all presentations.
- Presentation criteria: focus on things that matter, communicate you message in an easy and understandable way, it is important, that the audience can follow. Do not get lost in arcane details.
- Wiki criteria: volume, conciseness, clearness
Grading Criteria II

- The volume should reflect the amount of time and work you have spent on the topic.
- The conciseness refers to the level of details and the precision.
- The clearness: Is the red line of the work visible, is it easy to understand?

Therefore the wiki entries should have a meaningful structure and cover different levels of abstraction. You may reference white papers, manuals, tutorials and further reading. We record weekly notes for the final grading. Small deviations in a group can occur. If there are objections or group issues you can request an individual grading. In such a case please contact us as early as possible!
Data Set Presentation

▶ due to Oct 30th
▶ everybody has to present a data set
▶ we do NOT advise to take data sets from competitions – at least as your first choice
▶ present your data set in 3 min and answer the questions:
  ▶ **Why?** (do you find it interesting?)
  ▶ **What?** (is described by the data set, size, number of instances, contents, type of data, source of the data set . . . )
  ▶ **Goals?** (give some ideas for possible prediction tasks)
  ▶ put a more elaborate description in the wiki
▶ we will not analyze time series data, multi-relational data, proprietary or NDA-protected data or big data
What makes a good data set and where can I find one?

- decent size: should fit in your main memory
- can be up to several million instances
- more features are better
- anything that could be converted / represented as a table:
  - log messages
  - xml like, like Yelp
  - other short standardized texts
  - ...
- NGOs
- national and international data and statistics authorities
- companies, e.g. Netflix, Amazon, ...
Tools and Resources

- you are allowed to use any tool or programming language you have access to
- we recommend to have a look into: Python, Jupyter notebooks, Matplotlib and scikit-learn
- consider to check out Google colab
  https://colab.research.google.com
Requirements

▶ curiosity and common sense
▶ self-motivation
▶ basic computer skills like scripting to manipulate your data
▶ basic statistics and database knowledge may be helpful
Questions ?