Data Mining I

Student Projects
Student Projects

- **Goals**
  - Gain more practical experience with data mining
  - Get to know additional problem-specific
    - preprocessing methods
    - data mining methods

- **Expectation**
  - Select an interesting data mining problem of your choice
  - Solve the problem using
    - the data mining methods that we have learned so far plus some advanced problem-specific data pre-processing
    - other data mining methods which might be helpful for solving the problem and build on what we learned in class
Procedure

- Teams of three students
  1. realize a data mining project
  2. write a 12 page* summary of the project and the methods employed in the project
  3. present the project results to the other students (12 minutes presentation + 8 minutes discussion)

- Final mark for the course
  • 50 % written exam
  • 30 % written summary about the project
  • 20 % project presentation

* Including title, TOC, appendices etc.
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Tuesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.10.2014</td>
<td>Introduction to Student Projects</td>
<td>Project Work</td>
</tr>
<tr>
<td>26.10.2014 23:59</td>
<td>Submission of Project Outlines</td>
<td></td>
</tr>
<tr>
<td>27.10.2014</td>
<td>Feedback Student Projects</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>03.11.2014</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>10.11.2014</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>17.11.2014</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>23.11.2014 23:59</td>
<td>Submission of Project Reports</td>
<td></td>
</tr>
<tr>
<td>24.11.2014</td>
<td>-</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>01.12.2014</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
</tbody>
</table>
Where to find interesting Data Sets?

- The Data Hub
  - http://thedatahub.org/
  - Large European data set catalog
- Data.gov.uk, Data.gov.us
  - Public sector data provided by the UK and US governments
- KDD Cup and Data Mining Cup
  - http://www.data-mining-cup.de/
  - Data mining competitions providing data sets and solutions
- Stanford Large Network Dataset Collection
  - Web Graphs, Amazon Purchasing Data and Reviews
- Yahoo Webscope
  - http://webscope.sandbox.yahoo.com/
  - Topics: Web search, advertising, rating data
- Programmable Web
  - Website giving an overview about 4100 public Web APIs
Where to find Information about additional Methods?

1. Pang-Ning Tan, Michael Steinback, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.


Where to find Information about additional Methods?

– Check out the solutions to your problem that other people have tried.
  • for instance by looking at submissions of the KDD CUP or Data Mining Cup
  • or by using:
Some Project Ideas (not binding)

- **Web Log Mining**
  - Learn a classifier for the categorizing the visitors of your website.
  - What features matter? Number of pages visited, time on site, ..
    (Bing Liu Chapter 12.x)
  - Preprocess some web log data outside RapidMiner
  - Learn and evaluate classifier within RapidMiner

- **Wikipedia Contributors**
  - Examine the edit history of Wikipedia contributors
  - Cluster users by different attributes (no of edits, edits/day, topic, ...)
  - Learn a classifier for the categorizing Wikipedia contributors

- **Sentiment Analysis for Discussion Forum / Rating Site**
  - Are people positive or negative about topic / product? (Bing Liu 11.x)

- **SPAM Detection**
  - eMail, blog or discussion forum (Bing Liu 6.10, 11.9)
Some Projects realized in previous Semesters

- **Analysis of Training Data of a Fitness Center**
  - Find different customer groups by clustering exercise data
  - Find sequences of exercises that different customer groups like

- **Mannheim Police Reports**
  - Learn classifiers for police reports
  - Identify type of incident, severity of incident, location of incident

- **last.fm Playlist Analysis**
  - Cluster last.fm users according to the style of the songs they are listening to
  - Find commons sets of songs for the different clusters

- **Bundesliga Betting Rules**
  - Find rules that help you to predict the outcome of a bundesliga game

- **Sentiment Analysis of Tweets about Movies**
  - Learned classifier from IMDB movie reviews
  - Applied and tested with tweets afterwards

- **Classifying a Document’s Perspective**
  - using the example of Israeli – Palestinian Essays
Project Outlines

- 4 pages (exactly!) including title, using DWS master thesis layout
- due 26.10.2014 23:59
- send by eMail to Heiko, Volha, Robert & Oliver
- answer the following questions:

1. What is the problem you are solving?

2. What data will you use?
   - Where will you get it?
   - How will you gather it?

3. How will you solve the problem?
   - What preprocessing steps will be required?
   - Which algorithms you plan to use?
   - Be as specific as you can!

4. How will you evaluate, measure success?

5. What do you expect your results to look like?
Coaching Sessions

- We will give you tips and answer questions concerning your project.

- Registration via email is mandatory!
  - until Tuesday night!
  - including the questions that you like to discuss
  - including which session you prefer (Thursday B2/B3)

- We will assign you a time slot afterwards and inform you about the slot via email.

- Every team has to attend at least one coaching session!
Project Summaries

- 12 pages (exactly!) including title, TOC pages and references. Each extra page downgrades your mark by 0.3!
- due 23.11.2014 23:59
- send by email to Heiko, Volha, Robert & Oliver
- describe your solution including the steps to get there:
  1. Application domain and goals
  2. Structure and size of the data set
  3. Preprocessing
  4. Actual Data Mining
  5. Evaluation
  6. Results
- Requirements
  • You must use the DWS master thesis layout.
  • Please cite sources properly. Preferred citation style [Author, year].
  • Also submit your RapidMiner processes and (a subset) of your data.
Final Exam

- **Date:** 17.12.2014
- **Duration:** 60 minutes
- **Structure:** 5 - 6 open questions that
  - check whether you have understood the content of the lecture
  - require you to describe the ideas behind algorithms and methods
  - might require you to do some simple calculations