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 four 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 + 5 minutes discussion)

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

<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.10.2016</td>
<td>Introduction to Student Projects</td>
<td>Preparation of Project Outline</td>
</tr>
<tr>
<td><strong>Sunday, 30.10.2016, 23:59: Submission of Project Outlines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31.10.2016</td>
<td>Project Work</td>
<td>Feedback Project Outlines</td>
</tr>
<tr>
<td>07.11.2016</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>14.11.2016</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>21.11.2016</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>28.11.2016</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td><strong>Sunday, 04.12.2016, 23:59: Submission of Project Reports</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05.12.2016</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
  - Data mining competitions providing data sets and solutions
  - http://www.data-mining-cup.de/
- Kaggle
  - Website running commercial and educational data science competitions
  - Offers datasets as well as solutions for older competitions
  - https://www.kaggle.com/
- Programmable Web
  - Website giving an overview about 13000 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 as well as Kaggle
  - or search for relevant scientific papers using:
Some Projects realized in previous Semesters

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

- 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

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

- Bike rental analysis
  - Analysing temporal patterns
  - Demand forecasting
4 pages (exactly!) including title, using DWS master thesis layout (don’t use chapters!)
  • Include a project name and your team number on the first page!

due Sunday, 30.10.2016, 23:59

send by email to Heiko & 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 Monday 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 page and references ⇒ max. 9 pages text/tables/diagrams. Each extra page and each day of late submission downgrades your mark by 0.3!
  - Include your project name and your team number on the first page!
- due Sunday, 04.12.2016, 23:59, send by email to Heiko & Oliver
- Outline for project summaries:
  1. Application area and goals
  2. Structure and size of the data set
  3. Preprocessing
  4. Actual Data Mining
     1. including evaluation results
     2. including different approaches that you tried
  5. Discussion of 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.
Project Summaries Checklist

- You should look at and report P/R/F1 if your data is unbalanced, NOT only accuracy
- You should NOT balance your test data
- In general, you should NOT pre-process your test data in any way that requires knowing the class label
- You should NOT try to predict the class label based on attributes which are so closely related to the label that are unlikely to be available for unseen data
  - (example: don't predict rating of a movie or product using the number of votes for this same rating)
- You can balance your data in at least 2 ways – by sampling, or by filtering/appending the under-represented class
- Make sure you pre-process your data in the way required by the operator you use
- All pre-processing choices should be justified (e.g. "removing these attributes gives 2% increase in accuracy")
- Normalize your numeric data before you cluster (Important!)
Final Exam

- Date: 12.12.2015
- 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
    - which do not require a calculator
Team Assignment

- Find your team now!
- Then enter your team in the student/team matrix!
  - Only enter if you have a team (don’t make random crosses!)
  - There can only be **one cross per row** (you can’t be in two teams!)
  - There should be **four crosses per column** (four students per team!)

<table>
<thead>
<tr>
<th>Name / Team</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utzer, Ben</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Mustermann, Max</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Sampling, Susi</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Dent, Stu</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balance, Bobby</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Feature, Captain</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**No**