Data Mining I
Introduction to Student Projects

Heiko Paulheim
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 10 to 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
<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.10.2015</td>
<td>Introduction to student projects and group formation</td>
<td></td>
</tr>
<tr>
<td>02.11.2015</td>
<td>Submission of project outlines (Sunday, 23:59)</td>
<td>Discussion of project outlines</td>
</tr>
<tr>
<td>09.11.2015</td>
<td>Project work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>16.11.2015</td>
<td>Project work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>23.11.2015</td>
<td>Project work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>30.11.2015</td>
<td>Submission of project outlines (Sunday, 23:59)</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>07.12.2015</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>16.12.2015</td>
<td>Written Exam (Wednesday!)</td>
<td></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, Data Mining Cup, and Kaggle
  – http://www.data-mining-cup.de/
  – https://www.kaggle.com/competitions

• 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 Additional Information

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


Where to find additional information

• 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:

![Google Logo](image)
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

• US/UK 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
Some Projects Realized in Previous Semesters

• US Basketball Playoffs
  – Prediction of winners
  – Prediction of awards (e.g., outstanding players)

• Bike rental analysis
  – Analyzing temporal patterns
  – Demand forecasting

• Company KPIs
  – Analyzing resilience against crisis

• Gas station prices
  – Analyzing patterns
  – Making predictions and recommendations

• eBay car auctions
  – price prediction
Project Outlines

- 2-3 pages (sharp!) without title and TOC pages, DWS master thesis layout
- due 01.11.2015, 23:59
- send by eMail to Kiril, Oli, and Heiko
- 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

• Please send us an email that you want to attend a coaching session
  – until Monday night
  – including the questions that you like to discuss

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

• Each team needs to take part in a coaching session at least once
Project Summaries

• 10-12 pages (sharp!) without title and toc pages
• due 29.11.2015, 23:59
• send by email to Kiril, Oli, and Heiko
• 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 Rapid Miner processes and (a subset) of your data.
Checklist for Project Summaries

• You should look at and report P/R/F1 if your data is unbalanced, NOT only accuracy
• You can optimize the performance with respect to F1 or cost, 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 low-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"
Final Exam

• Date: 16.12.2013
• Duration: 60 minutes
• Structure: six 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 pocket calculator
Questions?