Data Mining

Introduction to the Student Projects
Outline

1. Requirements for Student Projects
2. Requirements for Project Reports
3. Final Exam
Student Projects

- **Goals**
  - Gain practical experience with the complete data mining process
  - 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, including
      - proper parameter optimization
      - problem-specific pre-processing and smart feature creation
    - additional data mining methods which might be helpful for solving the problem and build on what we learned in class
Procedure

- Teams of six students
  1. realize a data mining project
  2. write a 10 page summary of the project and the methods employed in the project
  3. present the project results to the other students (10 minutes presentation + 5 minutes discussion)

- Final mark for the course
  - 30 % written summary about the project
  - 10 % project presentation
  - 60 % written exam
<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>03.04.2019</td>
<td>Introduction to Student Projects</td>
<td>Preparation of Project Outline</td>
</tr>
<tr>
<td></td>
<td><strong>Sunday, April 7th 2019, 23:59: Submission of Project Outlines</strong></td>
<td></td>
</tr>
<tr>
<td>10.04.2019</td>
<td>Lecture Association Analysis</td>
<td>Exercise Association Analysis</td>
</tr>
<tr>
<td></td>
<td><strong>Feedback Student Projects</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(13:45-15:15)</em></td>
<td></td>
</tr>
<tr>
<td>06.05.2019</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>13.05.2019</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>20.05.2019</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td></td>
<td><strong>Sunday, May 26th 2019, 23:59: Submission of Project Reports</strong></td>
<td></td>
</tr>
<tr>
<td>29.05.2019</td>
<td>Presentation of Project Results</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(10:15 to 15:15)</em></td>
<td></td>
</tr>
<tr>
<td>03.06.2018</td>
<td>Exam</td>
<td></td>
</tr>
</tbody>
</table>
Where to find interesting Data Sets?

- **KDnuggets Dataset List**
  - References to various data catalogs and datasets

- **Data.gov, data.gov.uk, govdata.de**
  - Public sector data provided by the government bodies

- **Programmable Web**
  - Website giving an overview about 13000 public Web APIs

- **Kaggle**
  - Website running commercial and educational data science competitions
  - Offers datasets as well as solutions for older competitions
  - https://www.kaggle.com/
  - Please compare your results to results from the competition’s forum

- **KDD Cup and Data Mining Cup**
  - Data mining competitions providing data sets and solutions
  - http://www.kdd.org/kdd-cup
  - https://www.data-mining-cup.com
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 discussion groups
  - or search for relevant scientific papers using Google™
Some Project Ideas (not binding)

- **Web Log Mining**
  - Learn a classifier for the categorizing the visitors of your website.
  - Which 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 / Hoax Articles**
  - Examine the edit history of Wikipedia contributors
  - Cluster users by different attributes (no of edits, edits/day, topic, ...)
  - Or learn a classifier for the categorizing Wikipedia contributors

- **Sentiment Analysis for Discussion Forum / Rating Site / Tweets**
  - 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

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

- 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

- Analysis of Training Data of a Fitness Center
  - Find different customer groups by clustering exercise data
  - Find frequent combinations of exercises

- 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

- maximum 4 pages including title page, using DWS master thesis layout
  - Include a project name and your team number on the first page!
- due Sunday, April 7th 2019, 23:59
- send by eMail to Chris, Anna, 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?
     1. What preprocessing steps will be required?
     2. Which algorithms do you plan to use?
     • Be as specific as you can!
  4. How will you measure success? (Evaluation method)
  5. What do you expect your results to look like? (Model/Clusters/Patterns)
- Feedback about your project outlines: Wednesday, 10.04.2019, 13:45-15:15
Coaching Sessions

- We will give you tips and answer questions concerning your project.
- **Registration via email** to Oliver & Anna 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 Report

- 10 pages (exactly!) plus references page, no appendix ➔ document length: 11 pages
- Each extra page and each day of late submission downgrades your mark by 0.3!
- due Sunday, May 26th 2019, 23:59
- send by email to Chris, Anna & Oliver
- Outline for project report:
  1. Application area and goals
  2. Structure and size of the data set (minimum 1 page)
  3. Preprocessing and Mining
     - describe different approaches and parameter settings that you tried
     - including evaluation setup and evaluation results
     - including discussion of the results
- Requirements
  1. You must use the latex template of the Springer Computer Science Proceedings
  2. Please cite sources properly. Preferred citation style [Author, year]
  3. Also submit your RapidMiner processes and (a subset) of your data
  4. Include your names and your team number on the first page!
my title

author1 name¹ and author2 name¹,²

¹ University 1
e-mail1@gmail.com
² University 2
e-mail2@gmail.com


1 Introduction


http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines
Severe Errors to Avoid

1. Normalize numeric data before calculating any similarity metrics

2. If your data is unbalanced
   • balance your training data
   • do NOT balance your test data
   • report P/R/F1, not accuracy
Final Exam

- **Date:** Monday, June 3\textsuperscript{rd}
- **Duration:** 60 minutes
- **Structure:** 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
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 **six crosses per column** (six 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>x</td>
<td>x</td>
</tr>
<tr>
<td>Sampling, Susi</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Dent, Stu</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Balance, Bobby</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature, Captain</td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

No

Yes