Data Mining

Introduction to the Student Projects
Outline

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

– **Goals**
  - Gain practical experience with the complete data mining process
  - Get to know additional problem-specific
    - preprocessing methods
    - data mining methods

– **Expectation**
  - You select an interesting data mining problem of your choice
  - You 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
     (online, for instance via MS Teams or other collaboration tools)
  2. write a 10 page summary of the project and
     the methods employed in the project
     (online, for instance using Overleaf)
  3. present the project results to the other students
     • via MS Teams, 10 minutes presentation + 5 minutes discussion

- Final mark for the course
  • 20 % written summary about the project
  • 5 % project presentation
  • 75 % written exam (online, open book)
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.05.2021</td>
<td>Introduction to Student Projects and Group Formation</td>
<td>Preparation of Project Outline</td>
</tr>
<tr>
<td></td>
<td><strong>Sunday, May 16th 2021, 23:59: Submission of Project Outlines</strong></td>
<td></td>
</tr>
<tr>
<td>19.05.2021</td>
<td><strong>Feedback on Project Outlines (if required, via Zoom)</strong></td>
<td>Project Work</td>
</tr>
<tr>
<td>27.05.2021</td>
<td>Project Work</td>
<td>Feedback on demand (via Zoom)</td>
</tr>
<tr>
<td>02.06.2021</td>
<td>Feedback on demand (via Zoom)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Sunday, June 13th 2021, 23:59: Submission of Project Reports</strong></td>
<td></td>
</tr>
<tr>
<td>16.06.2021</td>
<td>Presentation of Project Results</td>
<td>Presentation of Project Results</td>
</tr>
<tr>
<td>23.06.2021</td>
<td>Final 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
- 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
- Kaggle
  - Website running commercial and educational data science competitions
  - Offers datasets as well as solutions for older competitions
  - https://www.kaggle.com/
  - You must compare your results to results from the competition’s forum!
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 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 categorizing Wikipedia contributors

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

- **Estimate House or Car Prices**
  - using different regression methods or transfer learning to localize method
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

- **Transfer Learning for 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 using Springer Computer Science Proceedings layout or Word
  - Include a project name and your team number on the first page!
- due Sunday, May 16th 2021, 23:59
- send by eMail to Chris, Anna, Ralph
- 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 if required: Wednesday, 19.05.2021, 10:15-11:45
Coaching Sessions

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

- **Registration via email** to Anna & Ralph 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.

- Coaching sessions will take place via Google Hangout
  - We will send you the URL of the Hangout together with the information about the time slot of your meeting.

- Every team has to attend at least one coaching session!
Some Project Management Hints

- Organize your project in multiple iterations
  - Every artefact will be improved over time!
- Parallelize tasks while keeping centrally track of results
  - e.g. one central document with results plus reference to exact version of the processes/notebooks/datasets that produced these results
  - sub-groups should explore specific ideas for a specified amount of time
- Define concrete milestones: When should what be finished?
  - e.g. 22.5.21 Data exploration results collected in single document
  - e.g. 26.5.21 Subgroup on sentiment lexica adds results to central document
- Infrastructure
  - use shared folder for result document, versions of data, processes, slideset (e.g. MS Teams, Google Drive, github)
- Get simple process running early on to have a baseline
Tasks within the Iterations of the Project

1. Data Exploration and Visualization

2. Data Preprocessing: Value normalization, deal with outliers, deal with missing values, feature generation, balance training data if necessary

3. Establish/update baseline (majority class, mean value)

4. Try different learning methods using different feature creation methods and feature combinations

5. Perform error analysis in order to understand what is going on!

6. Later iteration:
   1. Perform automatic hyperparameter optimization and attribute selection
   2. Employ more sophisticated evaluation setup: x-val + holdout vs. nested x-val
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, June 13th 2021, 23:59
- send by email to Chris, Anna, Alex & Ralph
- Outline for project report:
  1. Application area and goals (0.5 pages)
  2. Profile (structure and size) of your data set (minimum 1 page)
  3. Preprocessing and Mining
     - describe different approaches and parameter settings (parameter optimization) that you tried
     - including description of evaluation setup (split, x-val, nested-x-val?) and evaluation results
     - including discussion of the results and an analysis of the errors still made by the best method (minimum 2 pages)

- Requirements
  1. You must use the latex template of the Springer Computer Science Proceedings
  2. Please cite sources properly and use your references page
  3. Also submit your Python code or 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

Checklist

- **Business Understanding:**
  - What is the actual problem (in the domain)?
  - What is the target variable?
    - Classification/Regression/Clustering?

- **Data Understanding:**
  - What is the distribution of labels / target variable?
  - Are all attributes and their types listed and important attributes explained?
  - What is the quality of the data?
  - What does correlation analysis reveal about attribute importance?

- **Preprocessing**
  - Are missing values replaced (in case needed)?
  - Checked for outliers (and handled them)?
  - Validity tests of attributes (Height above sea level < 9000)?
  - Check for inconsistencies (age=42, birthday=03/07/1997)
  - Check for duplicates
  - Data normalization
  - Additional features generated?
  - Has binning been tried out?
  - Feature subset selection implemented?
Checklist

- **External Knowledge:**
  - Are additional datasets used?

- **ML approaches:**
  - Which ML approaches were tried out?
  - How did you optimize hyperparameters (which attributes/ in which range / nested-x-val)?
  - Do you have at least one baseline (majority class / mean value / domain specific …)?

- **Evaluation**
  - Do you use fix train/test split, x-validation, or nested x-validation?
  - Is eval stratified?
  - Cost matrix or not?
  - Analyse a symbolic model (how does the decision tree / rules /… look like?)
  - What features do have a high impact on the result?
  - What types of errors are done by the best model? (error analysis)

- **Result**
  - Is the result critically evaluated?
  - Is the best result compared against the baseline?
  - What does the result mean given the problem? Could you use the model in practice?
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

3. Implement the recommendations concerning model evaluation, hyperparameter selection and feature selection given on the summary slides
Questions?
Final Exam

- Date: June 23\textsuperscript{rd}
- Duration: 60 minutes
- Location: Online
- Structure: 6 open questions that
  - check whether you have understood the content of the lecture
    - we try to cover all major chapters of the lecture: clustering, classification, regression, association analysis, and text mining
  - require you to describe the ideas behind algorithms and methods
    - What is the advantage or problem of X?
    - How do methods react to special pattern in the data?
    - Given the following data. What happens?
  - might require you to do some simple calculations
    - you need to know the most relevant formulas
    - you are not allowed to use a calculator
Questions?