

Data Mining

Introduction to the Student Projects



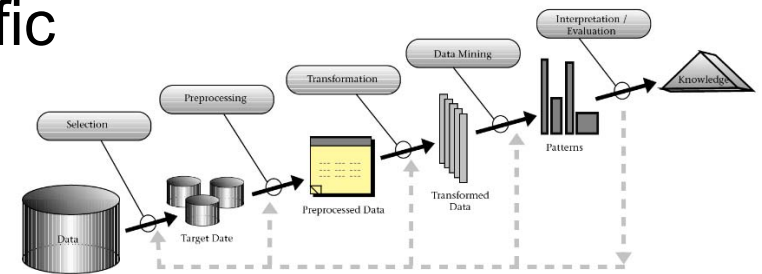
Outline

1. Requirements for Student Projects
2. Requirements for Project Reports
3. Final Exam
4. ~~Team Formation~~ (done already)

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 Google Hangout and Google Drive)
 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
 - Offline if possible again: 10 minutes presentation + 5 minutes discussion
 - Online if required: via WebEx, 10 minutes presentation + 5 minutes discussion
- Final mark for the course
 - 20 % written summary about the project
 - 5 % project presentation
 - 75 % written exam (hopefully offline in normal rooms again)

Schedule

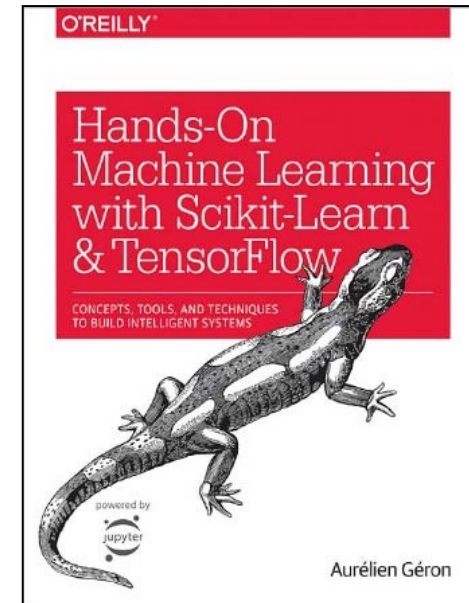
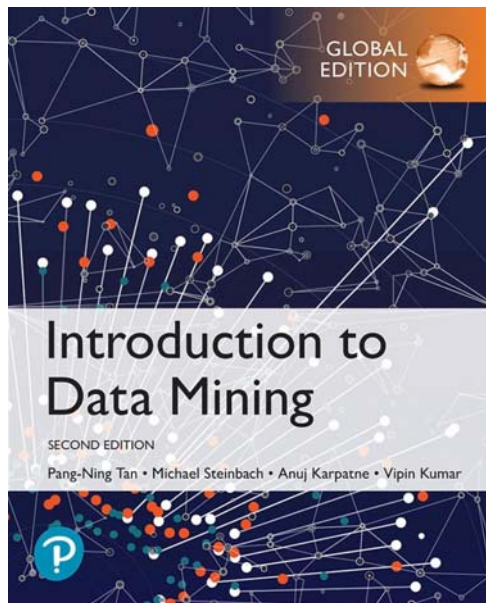
Week	Wednesday	Thursday
22.04.2020	Introduction to Student Projects and Group Formation	Preparation of Project Outline
Sunday, April 26th 2020, 23:59: Submission of Project Outlines		
29.04.2020	Feedback on Project Outlines (if required, via Hangout)	Project Work
06.05.2020	Project Work	Feedback on demand (via Hangout)
13.05.2020	Project Work	Feedback on demand (via Hangout)
20.05.2020	Project Work	
Wednesday, May 20th 2020, 23:59: Submission of Project Reports		
27.05.2020	Presentation of Project Results	Presentation of Project Results
08.06.2020	Final Exam	

Where to find interesting Data Sets?

- **KDnuggets Dataset List**
 - <https://www.kdnuggets.com/datasets/index.html>
 - 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!

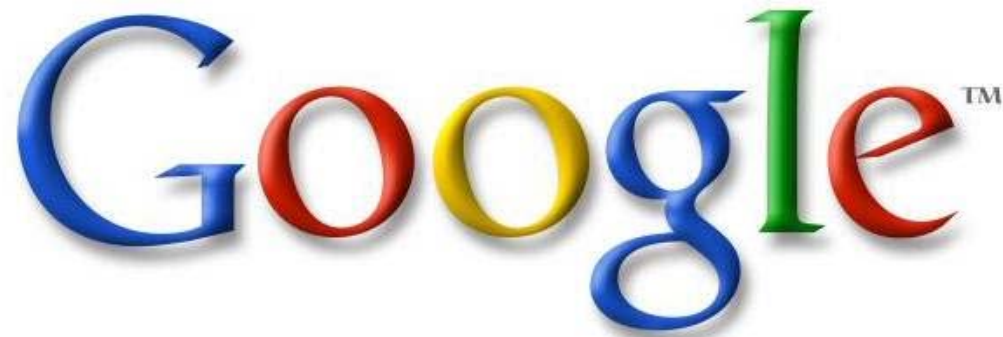
Where to Find Information about Additional Methods?

1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
2. Bing Liu: Web Data Mining, 2nd Edition, Springer.
3. Aurélien Géron: Hands-on Machine Learning with Scikit-Learn. O'Reilly.



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



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, April 26th 2020, 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, 29.04.2020, 13:45-15:15**

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!**

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 **Wednesday, May 20th 2020, 23:59**
- send by email to Chris, Anna & 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!

Template: Springer Computer Science Proceedings



<http://www.springer.com/de/it-informatik/Incs/conference-proceedings-guidelines>

Checklist

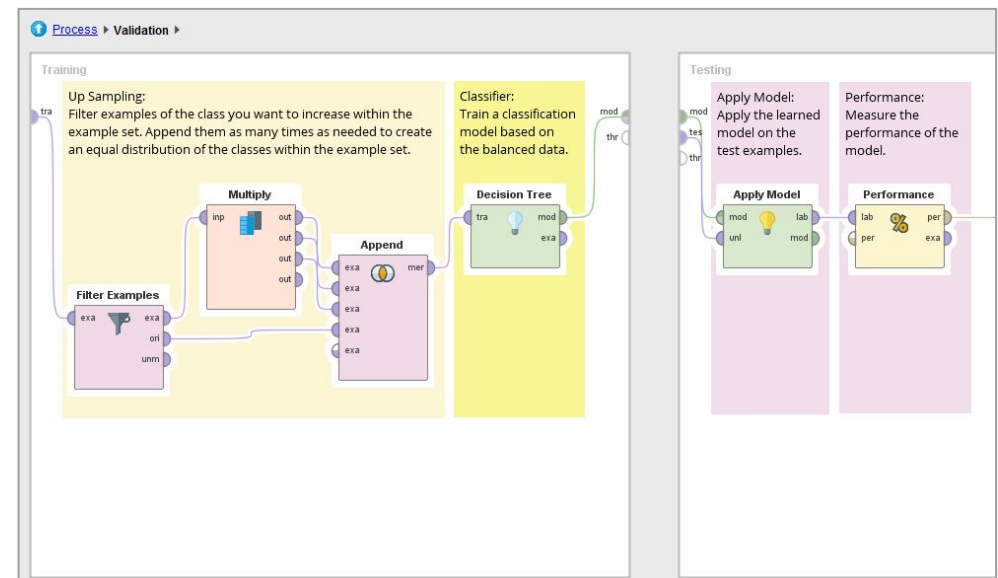
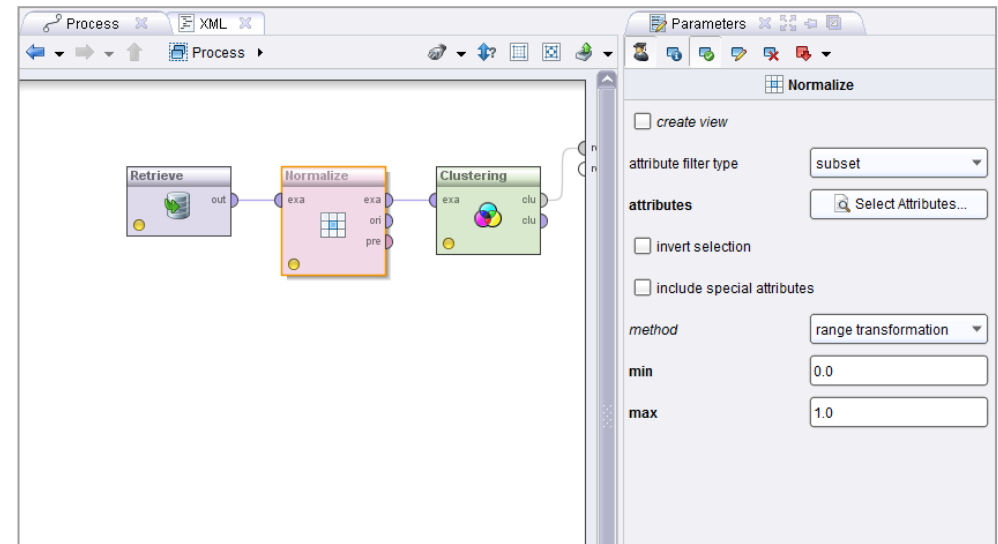
- Business Understanding:
 - What is the actual problem (in the domain)?
 - What is the target variable?
 - Clustering/Classification/Regression?
- Data Understanding:
 - What is the distribution of labels / target variable?
 - Are all attributes and their types listed and important attributes explained?
 - What is the quality of the data?
 - Are examples sorted (time series)?
- 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?
 - Correlation analysis implemented?
 - 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
 - Is there a train test split or 10 fold cross validation implemented?
 - 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 is critically evaluated?
 - Is the result analyzed 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 8th
- Duration: 60 minutes
- Location: We hope to be able to use the lecture halls again!
- 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?

Team Formation

- You are allowed to form teams of 6 students as you like!
 - You tell us which 6 students are in your team, we note this down
- We will form teams out of the remaining students who did not find a team by themselves.
- We hope to be able to do the team assignment in person in the session on 22.04.2020.
- If this won't be possible due to the Corona situation, we will ask you to send us the names of your team members by email before the kick-off session on 22.4.2020 which we would do via ZOOM or WebEx in this case.

Team Formation

- 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!)

Name / Team	1	2	3
Utzer, Ben		x	
Mustermann, Max	x	{	x
Sampling, Susi		x	
Dent, Stu	x		
Balance, Bobby		x	
Feature, Captain		x	

