Data Mining 1

Introduction to the Student Projects

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Outline

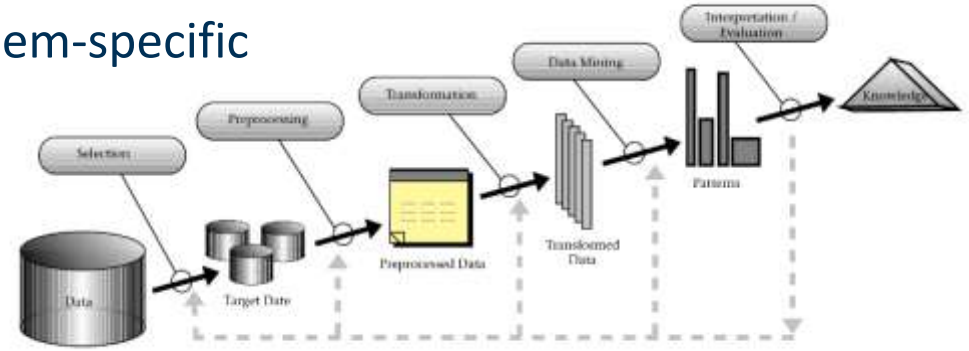
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

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

Student Projects

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **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 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

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Teams of **five** students
  - realize a data mining project
  - write a 12 page summary of the project and the methods employed in the project
  - present the project results to the other students (10 minutes presentation + 5 minutes discussion)

- Final mark for the course
  - 20 % written summary about the project
  - 5 % project presentation
  - 75 % written exam

Schedule

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Week	Wednesday	Thursday
26.10.2020	Introduction to Student Projects	Exercise: Classification 2
Monday, November 2nd, 2020, 23:59: Submission of Project Proposals		
02.11.2020	Lecture: Regression	Project feedback
09.11.2020	Project feedback	Exercise: Regression
16.11.2020	Lecture: Text Mining	Project feedback
23.11.2020	Project feedback	Exercise: Text Mining
30.11.2020	Lecture: Association Analysis	Presentation of project results
Wednesday, December 23th, 2020, 23:59: Submission of Project Report		

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Public sector data**
  - US government: https://www.data.gov
  - UK government: https://data.gov.uk
  - EU: https://www.europeandataportal.eu
  - CIA World Fact Book: https://www.cia.gov/library/publications/the-world-factbook/
  - Health data (over 125 years): https://www.healthdata.gov/
- **Data registries**
  - Datasets hosted on Amazon AWS https://registry.opendata.aws
  - Million Song Dataset, 1000 Genome Project, database of satellite imagery of Earth from NASA, Web Crawl
  - Google's Dataset Search: https://datasetsearch.research.google.com/
  - Microsoft Datasets: https://msropendata.com/
  - Dataset collection on Github: https://github.com/awesomedata/awesome-public-datasets
  - Data Hub: http://datahub.io
  - Linked Open Data Cloud: http://lod-cloud.net/

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Knowledge graphs**
  - Wikidata: https://www.wikidata.org
  - BabelNet: https://babelnet.org
  - DBpedia: http://wiki.dbpedia.org
- **Language resources**
  - WordNet: https://wordnet.princeton.edu
  - EuroWordNet: http://projects.illc.uva.nl/EuroWordNet/
  - Project Gutenberg (36.000 ebooks): http://www.gutenberg.org/
  - New York Times (starts 1851): http://developer.nytimes.com/docs
  - Wikitionary: https://www.wiktionary.org as KG: http://kaiko.getalp.org/about-dbnary/
- **Competitions**
  - Kaggle: https://www.kaggle.com/
  - Data Mining Cup: http://www.data-mining-cup.de
  - KDD Cup: https://www.kdd.org/kdd-cup
  - DrivenData: https://www.drivendata.org
  - CrowdAnalytix: https://www.crowdanalytix.com

Where to find interesting Data Sets?

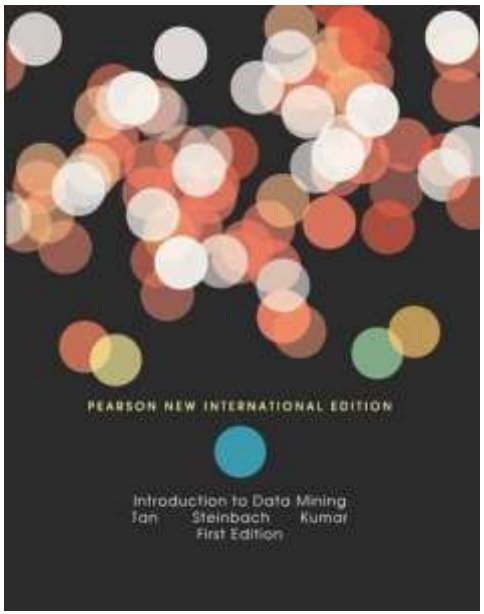
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

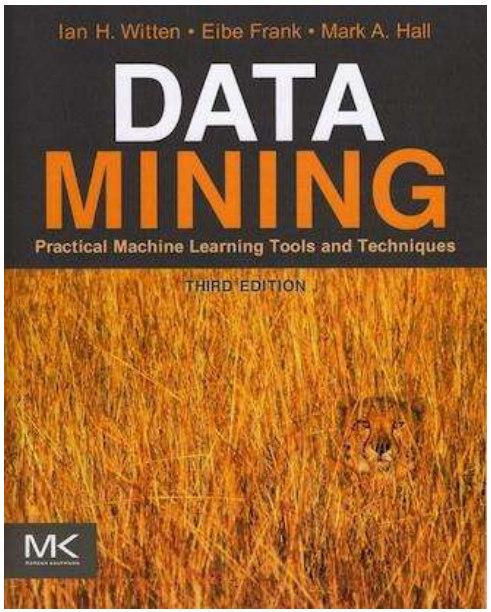
- **Covid19**
  - Johns Hopkins University https://github.com/CSSEGISandData/COVID-19
  - Our world in data: https://github.com/owid/covid-19-data
  - ECDC https://www.ecdc.europa.eu/en/covid-19/data
  - Harvard: https://dataverse.harvard.edu/dataverse/covid19

Where to find Information about additional Methods?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Pang-Ning Tan, Michael Steinback, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
- Ian H. Witten, Eibe Frank, Mark A. Hall: Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann.
- Bing Liu: Web Data Mining, 2nd Edition, Springer.





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 Project Ideas (not binding)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 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

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- maximum 4 pages including title page, using DWS master thesis layout
  - Include a project name and your team number on the first page!
- due **Monday, November 2nd, 2020, 23:59**
- send by eMail to Heiko, Nico, Ralph & Sven
- 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 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)

Coaching Sessions

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 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
- 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

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 12 pages (exactly!) including title/toc page and reference page
  - max. 10 pages, no appendix
  - Each extra page and each day of late submission downgrades your mark by 0.3!
- due Wednesday, December 23th, 2020, 23:59
- send by email to Heiko, Nico, Ralph & Sven

Project Report

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Outline for project report:

1. Application area and goals (Business Understanding)
2. Structure and size of the data set (minimum 1 page) (Data Understanding)
3. Preprocessing
4. Data Mining
   - (External Knowledge)
   - ML approaches
   - Evaluation
5. Results

Project Report

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Requirements

1. You must use the DWS master thesis layout.
2. Please cite sources properly. Preferred citation style [Author, year].
3. Also submit your RapidMiner processes/Python scripts and (a subset) of your data.
4. Include your project name and your team number on the first page!

Checklist

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

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

- Data Understanding:
  - Are examples sorted (time series)?
  - What is the distribution of labels / target variable?
  - Are all attributes and their types listed?
  - Are attributes explained?
  - What is the quality of the data?

Checklist

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 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?

- External Knowledge:
  - Are additional datasets used?

Checklist

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- ML approaches
  - How many different ML approaches were tried out?
  - Do you have at least one symbolic and one non symbolic approach?
  - 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?
  - Are the hyper parameters tuned (in which range / which attributes) ?
  - Are the tests systematic?
  - Analyse a symbolic model (how does the decision tree / rules /… looks like)
  - What features do have a high impact on the result?

Checklist

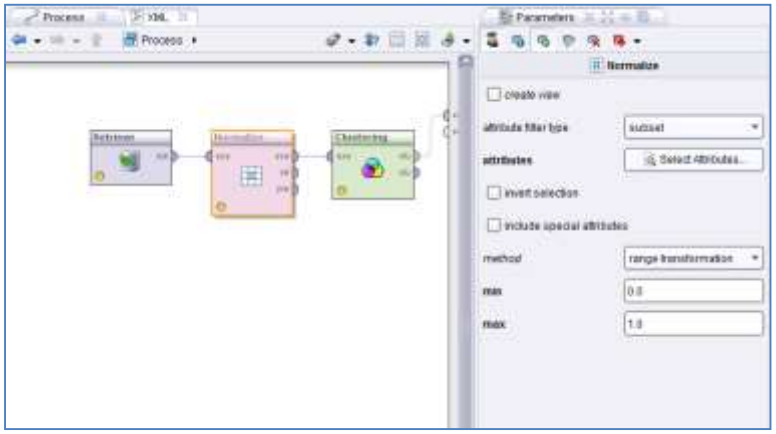
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- 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 it)

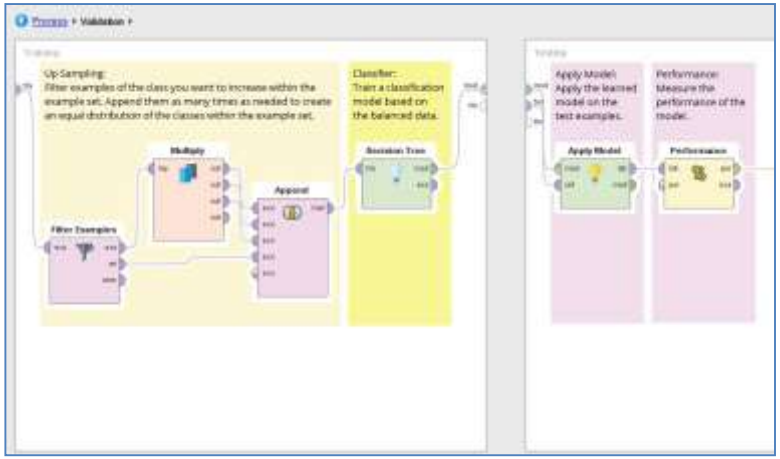
Deadly Errors to Avoid

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Normalize numeric data before calculating any similarity metrics



- If your data is unbalanced
  - balance your training data
  - do NOT balance your test data
  - report P/R/F1, not accuracy



Final Exam

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Date: Thursday, December 10th, 2020
- 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