Data Mining Cup 2022

Statistics: Michel Lang, Jonas Rieger, Steffen Maletz

Computer Science: Emmanuel Müller, Benedikt Böing, Simon Klüttermann

DMC 2022

- We are planning the event as an on-site course.
- We will switch to an online course only if the situation requires it.
- Predictive modeling competition from the field of online marketing
  - http://www.data-mining-cup.de/
  - Training dataset + unlabeled test data for prediction.
  - Optimize against specified quality measure
- International competition
  - 2021: 115 teams from 86 universities in 28 countries
  - 2020: 162/126/35, 2019: 149/114/28, 2018: 193/148/47, 2017: 202/150/48
- (Successful) history of Dortmund statisticians
  - 2010: Second Place , 2011: First Place, 2013: First and Second Place
  - 2020: First and 6th Place (joint team with computer science department),
  - several top 10 occurrences
- Prize money (2000/1000/500 €)

Statistical Methods

- EDA (Explorative Data Analysis)
- Preprocessing (Imputation, ...)
- Resampling and Evaluation
- Discriminant Analysis
- Nearest Neighbours
- Trees and Forests
- Support Vector Machines
- Regularized Linear Models
- Gradient Boosting
- Neural Networks
- Hyperparameter optimization
- Feature Selection
- Feature Generation
- Ensembles and Stacking
- [...]

Software

- Version management using GitHub
- Visualization (interactive)
- data.table / SQL
- Parallel computing (local/cloud)
- Machine Learning frameworks
  - e.g. mlr3 in R or scikit-learn in Python
- Modern ML packages
  - e.g. ranger, xgboost, glmnet
- Slack for communication
- Zoom for potential remote meetings

Requirements

- Master Statistik: Fallstudien I (recommended)
- Master Econometrics: Minor Introductory Case Studies
- Master Data Science:
  - All requirement courses (Introductory Case Studies, ...) must have been passed
  - Advanced Statistical Learning is recommended to be passed

- Computer Science: Big Data Analytics, Mathematics Courses

Course Plan

- 12 participants from Statistics department and 16 from Computer Science
- now: Remote Kickoff-Meeting
- March 1: Registration start DMC (2 team “leader” who register)
- March 21 - March 25, 9am - 4pm: In-person “Bootcamp” week
  - **Student presentation on selected methods/concepts** (max. 20 min)
  - March 21st, Kickoff-Meeting with Prof. Müller for students from CS department
  - April 5: **for students from Statistics department: Submission of short reports** (max. 10 pages)
- April 12: Start of competition, release of data and task by the prudsys AG
- Regular in-person meetings (2 per week), **active participation**
  - Wednesday 2pm - 4pm, CDI 121
  - Thursday 10am - 12pm, CDI 121
- June 28: End of competition, upload of predictions for test data
- (July 14: Award ceremony)
- August 31: **Statistics: Final Report** (~ 25 pages)

Examination Statistics

- presentation (max. 20 min) within the bootcamp week
  - explain the method/concept to the other students
  - it is not that relevant, that you are able to proof some (possibly complicated) theoretical properties
  - more important: properties that are relevant in practical usage, e.g. outlier handling...
  - how does your method/concept contributes to the Data Mining Cup?
- short report (max. 10 pages) - deadline: April 5 – no extension!
  - figures and tables count towards the page limit! Number of pages from the introduction until the last page of the conclusion is relevant for the page count; title page, contents page and bibliography do not count towards page limit
  - scientific explanation of the presented method/concept
  - if applicable: application of your method/concept to the example dataset "titanic" (https://www.kaggle.com/c/titanic/)
  - structure of report (examplarily): 1 Introduction, 2 Methods, 3 Application on Titanic Dataset, 4 Conclusion/Discussion
  - take the rules as teached in case studies (Fallstudien I), Introductory Case Studies or similar... (or maybe thesis) as a guideline
- active participation in competition and discussions
- final report (~ 25 pages, we will announce specific formalia for this report at the end of the competition) - deadline: August 31 – no extension!

Examination Computer Science

- active participation in competition and discussions
  - initiative for open tasks
  - imagination for what could be useful tasks
  - take and fill necessary roles in team
  - think both in and beyond your team
- final presentation at the end of the competition
  - explanation of task, teams and your role in the DMC
  - outline how your team's process going from early to later solutions
  - explain team's contributions to the final solution

To Do

- Wednesday 2pm - 4pm and Thursday 10am - 12pm ok?
- determine 2 team “leader” for registration (March 1)
- assignment of methods/concepts to students

- Questions?

- Keep yourself up to date using Slack. We will share all information there.