Instructor & class information

Instructor: Vencislav Popov, PhD

Contact: v.popov@psychologie.uzh.ch

Class time: Tuesdays, 10:15am – 11:45am

Class format: Online via Zoom

https://uzh.zoom.us/j/91354029952?pwd=aWZVT29TbTgvb3o5MnRsWDd0TlJtQT09

GitHub repository: https://github.com/venpopov/DataSciencePsychUZH

Class overview

This course will present a broad overview of advanced statistical methods required for research in modern psychology. Students will get a hands on experience with the entire process of data analysis, starting with project organization, data preprocessing, data exploration and visualization, statistical inference and prediction, avoiding overfitting, model selection, and project presentation. An emphasis will be put on understanding the general linear model and its extensions such as mixed-effects modeling. A strong focus will be put on creating reproducible workflows that take the student from a raw dataset to a completed research product. Students will get hands on experience with processing and analyzing data in the R statistical environment. Students will also become familiar with tools that aid reproducibility such as github and version control, Rstudio projects, notebooks and markdown files.

Learning goals

- understand and practice proper data organization, archiving and cleansing
- learn how to create reproducible data processing workflows
- learn how to use extensive visualization techniques in order to gain a deeper understanding of your data
- understand the difference between inference and prediction, between regression and classification problems, and learn how to select the proper statistical tools for each
- understand model overfitting and how to avoid it

Class structure

Each class will involve a mixture of segments of lectures, breakout discussions and Q&A sessions. Except for the first week, students are expected to have finished the assigned readings and tutorials prior to the corresponding class. Students should submit two questions on the assigned readings by 2pm on the Monday before each class. Classes will have the following structure:

- 1-3 short lecture segments (20-40 minutes total)
- 1-3 Q&A sessions interspersed between lecture segments (20 minutes)
- Small groups break-out discussion sessions (see below for more information; 20 minutes)
- Summary of break-out sessions (10 minutes)

Discussion sessions

After the lecture segments and Q&A sessions, students will split into separate breakout groups of 4-5 to discuss the topic of the day. Each group should select one student who will summarize the discussion of their group for the rest of the class during the final summary sessions. Each group should after class submit a short written summary of their discussion. *If a student is unable to be present for the in-class discussions, they should contact the instructor beforehand to organize an alternative for in-class participation.*

Grading

Data Science is learned through practice and the grading of this course will reflect this. Grades will be composed of:

- 30% homeworks
- 15% submission of two questions on the weekly readings *prior to class*
- 15% participation in class discussions
- 40% final project

Final project

At the end of this class you will have to submit a final project. This project will involve an analysis of a dataset of your choosing, using the techniques acquired in the class. Ideally, the dataset you choose will either be part of research you are currently involved in, or related to your interests. You will have to select a dataset for your final project by April 13th. You should submit a short written summary of why you have chosen this dataset for your final project before the start of the April 13th class. The dataset must be sufficiently complex to allow you to use a variety of the methods presented in the class. You will be notified by April 19th whether your final project proposal is approved. If you do not have access to a dataset from your research, please contact the instructor to help you find an appropriate publicly available dataset for your final project. To complete the final project, you must submit a 10 minute pre-recorded video presentation by May 28th and a written description of the project in the form of a Jupyter notebook by June 6th.

Materials

The required readings must be completed ***before*** each class. These readings will involve a mixture of textbook chapters, journal articles, and/or tutorials (please see the Class schedule). Tutorials and journal articles will be available on OLAT and the class github page

Required textbook: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer. Freely available for download at https://www.statlearning.com/

- Most weekly required readings come from this textbook

Recommended textbook: Wickham, H. & Grolemund, G. (2016). *R for Data Science.* O'Reilly. Freely available at https://r4ds.had.co.nz/

- No required readings, but a great reference for how to use many of the techniques discussed in class with the R statistical programming language

Tutorials: the practical side of the course will involve tutorials related to the homeworks. These tutorials will be available as Jupyter notebooks on the class GitHub page

GitHub repository: https://github.com/venpopov/DataSciencePsychUZH

Class schedule

Week	Date	Topic	Readings	Homework Due
1	23.02.2021	Intro to Data Science		
2	02.03.2021	Data organization & reproducibility	Goodman et al (2016) Sandve et al (2013) Optional: Perkel (2018)	Homework 1
3	09.03.2021	Data structures and data cleaning	Wickham (2014); Mueller & Freytag (2003)	Homework 2
4	16.03.2021	Data visualization	TBD	Homework 3
5	23.03.2021	The bias-variance trade-off	James et al. Chapters 1 & 2	
6	30.03.2021	Linear models	James et al. Chapter 3	
7	13.04.2021	Mixed-effects models	TBD; http://mfviz.com/hierarchical-models/	Homework 4
8	20.04.2021	Classifiers (Part 1)	James et al. Chapter 4	
9	27.04.2021	Classifiers (Part 2)	James et al. Chapter 4	Homework 5
10	04.05.2021	Cross-validation	James et al. Chapter 5	
11	11.05.2021	Bootstrapping	James et al. Chapter 5	Homework 6
12	18.05.2021	Power analyses via simulation	Smith & Little (2018)	
13	25.05.2021	Model selection	James et al. Chapter 6	Homework 7
14	01.06.2021	Dimensionality reduction	James et al. Chapter 10	