Description Course Winter School: Statistical Learning and Prediction

Dates: February 1 & 2, 2022 (10-18h, Leiden, Pieter de la Court Building)

Description: Statistical learning refers to a vast set of tools for understanding data. Two classes of such tools can be distinguished: "supervised" and "unsupervised". Supervised statistical learning involves building a statistical model for predicting an output (response, dependent) variable based on one or more input (predictor) variables. There are many areas of psychology where such a predictive question is of interest. For example, finding early markers for Alzheimer's or other diseases, selection studies for personnel or education, or prediction of treatment outcomes. In unsupervised statistical learning, there are only input variables but no supervising output (dependent) variable; nevertheless we can learn relationships and structures from such data using cluster analysis and methods for dimension reduction. In this course we aim to give the student a firm theoretical basis for understanding and evaluating statistical learning techniques and teach the students skills to apply statistical learning techniques in empirical research. Techniques often referred to as 'machine learning' (though we prefer the name 'statistical learning', because the basic problems of sampling and probability have not changed with the newer algorithms) will be discussed, including: cross-validation, elastic net regression, clustering (k-means, hierarchical), partial least squares, k-nearest neighbours, decision trees, gradient boosting, random forests, smoothing splines and support vector machines.

Target group: PhD candidates

Teachers: Marjolein Fokkema, Tom Wilderjans

Method: Lab sessions, in which lectures are combined with computer assignments.

Preparation: Students receive a list of online videos and associated parts and exercises of the book of James et. al, 2021 (see below) that should be prepared before the course starts. This will take about 20 – 25 hours, depending on current statistical knowledge and R skills.

Assessment method: Active participation

Reading list

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R (Second Edition). New York: Springer. A free copy can be obtained online, https://www.statlearning.com/

Additional suggested material (not required)

- Berk, R. A. (2008). Statistical learning from a regression perspective. New York, Springer. (freely accessible within the Leiden network)
- Kuhn, M. & Johnson, K. (2013). Applied predictive modelling. New York, Springer. (freely accessible within the Leiden network)