EUI DEPARTMENT OF POLITICAL AND SOCIAL SCIENCES

Introduction Machine Learning for the Social Sciences

3rd term workshop 2021

Course Details:

Workshop:	5-day Workshop 9.00-12.00 am, 7th - 11th June Lukas F. Stoetzer
E-mail:	lukas.stoetzer@hu-berlin.de

Course Description:

This workshop provides an introduction to machine and statistical learning methods for postgraduate social scientists. The workshop covers the basic conceptualizations of statistical learning. It introduces supervised learning methods for regression tasks and classification tasks. The course also gives a brief overview of unsupervised learning methods.

The main goal is to provide students with an overview over common supervised and un-supervised learning techniques. It also provides an environment for students to learn the necessary skills to conduct first machine learning analyses using R.

Readings:

The workshop closely follows the following text book:

James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning with Applications in R.* New York: Springer, 2013.

The textbook gives an easy to follow introduction to the material and is available online. Students can also find pre-recorded lecture videos by the authors on this page. In addition, the ISLR Package contains most of the datasets used in the book.

Students who are interested in more advanced discussions of the material can consult:

Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2 ed. New York: Springer. 2009

Prerequisites:

There is formally no prerequisite for the workshop. However, students should have taken a statistic class that covers multivariate regression analysis and have some experience with R (or alternatively python) to solve the afternoon exercises.

Course Structure:

The workshop has five morning lecture/seminar sessions that are held each day from 9.00 - 12.00. The lecture will be held online via Zoom. As there is plenty of material to cover, students should read the respective book material beforehand and come prepared to the seminar sections.

In the afternoon, students are given a few exercises from the book that allow them to apply the learned methods to existing datasets and test their understanding of the conceptual material. As there are no lab sessions for the workshop, students should learn about the necessary R-Code by themselves. The book gives a good description at the end of each chapter and there are also pre-recorded lab-session videos online. It potentially makes sense to work on the exercises in small study groups.

Software:

We will work with the open-source statistical programming language R. It is particularly suited for carrying out basic statistical learning methods. If students wish to rely on other software programs, like python, they will find enough openly accessible material online.

Course Requirements:

Students have to hand in a machine learning project until September 2020, for which they will receive personalized feedback from the instructor. **For credits, students have to send a proposal of their final project at the end of the class on Tuesday June 11th**. This can be a short outline of what they intend to do for their final project (no more than one page). Students are free to choose any topic for which the methods introduced during the workshop are applicable. It makes sense to look for applications in their field of substantive interest (sociology, political science, history, economics).

Workshop Outline:

Day 1: Introduction

On the first day we will cover conceptual issues, talking about the difference between prediction versus explanation, supervised versus unsupervised Learning, the bias/variance trade-off, cross-validation and test-data and discuss first simple techniques such as regression, nearest neighbor and Bayes Classifier.

- Readings

James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. New York: Springer, Ch. 2 (15 - 52) and Ch.5 (176 - 184)

- Afternoon Study:
  - Conceptual: Ch.2 Ex. 3, 7 and Ch.5 Ex.3
  - Application in R: Ch.2 Ex. 10 and Ch.5 Ex.8

Day 2: Supervised Learning: Regression & Classification

On the second day we will discuss multivariate linear regression (a method you probably know from other stats classes) as one foundation of statistical learning. We discuss some parametric extensions to make it more flexible, like polynomials and interactions. We compare linear regression with nearest neighbors to see the merits of parametric assumptions in statistical learning. We will further discuss classification tasks, discussing first simple methods logistic regression and linear/quadratic discrimination analysis. These methods form the basis for some of the supervised learning techniques we will discuss in the next two days.

- Readings

  James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. New York: Springer, Ch. 3 (59 - 104) and Ch.4 (127 - 154)

- Afternoon Study:
  - Application in R: Ch.3 Ex. 15 and Ch.4 Ex.10

Day 3: Supervised Learning: Sparse and flexible Regressions

On the third day we discuss two statistical learning techniques for regression models. We start with selection and regularization, talking about the subset selection, shrinkage methods and dimension reduction methods. We further discuss ways to move beyond Linearity, focusing on regression splines, smoothing splines, local regression and generalized additive models.

- Readings

  James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. New York: Springer, Ch. 6 (203 - 243) and Ch.7 (265 - 287)

- Afternoon Study:
  - Conceptual: Ch.6 Ex. 3, 4
  - Application in R: Ch.6 Ex. 11 and Ch.7 Ex.9

Day 4: Supervised Learning: Tree based methods and Support Vectors Machines

On the fourth day we focus on two statistical learning techniques that have become very valuable in machine learning tasks for classification and regression. We first discuss tree- based methods, starting with decision trees and talk about bagging, random forests and boosting. Second, we discuss support vectors machines.

- Readings

  James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. New York: Springer, Ch. 8 (303 - 323) and Ch.9 (337 - 356)

- Afternoon Study:
  - Conceptual: Ch.8 Ex. 1 and Ch.9 Ex.3
  - Application in R: Ch.8 Ex. 9 and Ch.9 Ex.8

Day 5: Unsupervised Learning: Principal Component Analysis and Clustering

On the final day we will look at a number of methods for unsupervised learning tasks. We focus on Principal Component Analysis and Clustering Methods (Including K-Means and Hierarchical Clustering).

- Readings

  James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. New York: Springer, Ch. 10 (373 - 401)

- Afternoon Study:
  - Conceptual: Ch.10 Ex. 3
  - Application in R: Ch. 10 Ex. 9