Web Scraping & Machine Learning

Bundesbank Workshop

Prof. Dr. Frauke Kreuter
frauke.kreuter@uni-mannheim.de
A5, 6, entrance C, room 216
Phone: +49 (0)621 181-2003

Dr. Christoph Kern
c.kern@uni-mannheim.de
A5, 6, entrance C, room 011
Phone: +49 (0)621 181-2298

Malte Schierholz, M.Sc.
schierholz@uni-mannheim.de
A5, 6, entrance A, room 119
Phone: +49 (0)621 181-2791

Sebastian Sternberg, M.A.
ssternbe@mail.uni-mannheim.de
A 5, 6, entrance C, room 221
Phone: +49 (0)621 181-2542

Course Description: Given the intense activities and interactions on a multitude of web pages, vast amounts of data are available from various web resources. With the emergence of Big Data, these resources play an increasingly important role in scientific research. However, in order to collect and analyze data from the web, specific computational tools are needed. In addition, new data sources can also induce a shift in analytical goals, putting more emphasis on exploratory and/or predictive modeling.

This workshop provides an introduction to web scraping (I), supervised (II) and unsupervised (III) machine learning using R. The first part of the course exemplifies how data can be captured from the web efficiently and discusses the most common standards of data exchange (XML, JSON, APIs). The second part introduces supervised machine learning as a potential means for analyzing data from a prediction perspective. In this context, classification and regression trees, bagging, random forests and boosting methods will be presented. The third part of this course introduces unsupervised learning techniques such as principle component analysis and clustering methods, with which patterns in the data can be detected. In the practical sessions of both machine learning modules, scraped data from the first part of the course will be used.

References:

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2017). *Big Data and Social Science: A Practical Guide to Methods and Tools.* Boca Raton, FL: CRC Press Taylor & Francis Group.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York, NY: Springer.

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

Course Outline

Date	Slot	Content
15.11.	10:00–12:30	Introduction (Frauke Kreuter) Web Scraping (Malte Schierholz) • HTML, XML, JSON, APIs • Regular Expressions • Practical session
	13:30–17:00	Supervised Learning I (Christoph Kern) • Machine Learning Basics (CV, Performance Measures) • Decision Trees (CART) • Practical session
16.11.	10:00–12:30	Supervised Learning II (Christoph Kern, Sebastian Sternberg) • Bagging • Random Forests
	13:30–17:00	• Boosting (AdaBoost, GBM, XGBoost) • Practical session
17.11.	10:00–12:30	Unsupervised Learning (Sebastian Sternberg, Christoph Kern) • Introduction • PCA, Distance Measures, Multidimensional Scaling • Practical session
	13:30–17:00	• Clustering Methods (K-Means, Hierarchical Clustering) • Practical session