Web Scraping & Machine Learning

20.10.2017

INNcentive Workshop für NachwuchswissenschaftlerInnen an der Universität Bremen

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

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 analytical and computational tools are needed.

This workshop provides an introduction to web scraping (I) and supervised machine learning (II) 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, random forests and boosting methods will be presented, drawing on scraped data from the first part of the course.

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

Slot		Content
1	11:00–12:30	Web Scraping (Malte Schierholz, M.Sc.) • HTML • XML • JSON • APIs • Regular Expressions
2	13:30–15:00	Supervised Learning I (Dr. Christoph Kern) • Machine Learning Basics • Cross-Validation • Performance measures • CART
3	15:30–17:00	Supervised Learning II (Dr. Christoph Kern) • Random Forests • Boosting Text Mining (Malte Schierholz, M.Sc.)