Web Scraping, Machine Learning & Deep Learning
Bundesbank Workshop

25th - 29th March 2019

Providers:

Prof. Frauke Kreuter   Dr. Christoph Kern
Marcel Neunhoeffer   Sebastian Sternberg

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 as well as deep learning (IV) 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 predictive 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. The fourth part is dedicated to recent advances in deep learning, including an introduction to the anatomy of neural networks, an introduction to Keras (a popular high-level neural network API). The last day of the workshop is dedicated to advanced deep learning techniques such as applications to text and sequence classification problems and Generative Deep Learning, a promising neural network type to produce artificial data. This workshop places particular emphasis on the practical application of these methods, which is why every module is accompanied by hands-on applications.

Prerequisites:

For the workshop, we will use the statistical software *R*. *R* is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. We expect the participants to familiarize themselves with *R* prior to the workshop, though knowledge of *R* is not a strict prerequisite but an advantage.

- Participants download *R* for free at `http://www.r-project.org/`.
- RStudio (the graphical user interface for *R*) can be downloaded at urlhttp://www.rstudio.com/.
- Free tutorials for *R* are available online, for instance at `https://www.rstudio.com/online-learning/` or `https://www.datacamp.com/courses/free-introduction-to-r`.

Workshop-Agenda

Day	**Time Slot**	**Content**
1	10.00-10.30	Workshop Opening (Prof. Frauke Kreuter) • Introduction to Big Data Analysis
	10.00-12.00	Web Scraping (Dr. Christoph Kern) • HTML, XML, JSON, APIs • Regular Expressions • Practical session
	13.00-16.00	Supervised Learning I (Dr. Christoph Kern) • Machine Learning Basics • Decision Trees (CART) • Practical Session
2	10.00-16.00	Supervised Learning II (Dr. Christoph Kern) • Bagging • Random Forests • Boosting (AdaBoost, GBM, XGBoost) • Practical Session
3	10.00-12.00	Unsupervised Learning I (Sebastian Sternberg) • Introduction • Principal Component Analysis, Distance Measures • Practical Session
	13.00-16.00	Unsupervised Learning II (Sebastian Sternberg) • K-Means Clustering • Hierarchical Clustering • Practical Session
4	10.00-16.00	Deep Learning (Marcel Neunhoeffer) • Introduction • Anatomy of Neural Networks • Introduction to Keras • Practical Session
5	10.00-16.00	Advanced Deep Learning (Marcel Neunhoeffer & Sebastian Sternberg) • Deep Learning for Text and Sequences • Practical Session • Generative Deep Learning (Generative Adversarial Networks) • Practical Session

References

[Allaire and CholletAllaire and Chollet2018] Allaire, J. and F. Chollet (2018). *Deep Learning with R.* Manning Publications Co.

[Hastie, Tibshirani, and FriedmanHastie et al.2011] Hastie, T., R. Tibshirani, and J. Friedman (2011). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* (2 ed.). New York: Springer.

[James, Witten, Hastie, and TibshiraniJames et al.2013] James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning.* New York: Springer.

[Kuhn and JohnsonKuhn and Johnson2013] Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling.* New York: Springer.