Discipline: Information Systems / Operations Research

1. Language

English

2. Title

Machine Learning

3. Lecturer

Professor Dr. Stefan Lessmann, School of Business and Economics, Humboldt-University of Berlin

https://www.wiwi.hu-berlin.de/de/professuren/bwl/wi/personen/hl

stefan.lessmann@hu-berlin.de

4. Date and Location

24. – 27. August 2020

Harnack House, Conference Venue of the Max Planck Society,

Ihnestr. 16-20, 14195 Berlin, Germany

https://www.harnackhaus-berlin.mpg.de/

5. Course Description

5.1 Abstract and Learning Objectives

The course exposes participants to recent developments in the field of machine learning and discusses their ramifications for business and economics. Machine learning comprises theories, concepts, and algorithms to infer patters from observational data. The prevalence of data ("big data") have led to an increasing interest in corresponding methodology to leverage existing data assets for improved decision-making and business process optimization. Concepts such as business analytics, data science, and artificial intelligence are omnipresent and ground to a large extend on machine learning. Familiarizing course participants with these concepts and enabling them to purposefully apply cutting-edge methods to real-world decision problems in management, policy development, and research is the overarching objective of the course. Accordingly, the course targets PhD students and young researchers who want to employ machine learning in their research. A clear and approachable explanation of relevant methodologies and recent developments in machine learning paired with a batterie of practical exercises using contemporary software libraries of (deep) machine learning will ready participants for design-science or empirical-quantitative research projects.

5.2 Content

The course provides participants a comprehensive overview of the state-of-the-art in machine learning and its applications in business and economics. To that end, the course splits into three parts.

Part I revisits fundamental concepts of machine learning including approaches for unsupervised, supervised, and reinforcement learning. Further topics of the first part comprise a discussion of the connections between machine learning and more traditional data analysis paradigms such as statistics and econometrics and the fundamental differences between data-driven models for descriptive, explanatory, predictive, and normative decision support. The overall objective of Part I is to re-introduce selected fundamentals of machine learning and relevant machine learning algorithms.

Part II examines recent developments in the scope of deep learning using artificial neural networks. Promising autonomous feature extraction, deep learning advances conventional approaches for (un-)supervised machine learning toward artificial intelligence. Deep learning has become the de facto standard for processing large unstructured data sources such as text and images. Following an introduction of deep neural networks, the course concentrates on approaches for processing sequential data using the example of textual data. Considering a piece of text as a sequence of individual tokens (i.e., words) ensures that the techniques covered in the course are readily applicable to other types of sequential data such as time series. At the same time, participants have an opportunity to explore state-of-the-art approaches for natural language processing.

Part III (Deep) machine learning algorithms have proven their ability to process large and heterogeneous high-dimensional data sets. Emphasizing scalability as design principle, machine learning has to a large extent focused on the extraction of correlational patterns. Econometricians have long criticized the inability of machine learning algorithms to capture causal relationships between variables of interest. Against this background, the third part of the course examines recent developments in the scope of causal machine learning. Considering the example of decision models in marketing, the course briefly revisits some fundamentals related to causal inference and elaborates on selected causal machine learning algorithms such as causal forests (Athey & Imbens, 2019) or the x-learner (Künzel et al., 2019).

5.3 Schedule (including start and end time)

<u>Day I</u>

- 08:00 – 09:00 Arrival of participants and registration
- 09:00 – 09:15 Welcome
- 09:15 – 10:45 Session I.1: Fundamentals of machine learning
- 10:45 – 11:00 Coffee break
- 11:00 – 12.30 Session I.2: Algorithms and libraries for supervised learning
- 12:30 – 13:30 Lunch break
- 13:30 – 15:00 Session I.3 Model validation and interpretability
- 15:00 – 15:15 Coffee break
- 15:15 – 17:30 Day I modeling task: Prediction of retail credit risk

Day II

- 09:00 – 09:15 Welcome and recap day I
- 09:15 – 10:45 Session II.1: Fundamentals of deep learning and neural networks
- 10:45 – 11:00 Coffee break
- 11:00 – 12.30 Session II.2: Deep learning for text data analytics I
- 12:30 – 13:30 Lunch break
- 13:30 – 15:00 Session II.3 Deep learning for text data analytics II
- 15:00 – 15:15 Coffee break
- 15:15 – 17:30 Day II modeling task: Prediction of online review sentiment

Day III

- 09:00 – 09:15 Welcome and recap day II
- 09:15 – 10:45 Session III.1: State-of-the-art approaches for text data
- 10:45 – 11:00 Coffee break
- 11:00 – 12.30 Session III.2: From machine learning to causal inference
- 12:30 – 13:30 Lunch break
- 13:30 – 15:00 Session III.3 Algorithms for causal machine learning
- 15:00 – 15:15 Coffee break
- 15:15 – 17:30 Day III modeling task: Uplift models for e-commerce analytics

Day VI

- 09:00 – 09:15 Welcome and recap day III
- 09:15 – 10:45 Session IV.1: Ethics in data analytics, machine learning and AI
- 10:45 – 11:00 Coffee break
- 11:00 – 12.00 Session IV.2: Machine learning research opportunities
- 12:00 – 13:00 Lunch break
- 13:00 – 14:00 Session IV.3 Presentation of modeling tasks for the final assignment
- 14:00 – 17:00 Session IV.4 Group work on machine learning assignment
- 17:00 – 17:30 Closing session

5.4 Course format

The course adopts a multi-facetted teaching concept combining conceptual lectures, discussion of research papers, reviews of programming codes and hands-on exercises using Python. Each of the three core parts is associated with modeling exercises using real-world data sets from fields such as marketing and credit risk analytics. The data will be provided in the course. In addition, the final exam will give students an opportunity to carry out an independent data analytic modeling task on their own data. This way, participants can readily apply the concepts covered in the lectures in realistic decision and research settings. The course language is English.

6. Preparation and Literature

6.1 Prerequisites

Master-level education in Business, Economics, Computer Science, Engineering or a related field.

Course participants should have experiences in computer programming, preferably in languages such as Matlab, Python, or R, which are commonly used for statistical computing. Practical exercises and assignments will use the Python programming language. Therefore, familiarity with Python and Jupyter Notebooks is particularly beneficial, but can also be obtained in the scope of a mini-assignment, which precedes the course.

6.2 Essential Reading Material

- Athey, S., & Imbens, G. (2019). Machine Learning Methods Economists Should Know About. CoRR, arXiv:1903.10075v1. https://arxiv.org/abs/1903.10075
- Dalessandro, B., Perlich, C., & Raeder, T. (2014). Bigger is better, but at what cost? Estimating the economic value of incremental data assets. Big Data, 2(2), 87-96. http://dx.doi.org/10.1089/big.2014.0010
- Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big Data, 6(1), 13-41. http://dx.doi.org/10.1089/big.2017.0104
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10), 4156-4165. https://arxiv.org/abs/1706.03461
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. http://dx.doi.org/10.1038/nature14539
- Nielsen, M. A. (2015). Neural Networks and Deep Learning. Free online book: Determination Press. http://neuralnetworksanddeeplearning.com/index.html
- Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of Causal Inference. Cambridge, MA, USA: MIT Press. Full-text available via https://mitpress.mit.edu/books/elements-causal-inference
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. Sebastopol, CA, USA: O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/

6.3 Additional Reading Material

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning: MIT Press. https://www.deeplearningbook.org/
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. New York: Springer. http://appliedpredictivemodeling.com/
- Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2019). Targeting customers for profit: An ensemble learning framework to support marketing decision-making. Information Sciences, online first, https://doi.org/10.1016/j.ins.2019.05.027
- Varian, H. R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3-28. http://www.aeaweb.org/articles?id=10.1257/jep.28.2.3

6.4 To prepare

Participants are expected to study the essential reading material. Familiarity with literature from the additional reading material list is beneficial. The PhD course *Data Science as a Research Method*, which is also offered in the VHB ProDok lecture series, provides an excellent foundation for the course.

To prepare for the practical exercises and course assignment, participants are required to familiarize themselves with the Python programming language and Jupyter notebooks. To that end, participants will receive a set of programming tasks – mini-assignment – prior to the course. Completion of these tasks will ensure that participants are well-prepared for programming exercises. Participants will receive the mini-assignment 4 weeks before the course and can submit their solution up to one week before the course to obtain some feedback. Participation in the mini-assignment and submission of a solution is highly recommended. However, the mini-assignment will not be graded and is not a mandatory prerequisite to participate in the course.

Mini-Assignment timeline:

Announcement of mini-assignment task: 20. of July 2020

Deadline for submission of a solution in the form of a Jupyter notebook: 15. August 2020

Submissions that fail meeting this deadline will not receive any feedback

7. Administration

7.1 Max. number of participants

The number of participants is limited to 20.

7.2 Assignments

A mini-assingment precedes the course. Completion of that mini-assignment is highly recommended but not mandatory. Valid submissions will receive some feedback.

During the course, there will be a couple of in-class exercises, allowing participants to verify their ability to apply concepts covered in the course to real-word settings. The exercises will not be graded.

7.3 Exam

After the course, participants are required to complete a machine learning assignment and write-up results in the form of a Jupyter Notebook. The schedule of the course leaves some time to start working on the assignment during the course. The notebook will be graded.

7.4 Credits

The course is eligible for 6 ECTS

8. Working Hours

Working Hours	**Stunden**
Mandatory readings	40 h
Mini-assignment related to Python programming (to be completed before the course)	40 h
Active participation in class	30 h
Final exam (practical assignment to be completed and written-up after the course)	70 h
SUMME	**180 h**
ECTS: 6	