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

Three-week online course

April 6th – 23

The course will be offered in an electronic format. Participants receive pre-recorded videos of lecture and tutorial sessions to familiarize themselves with relevant machine learning concepts and their practical application using Python. In addition, several video conferences will be held over the three week course period to support participants with their mastery of course concepts and to facilitate discussion and networking.

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 patterns from observational data. The prevalence of data (“big data”) has led to an increasing interest in the 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 in decision-makers’ mindset and ground to a large extent 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 Ph.D. students and young researchers with a general interest in algorithmic decision-making and/or concrete plan 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 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 objective of Part I is to (re-)introduce selected fundamentals of machine learning and relevant machine learning algorithms. We emphasize techniques for supervised machine learning, which are most relevant for machine learning-oriented research in business and economics.

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 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 covers selected topics in machine learning research. (Deep) machine learning algorithms have proven their ability to process large and heterogeneous high-dimensional data sets. Emphasizing scalability as a 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). Another typical critic machine learners are facing concerns a lack of interpretability. Machine learning models are often considered black boxes. However, recent research has proposed a set of explanation methods for understanding and diagnosing such models. Acknowledging the cruciality of explaining model-based recommendations in many applications of machine learning, Part III of the course will look into the field of interpretable machine learning and equip students with a solid understanding of the options to explain model predictions.

5.3 Course Schedule

The course consists of ten lecture (L) and five programming sessions (P), which split into the three topic blocks introduced above. In addition, several live meetings are offered via Zoom to discuss the course content and questions.

The course follows a master schedule, which suggests a period of one week to complete the sessions of each course block. Consequently, the online course is offered over a period of three weeks. The agenda

of each course block and week together with the approximated time invest to complete sessions is provided below.

Participants can depart from the master schedule and are invited to complete course sessions at their own pace. However, the video sessions are offered over the three-week window of the course.

Block I: Fundamentals of Machine Learning

- L.I.1: Introduction to machine learning (60 min)
- L.I.2: Basic algorithms for supervised learning (90 min)
- P.I.1: Data integration and preparation (90 min)
- L.I.3: Machine learning model validation (90 min)
- L.I.4: Advanced learning algorithms (90 min)
- P.I.2 Prediction of retail credit risk (90 min)

Block II: Artificial Neural Networks for deep learning and text analytics

- L.II.1: Introduction to neural networks (90 min)
- P.II.1: Neural networks in Python (60 min)
- L.II.2: Neural networks for sequential & textual data (90 min)
- P.II.2: Fundamentals of natural language processing (60 min)
- L.II.3: State-of-the-art models for text analysis (90 min)
- P.II.3 Prediction of online review sentiment (90 min)

Block III: Selected topics in machine learning research

- L.III.1: Interpretable machine learning (90 min)
- L.III.2: From machine learning to causal inference (90 min)
- L.III.3: A more formal perspective on causal ML (90 min)
- P.III.1: Uplift models for e-commerce analytics (90 min)

Weekly schedule of video conferences (via Zoom) over the course period:

- Monday, 10.00 – 12.00: Introduction to the current course block
- Wednesday, 17.00 – 18.00: After-work discussion, room for student presentation
- Friday, 10.00 – 12.00: Discussion of the current block

The above schedule repeats every week. In general, video conferences are offered to support course participants in their learning experiences. All conference sessions and especially those on Wednesday evenings shall also foster discussion, an exchange of ideas, and facilitate networking. Participants are invited to attend all video conferences. However, attendance of the sessions is not mandatory. In fact,

one reason for offering a relatively large amount of online time is to offer participants flexibility. For example, questions related to one block of the course can be asked in the Friday session, which completes that blog or the Monday session introducing the next course block.

Provided sufficient interest among participants, the Wednesday evening session will include a short presentation of a research project from one or a few participant(s) in which s/he has used one of the machine learning methods covered in the corresponding course block.

5.4 Course format

The course adopts a multi-faceted teaching concept combining conceptual lectures, discussion, 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 their research. 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 will benefit from some experiences with 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 the course.

6.2 Essential Reading Material

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. New York: Springer. http://appliedpredictivemodeling.com/
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. http://dx.doi.org/10.1038/nature14539
- 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

6.3 Additional Reading Material

- 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
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning: MIT Press. https://www.deeplearningbook.org/

- 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
- Athey, S., & Imbens, G. (2019). Machine Learning Methods Economists Should Know About. CoRR, arXiv:1903.10075v1. https://arxiv.org/abs/1903.10075
- 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
- Knaus, M. C., Lechner, M., & Strittmatter, A. (2018). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. CoRR, (arXiv:1810.13237).
- 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
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. Sebastopol, CA, USA: O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/
- 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 essential reading material. Familiarity with literature from the additional reading material list is beneficial. The Ph.D. 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 receive a set of data science tasks and solutions to prepare themselves for programming sessions. Two Jupyter notebooks explaining the tasks are available at:
https://github.com/stefanlessmann/VHB_ProDoc_ML

Participants will receive a demo solution for these tasks four weeks before the start of the course and can compare their codes against that solution.

7. Administration

7.1 Max. number of participants

The number of participants is limited to 20.

7.2 Assignments

None

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. Typically, each participant will work on a different modeling

task. Ideally, the assignment task connects to a research project that the participant is involved. Alternative assignment topics include the replication of a published machine learning paper or working on a Kaggle competition (http://www.kaggle.com). The schedule of the course leaves room for discussing possible topics for the assignment. Student will submit their solution to the assignment roughly six weeks after the end of the course period. The submitted notebooks will be graded according to the quality of the exposition, the complexity of the modeling tasks, and the degree to which machine learning concepts have been used successfully.

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	