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

September 20th – 23th, 2022

The course will be offered over a four-day period comprising lecture, tutorial, and discussion sessions.

Harnack-Haus
Ihnestr. 16-20
14195 Berlin

5. Course Description

5.1 Abstract and Learning Objectives

The course exposes participants to recent developments in the field of machine learning (ML) and discusses their ramifications for business and economics. ML comprises theories, concepts, and algorithms to infer patterns from observational data. The prevalence of data (“big data”) has led to an increasing interest in ML 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 ML. Familiarizing course participants with these concepts and enabling them to apply cutting-edge ML algorithms 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 plans to employ ML in their research. A clear and approachable explanation of relevant methodologies and recent ML developments paired with a batterie of practical exercises using contemporary software libraries for (deep) ML 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 ML and its applications in business and economics. To that end, the course splits into three parts.

Part I introduces ML and discusses connections to other data analysis paradigms such as statistics and econometrics. We also elaborate on the fundamental differences between data-driven models for descriptive, explanatory, predictive, and normative decision support. Thereafter, we revisit traditional methods such as kMeans, logistic regression, and decision tree, and discuss their extensions in state-of-the-art ML algorithms. We emphasize techniques for supervised machine learning, which are arguably especially relevant for ML-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 ML approaches 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 neural networks, the course concentrates on deep learning approaches for natural language processing. While concentrating on the example of text processing, the techniques covered in the course are readily applicable to other types of sequential data such as, for example, time series.

Part III covers selected topics in ML research. (Deep) machine learning algorithms have proven their ability to process large and heterogeneous high-dimensional data sets. Emphasizing scalability as a design principle, ML has largely focused on the extraction of correlational patterns. Econometricians have long criticized the inability of ML techniques to capture causal relationships. Against this background, the third part of the course examines recent developments in the scope of causal ML. Considering the example of decision models in marketing, the course briefly revisits some fundamentals related to causal inference and elaborates on selected causal ML algorithms. Another typical critic machine learners face concerns a lack of model interpretability. ML 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 fields, Part III of the course will investigate the field of interpretable ML and equip students with a solid understanding of the options to explain model predictions.

5.3 Course Schedule

The course consists of several lecture (L) and programming (P) sessions.

Pre-course stage			
		Study papers from reading list	
		Familiarize with Python and Jupyter notebooks	
		Study pre-course notebooks, which exemplify ML foundations	
Day 1			
		Arrival of participants	
09:00	10:30	Welcome and introduction	
10:30	11:00	Coffee break	
11:00	12:30	L.I.1	Introduction to machine learning
12:30	13:30	Lunch break	
13:30	15:15	L.I.2	Basic algorithms for supervised learning
15:15	15:45	Coffee break	
15:45	17:30	L.I.3	Machine learning model validation
Day 2			
09:00	10:30	L.I.4	Advanced algorithms for supervised learning
10:30	11:00	Coffee break	
11:00	12:30	P.I.1	Prediction of retail credit risk
12:30	13:30	Lunch break	
13:30	15:15	L.II.1	Introduction to neural networks
15:15	15:45	Coffee break	
15:45	17:30	P.II.1	Neural networks in Python
Day 3			
09:00	10:30	L.II.2	Neural networks for sequential & textual data
10:30	11:00	Coffee break	
11:00	12:30	P.II.2	Fundamentals of natural language processing
12:30	13:30	Lunch break	
13:30	15:15	L.II.3	State-of-the-art models for text analysis
15:15	15:45	Coffee break	
15:45	17:30	P.II.3	Prediction of online review sentiment
Day 4			
09:00	10:30	L.III.1	Interpretable machine learning
10:30	11:00	Coffee break	
11:00	12:30	L.III.2	Causal machine learning
12:30	13:30	Lunch break	
13:30	15:15	L.III.3	Discussion of the course assignment
15:15	15:30	Closing remarks and farewell	
Post-course stage			
4 to 6 weeks		Development of a Jupyter notebook demonstrating the use of ML in research. Specific tasks will be agreed with participants and should ideally display a strong link to the participant's Ph.D. topic.	

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 pre-course stage.

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
- 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

- 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
- 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 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 pre-course notebooks that exemplify selected foundations of Python programming and ML. The pre-course pack is available at https://github.com/stefanlessmann/VHB_ProDoc_ML.

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 corresponds to a scope of 6 LP/ECTS.

8. Working Hours

Working Hours	**Stunden**
Mandatory readings	30 h
Preparation for programming part / study of pre-course Jupyter notebooks	30h
Active participation in class	30 h
Final exam (practical assignment to be completed and written-up after the course)	90 h
TOTAL	**180 h**