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 17 – 20, 2024
   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 extract patterns from observational data. The prevalence of data (“big data”) has led to a sure in the 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 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 prescriptive decision support. Thereafter, we revisit important ML practices and algorithms; from established industry workhorses like logistic regression to state-of-the-art boosting machines. The course emphasizes techniques for supervised machine learning, which we consider 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. 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 (NLP). Examining the kind of models that run the auto-complete function of the Google search engine or on your smartphone, participants obtain a solid understanding of modern NLP including word embeddings, transfer learning, and cutting-edge transformer architectures like the famous BERT model. While concentrating on example from the realms of NLP, we also establish the similarity between text and other forms of sequential data. This enables participants to readily apply the learnt concepts to a range of other types of data, most prominently time series.

Part III covers selected topics in ML research. (Deep) ML 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 selected marketing decision models as an example, the course revisits some fundamentals related to causal inference and elaborates on recently proposed techniques for causal ML. A related critic machine learners face concerns a lack of model interpretability. ML models are often black boxes. 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 explainable AI 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

### Day 1
- **09:00** to **10:30**: Arrival of participants
- **10:30** to **11:00**: Welcome and introduction
- **11:00** to **12:30**: Lecture (L.I.1) - Introduction to machine learning
- **12:30** to **13:30**: Lunch break
- **13:30** to **15:30**: Lecture (L.I.2) - Basic algorithms for supervised learning
- **15:30** to **16:00**: Coffee break
- **16:00** to **17:30**: Programming (P.I.1) - Data integration & preparation using Python

### Day 2
- **09:00** to **10:30**: Lecture (L.I.3) - Machine learning model validation
- **10:30** to **11:00**: Coffee break
- **11:00** to **12:30**: Lecture (L.I.4) - Advanced algorithms for supervised learning
- **12:30** to **13:30**: Lunch break
- **13:30** to **15:30**: Programming (P.I.2) - Prediction of retail credit risk
- **15:30** to **16:00**: Coffee break
- **16:00** to **17:30**: Lecture (L.II.1) - Introduction to neural networks

### Day 3
- **09:00** to **10:30**: Lecture (L.II.2) - NLP foundations & Word2Vec
- **10:30** to **11:00**: Coffee break
- **11:00** to **12:30**: Lecture (L.II.3) - State-of-the-art models for text analysis
- **12:30** to **13:30**: Lunch break
- **13:30** to **15:30**: Programming (P.II.1) - Fundamentals of natural language processing
- **15:30** to **16:00**: Coffee break
- **16:00** to **17:30**: Programming (P.II.2) - Prediction of online review sentiment

### Day 4
- **09:00** to **10:30**: Lecture (L.III.1) - Interpretable machine learning
- **10:30** to **11:00**: Coffee break
- **11:00** to **12:30**: Lecture (L.III.2) - Causal machine learning
- **12:30** to **13:30**: Lunch break
- **13:30** to **15:30**: Closing session: Discussion of the course assignment & next steps

### 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 programming demos and exercises using real-world data sets from fields such as marketing and credit risk analytics. The data will be provided in the course.

The final assignment will give students an opportunity to further develop their practical skills by working on an ML-related task, which can be connected to their Ph.D. 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.

Practical exercises and the course assignment involve Python programming. We assume that course participants are familiar with Python and the Python ecosystem for data science including, for example, Libraries like NumPy, Pandas, and Sci-Kit learn, and, importantly, Jupyter Notebooks. A basic understanding of these technologies is sufficient.

6.2 Essential Reading Material


6.3 Additional Reading Material


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 might find the following textbook useful:


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 computational essay (i.e., 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](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

<table>
<thead>
<tr>
<th>Working Hours</th>
<th>Stunden</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandatory readings</td>
<td>20 h</td>
</tr>
<tr>
<td>Preparation for programming part / study of pre-course Jupyter notebooks</td>
<td>40 h</td>
</tr>
<tr>
<td>Active participation in class</td>
<td>30 h</td>
</tr>
<tr>
<td>Final exam (practical assignment to be completed and written-up after the course)</td>
<td>90 h</td>
</tr>
<tr>
<td>SUMME</td>
<td>180 h</td>
</tr>
</tbody>
</table>