PADERBORN UNIVERSITY

Organizational Course Information

Explainable Artificial Intelligence

Dr. Stefan Heindorf

Outline

- Organization
- Introduction

General information

Course L.079.05806

- Lectures, mini project. Dr. Stefan Heindorf
- Tutorials. Parsa Abbasi
- Languages. English, Python

Information

- https://en.cs.uni-paderborn.de/ds-jrg/teaching/courses/xai-s25
- https://paul.upb.de and https://panda.upb.de

Time and location

- Lectures (as of April 8). Tuesday 16:15 – 17:45, Room F1.110
- Tutorials (as of April 14 bi-weekly). Monday 09:15 – 11:45, Room F2.211
- Mini Project (as of April 15). Tuesday 17:45 – 19:15, Room F1.110

Consultation

- After lecture
- Via Panda forum
- Set up appointment with me via email (heindorf@upb.de) [in exceptional cases]

Web resources

PAUL

- General. Standard course information
- Registration. Module, course, course achievement, exam

PANDA

- General. All announcements, asynchronous Q&A (forum), anonymous feedback
- Lectures and mini project. Slides
- Assignments. Sheets

Jupyter Server

- Local installation of “python” and “jupyter lab” (e.g., via Anaconda)
- Google Colab (https://colab.research.google.com)

Course achievement: mini-project

Requirements

- Course content
- Working knowledge of Python

Pass for mini-project

- Link to commented and executable code (GitLab/GitHub)
  (only main branch is considered)
- README with installation and usage details (as markdown)
- Project report with results (as pdf, uploaded to GitLab/GitHub)
- Teams of up to three persons
- Deadline: June 30, 2025
- Submission: mail to heindorf@upb.de, Subject: [XAI] Mini Project Submission, Name, imt accounts, and matriculation number of team members

Plagiarism / AI Tools

- Your group needs to solve the task on your own
- You need to specify your sources and resources
- Plagiarism Leaflet
- Guideline for the use of text-generating AI tools

Tutorials

- Discuss exercise sheets (typically 2-3 tasks)
- Solutions provided by students
- If nobody has a solution, you can use the time to work on task on-site (no official solutions provided)
- Bonus points
  - Present solution in tutorial
  - Hand in solution to exercise and permission to share
  - Email to parsa.abbasi@uni-paderborn.de
    Subject: [XAI] Exercise Sheet <Number>: Solution
  - Deadline: day of the tutorial, before end of tutorial
  - Team size: 1 person
  - Bonus: improve grade by one increment (0.3/0.4) if solution is satisfactory

Presentations of research papers

Students will present research papers in lecture

- 6 research papers (from predefined list or own suggestions)
- 3 lectures towards the end of the course (beginning of June)

Each research paper

- 25 minute presentation (with slides uploaded to Panda)
- 10 minute activity (e.g., quiz on paper, simple task)
- 10 minutes questions

All students are expected to prepare & participate in class discussions for all papers

- Read the paper in advance
- Prepare questions about the paper
- What do you like/dislike about the paper?

Organization

- Bonus: improve grade by one increment (0.3 / 0.4)
- Team size: up to 3 persons

Registration for research papers

- Sign up for research paper via email **after** Wednesday, April 16, 8:00
  - Email: heindorf@upb.de, Subject: [XAI] Paper Presentation: Registration
  - Send names of all team members and paper title
  - First come, first serve (I will confirm your topic via email and send you the date)
- Send your (preliminary) outline and responsibilities until April 27
  - Email: heindorf@upb.de, Subject: [XAI] Paper Presentation: Outline
  - Outline: Table of contents (TOC) of presentation
  - For each TOC entry, specify the estimated number of minutes
  - For each TOC entry, specify the person responsible
- Send draft of slides to me until June 11
  - Email: heindorf@upb.de, Subject: [XAI] Paper Presentation: Draft
  - Feel free to send (unofficial) draft and questions earlier

Tips for presentations of research papers

Tips and support for paper presentations

- General tips for presentations: https://webis.de/downloads/lecturenotes/lecturenotes-generic/unit-en-oral-presentations.pdf
- You can talk to me if you have specific questions
- You can send your slides to me to receive feedback in advance

Paper presentations

- Most important: convey intuition and general idea
- Provide good summaries of what's in the paper
- Figures and examples help

Exam

Requirements

- Pass the course achievement (“Studienleistung”): mini-project

Format

- Oral exam (online or in person)
- Content = lecture + exercises + mini project + paper presentations
- Preliminary dates
  - Tuesday, July 22
  - Tuesday, August 26
  - Wednesday, September 10
  - Thursday, September 11

Appointment

- Email: to mone@uni-paderborn.de
- Subject: [XAI] Oral Examination
- State whether you would prefer an examination online or in person
- State your matriculation number

Bonus (max. 0.7 intotal)

- Improve grade by up to one increment (0.3/0.4) for presentation in tutorials
- Improve grade by up to one increment (0.3/0.4) for paper presentation
- Bonus only applicable if exam passed with at least 4.0

Topic

This course

- Explainable artificial intelligence
- Builds upon machine learning (ML)
- Knowledge of basics in ML expected
  There will be a high-level recap, but not more
- Programming skills expected (Python)

Recommended courses before (alternatively)

- Machine learning for biometrics. Dr. Philipp Terhörst
- Foundations of knowledge graphs. Prof. Dr. Axel-Cyrille Ngonga Ngomo

Goals of this course

- Learn concepts and methods to explain machine learning models
- Explain black-box models
- Explain white-box models

Learning outcomes

After completing the module, students will be able to

- Recognize and discuss the importance of interpretability
- Explain and apply important explanation methods (e.g., interpretable models, model-agnostic methods, and model-specific methods)
- Recognize characteristics of datasets, machine learning tasks, and machine learning models in application problems and argue which explanation method is appropriate for a given problem
- Implement simple explanation methods from scratch extend and modify existing explanation methods
- Discuss problems and proposed solutions with experts in the field
- Read and discuss research literature in the area of XAI

Registration for module, course, and exam

To complete this course and module, you need to register for:

- Module: until April 30
- Course: until April 30
- Course achievement (Studienleistung): April 21 until May 21
- Exam: April 21 until May 21

General advice

- All registrations are done in PAUL. Each requires two clicks („Register“ and „Submit“)
- General Rule: If you see anything in PAUL that you can register for within this course or module, you should do so
- Regularly check the email address that PAUL sends its messages to

Contacts

- If anything looks suspicious in PAUL, contact the examination office
- If you need advice, contact study-service-cs@uni-paderborn.de or see office hours: https://cs.uni-paderborn.de/en/studies/study-service
- Important deadlines are also announced on the Fachschaft Discord server

Registration for course achievement (“Studienleistung”)

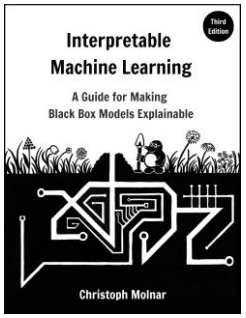
- You need to register for the course achievement between April 21 and May 21
- De-registration from the course achievement is possible until July 4
- If you did not pass the course achievement, you will be de-registered from the exam automatically
- After a course achievement has been entered into PAUL (even if it is entered as failed), the module can only be de-registered under certain conditions
- To pass the course achievement, you need to pass the mini project

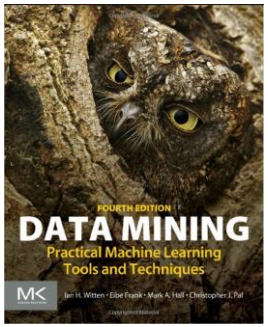
Registration for exam

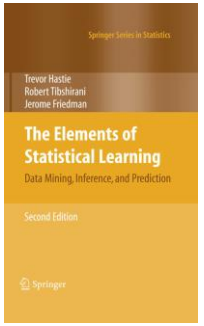
- Registration for the first phase is possible from April 21 and May 21
- Registration for the second phase is possible from September 1 till 5
- De-registration from the exam is possible until two days before the exam takes place
- Oral examinations will be offered between July and September

Literature and code basis

Books







- Interpretable Machine Learning (Molnar 2025) https://christophm.github.io/interpretable-ml-book/
- General ML books (Witten et al. 2016; Hastie 2009) https://link.springer.com/book/10.1007/978-0-387-84858-7
- Available in library and online

Conference and journal papers

- References to papers will occur in course content
- Most papers can be found online (e.g., at Google Scholar)

Code

- Different ML libraries available freely
- Papers often provide URL where code can be found
- Still, extensive own implementation needed in programming tasks

Lecture schedule (tentative)

Introduction

- April 8 Lecture 1 Organization

Interpretable models

- April 15 Lecture 2 Introduction to XAI
- April 22 Lecture 3 Linear regression
- April 29 Lecture 4 LASSO, Logistic regression, Decision trees
- May 6 Lecture 5 Decision tree, decision rules

Black-box models

- May 13 Lecture 6 Global model-agnostic methods (PDP, permutation feature importance surrogates)
- May 20 Lecture 7 Local model-agnostic method: ICE, LIME, Anchors
- May 27 Lecture 8 Local model-agnostic method: Shapley Values & SHAP
- June 3 Lecture 9 Local model-agnostic method: counterfactuals

Neural network models

- June 10 Lecture 10 Neural network interpretation (feature visualization, saliency maps, concepts)
- June 17 Lecture 11 To be announced

Paper presentations

- June 24 Lecture 12 Presentation of research papers (1 and 2)
- July 1 Lecture 13 Presentation of research papers (3 and 4)
- July 8 Lecture 14 Presentation of research papers (5 and 6)
- July 15 Lecture 15 Questions

Assignment and tutorial schedule

Assignment 0
- Publication: April 7
- Tutorial: April 14

Assignment 1
- Publication: April 14
- Tutorial: April 28

Assignment 2
- Publication: April 28
- Tutorial: May 12

Assignment 3
- Publication: May 12
- Tutorial: May 26

Assignment 4
- Publication: May 26
- Tutorial: June 16 (not June 9 due to public holiday)

Assignment 5
- Publication: June 9
- Tutorial: June 23

Assignment 6
- Publication: June 23
- Tutorial: July 7

Mini project schedule

- April 15 **Introduction of task**
- April 22 **Introduction of Graph Neural Networks**
- April 29 Work on task
- May 6 Work on task
- May 13 Work on task
- May 20 Work on task
- May 27 Work on task
- June 3 Work on task
- June 10 Work on task
- June 17 Work on task
- June 24 Work on task

(Anonymous) feedback on lecture and exercises

- Feel free to talk to me anytime something about the course bothers you
- You can do so in person, after the lecture, via email, …
- You can also do so anonymously in Panda

- I cannot see who submitted the feedback in Panda!
- I will do my best to take your feedback into account (for the remainder of the course / next year)

Question 1

- What do you like about the course? Please be specific and include examples.

Question 2

- What suggestions for improvement do you have for this course? Please be specific and include examples.

Presentation of research papers in lecture (1)

General

- Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” In *Nature machine intelligence*. 2019. https://www.nature.com/articles/s42256-019-0048-x.pdf
- Lapuschkin, Sebastian, et al. “Unmasking Clever Hans predictors and assessing what machines really learn.” In *Nature communications.* 2019. https://www.nature.com/articles/s41467-019-08987-4

Evaluation

- Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. “From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai.” In *ACM Computing Surveys.* 2023. https://dl.acm.org/doi/pdf/10.1145/3583558

Presentation of research papers in lecture (2)

Model-agnostic methods

- Sundararajan, Mukund, and Amir Najmi. “The many Shapley values for model explanation.” In *ICML*. 2020. https://proceedings.mlr.press/v119/sundararajan20b/sundararajan20b.pdf

Neural networks

- Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. “Concept bottleneck models.” In *ICML*. 2020. http://proceedings.mlr.press/v119/koh20a/koh20a.pdf

Conversational XAI

- Mindlin, Dimitry, Fabian Beer, Leonie Nora Sieger, Stefan Heindorf, Philipp Cimiano, Elena Esposito, and Axel-Cyrille Ngonga-Ngomo. “Beyond One-Shot Explanations: A Systematic Literature Review of Dialogue-Based XAI Approaches.” In *Artificial Intelligence Review*. 2024. https://link.springer.com/content/pdf/10.1007/s10462-024-11007-7.pdf

Presentation of research papers in lecture (3)

Applications

- Shu, Kai, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. “defend: Explainable fake news detection.” In *KDD.* 2019. https://dl.acm.org/doi/pdf/10.1145/3292500.3330935
- Lundberg, Scott M., Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston et al. “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.” In *Nature biomedical engineering.* 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6467492/pdf/nihms-1505578.pdf

Explainable AI for non-tabular data

- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D. and Du, M. Explainability for large language models: A survey. In *ACM Transactions on Intelligent Systems and Technology*. 2024. https://dl.acm.org/doi/10.1145/3639372
- Heindorf, Stefan, Lukas Blübaum, Nick Düsterhus, Till Werner, Varun Nandkumar Golani, Caglar Demir, and Axel-Cyrille Ngonga Ngomo. “Evolearner: Learning description logics with evolutionary algorithms.” In *WWW*. 2022. https://arxiv.org/abs/2111.04879
- Further topics are possible. Please come talk to me if you have other ideas