

Today

- Kick-Off Meeting
- Some Formalities
- Short Overview of the Topics
- We are very few
- →But this does not mean that we can't have a good seminar
- Organised by
 - Simon Klüttermann
(simon.kluettermann@cs.tu-dortmund.de)
 - Chiara Balestra
(chiara.balestra@cs.uni-dortmund.de)



Objective of this Seminar

- Introduction to some fundamental research problems
 - Researching current scientific ideas
 - Understanding benefits and drawbacks of state-of-the-art techniques
 - Writing a clear and concise scientific report
 - Presenting and discussing your findings

→Great start for a bachelor thesis.... →maybe just talk to your supervisor about this



Timeline

- 1 Presentation in Class (Last week of June)
 - 2 Discussion of your Findings (afterwards)
 - 3 Writing of your Report (till 15.07.2022 23:59)
- All parts required!

 - Everything will be done in english. If this is a problem for you, please write us.



Tasks of this Seminar

- 1 Choose a couple of topics from our list, you will be assigned to one of them
 - 2 Read and understand the chapter/paper given to you
 - 3 Find, read and understand related literature. It is probably impossible to get a good picture about your topic from just one paper (and chapter)
 - 4 Critically analyze the suggested ideas and compare them to the literature
- Final Results:
 - Presentation (25-30min +10min discussion)
 - Written Report (at least 6 Pages double column, ACM template)



Research Culture

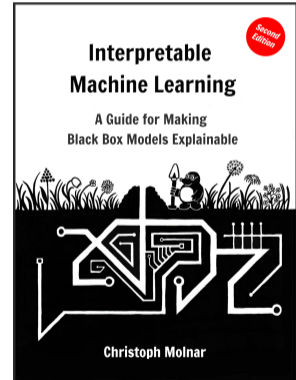
This course is Research oriented

- Feel free to ask as many Questions as you want
- If you want to discuss your Topic with somebody, make an appointment with your Supervisor
- the same holds for your Presentation/Report
- Any Feedback is always appreciated



Topics

- Based on the Book "Interpretable Machine Learning" by Christoph Molnar
- Freely available at christophm.github.io/interpretable-ml-book
- Some Topics contain programming assignments. We suggest using google colab for these.



Topic 1: Shapley Values

Chapter: 9.2,9.5 and 9.6 + 2/3 papers

Supervisor: Chiara Balestra (chiara.balestra@cs.uni-dortmund.de)

Either Computer Sciency

- **Shapley Values for Feature Selection: The Good, the Bad, and the Axioms** (Fryer et al. 2020)
- **Explaining Models by Propagating Shapley Values of Local Components** (Chen et al. 2020)
- **GraphSVX: Shapley Value Explanations for Graph Neural Networks** (Duval et al. 2021)

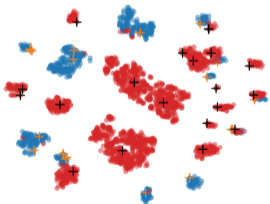
Or Medical Application

- **Identifying mortality factors from Machine Learning using Shapley values – a case of COVID19** (Smith et al. 2021)
- **Explaining multivariate molecular diagnostic tests via Shapley values** (Roder et al. 2021)



Topic 2: Prototypes

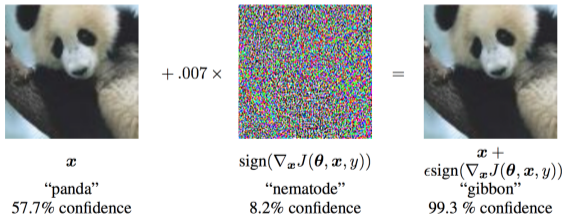
Interpreting Convolutional Sequence Model by Learning Local Prototypes with Adaptation Regularization (Ni, Chen, Cheng, et al., 2021) **Chapter: 8.7 Supervisor: Bin Li** (bin.li@cs.uni-dortmund.de)



- Represent some model output by well fitting data instances

Topic 3: Adversarial Examples

Practical Black-Box Attacks against Machine Learning (Papernot, McDaniel, et al., 2016)
Chapter: 10.4 Supervisor: Benedikt Böing (benedikt.boing@cs.uni-dortmund.de)



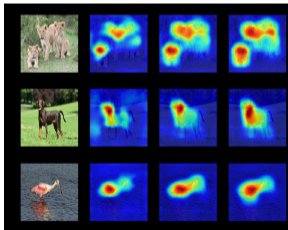
- Slight changes in a neural network can change its output drastically

Topic 4: Pixel Attribution

Efficient Saliency Maps for Explainable AI (Mundhenk, Chen, et al., 2019)

Chapter: 10.2 **Supervisor:** Simon Klüttermann

(simon.kluettermann@cs.uni-dortmund.de)



- Different parts of an image have different effect/importance on the classification of an image
- Programming task: Generate one Saliency Map yourself!



- 1: Shapley Values
- 2: Prototypes
- 3: Adversarial Examples
- 4: Pixel Attribution