



Proseminar Interpretable Machine Learning



Today

- Kick-Off Meeting
- Some Formalities
- Short Overview of the Topics

- Organised by
- Chiara Balestra
(chiara.balestra@cs.uni-dortmund.de)
- Simon Klüttermann
(simon.kluettermann@cs.tu-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 Kick-Off Meeting
 - 2 Choose Topics till 10.10.2022
 - 3 Talk to your supervisor once till 01.11.2022
 - 4 5min Presentations of your Topic on 01.11.2022
 - 5 Write your abstract till 15.11.2022
 - 6 Presentation in Class (end of January/beginning February)
 - 7 Discussion of your Findings (afterwards)
 - 8 Writing of your Report (till two weeks later)
- 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
- Presentation course (included):
 - 5min presentation of your topic
 - Write an abstract and get feedback
 - Final Results:
 - Presentation (25-30min +10min discussion)
 - Written Report (at least 6 Pages double column, ACM template)



Research Culture

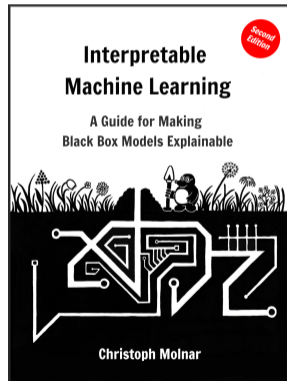
This course is Research oriented

- 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
- suggestion: Every two weeks!
- But at least once:
 - before your 5min Presentation,
 - before your Presentation,
 - before your Report to discuss the Presentation



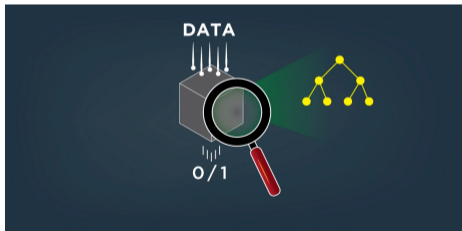
Topics

- 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: Why do we care about IML?

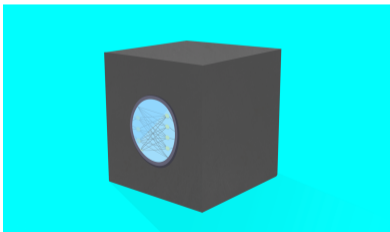
Chapter: 3.1-3.3 **Supervisor:** Daniel (daniel.wilmes@cs.uni-dortmund.de)



- e.g. Is IML required?

Topic 2: How to do IML research?

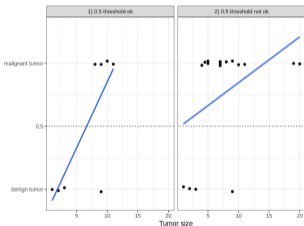
Chapter: 3.4-3.6 **Supervisor:** Daniel (daniel.wilmes@cs.uni-dortmund.de)



- e.g. How to evaluate Interpretability

Topic 3: Linear Models

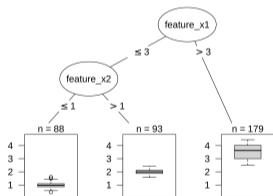
Chapter: 5.1-5.3 Supervisor: Jelle (jelle.huentelmann@cs.tu-dortmund.de)



- Simple Models are simple to explain
- Programming task: Do a linear regression on a simple dataset!

Topic 4: Decision Trees

Chapter: 5.4 Supervisor: Jelle (jelle.huentelmann@cs.tu-dortmund.de)

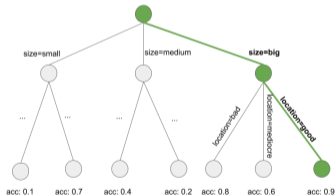


- Programming task: Train a decision Tree on a simple dataset!
- Could be combined with 5



Topic 5: Rule Based Methods

Chapter: 5.5-5.6 Supervisor: Jelle (jelle.huentelmann@cs.tu-dortmund.de)

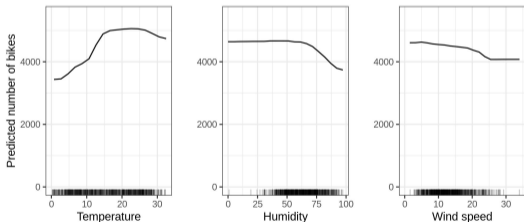


■ Could be combined with 4



Topic 6: Partial Dependence Plot

Chapter: 8.1 Supervisor: Carina (carina.newen@cs.uni-dortmund.de)

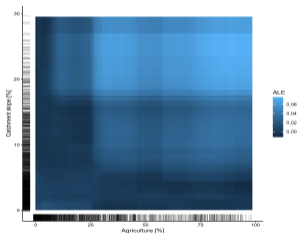


- How much does changing a feature change the output?
- Could be combined with 7



Topic 7: Accumulated Local Effects

Chapter: 8.2 Supervisor: Carina (carina.newen@cs.uni-dortmund.de)

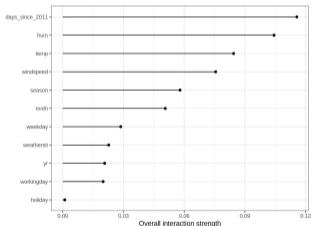


- How much effect does changing a feature have on the average prediction
- Could be combined with 6



Topic 8: Feature Interactions

Chapter: 8.3 Supervisor: Carina (carina.newen@cs.uni-dortmund.de)

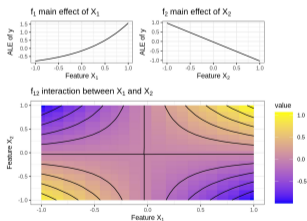


- In general features are not independent
- Measure the effect of interactions between them



Topic 9: Functional Decomposition

Chapter: 8.4 Supervisor: Daniel (daniel.wilmes@cs.uni-dortmund.de)

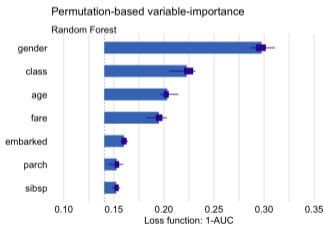


- Describe a function by feature interactions and their interactions



Topic 10: Permutation Feature Importance

Chapter: 8.5 Supervisor: Bin (bin.li@tu-dortmund.de)

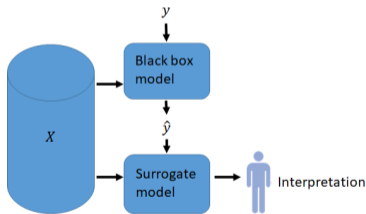


- How much does a feature change, if we permute its values



Topic 11: Global Surrogates

Chapter: 8.6 Supervisor: Bin (bin.li@tu-dortmund.de)

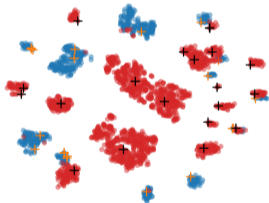


- Replace a complicated model by an interpretable one



Topic 12: Prototypes

Chapter: 8.7 Supervisor: Bin (bin.li@tu-dortmund.de)

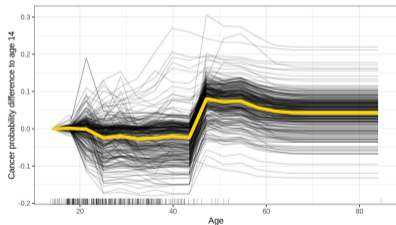


- Represent some model output by well fitting data instances



Topic 13: Individual Conditional Expectation

Chapter: 9.1-9.2 **Supervisor:** Chiara (chiara.balestra@cs.uni-dortmund.de)

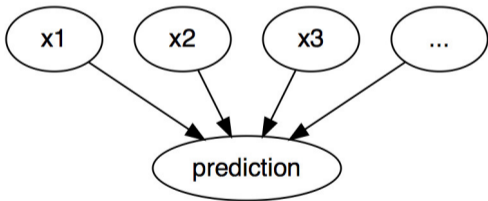


- Show the effect one feature has on the prediction



Topic 14: Counterfactual Explanations

Chapter: 9.3-9.4 **Supervisor:** Chiara (chiara.balestra@cs.uni-dortmund.de)

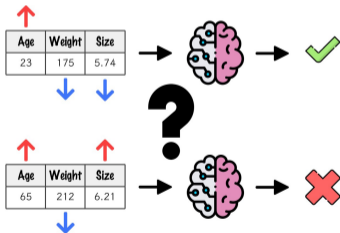


- What to do to change a prediction?



Topic 15: Shapley Values

Chapter: 9.5-9.6 **Supervisor:** Chiara (chiara.balestra@cs.uni-dortmund.de)

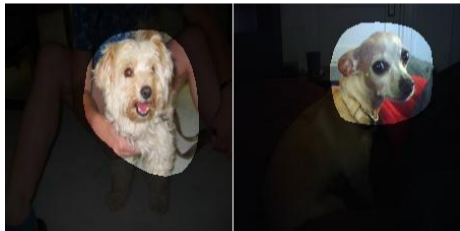


- Use game theory to explain the output of a model



Topic 16: Learned Features

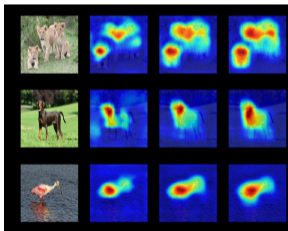
Chapter: 10.1 **Supervisor:** Benedikt (benedikt.boeing@cs.tu-dortmund.de)



- Conv. NN contain Intepretable Features
- Programming task: Visualise your own classifier!

Topic 17: Saliency Maps

Chapter: 10.2 **Supervisor:** Simon (simon.kluettermann@cs.uni-dortmund.de)

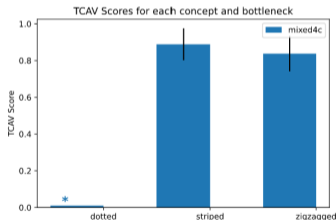


- Different parts of an image have different effect/importance on the classification of an image
- Programming task: Generate one Saliency Map yourself!



Topic 18: Concept Detection

Chapter: 10.3 Supervisor: Simon (simon.kluettermann@cs.uni-dortmund.de)

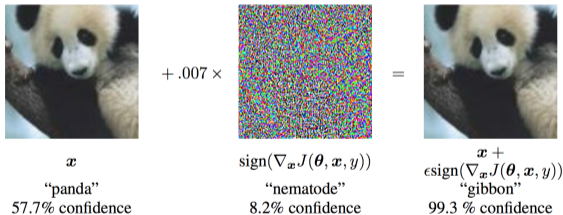


- Replace Features by Concepts



Topic 19: Adversarials

Chapter: 10.4 Supervisor: Benedikt (benedikt.boeing@cs.tu-dortmund.de)

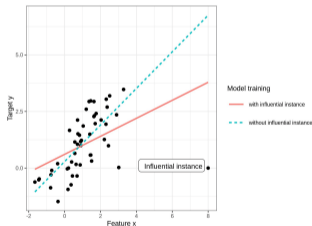


- Slight changes in a neural network can change its output drastically



Topic 20: Influential Instances

Chapter: 10.5 Supervisor: Simon (simon.kluettermann@cs.uni-dortmund.de)



- Single examples can change the output of a NN drastically

1: Why IML	2: How IML	3: Linear ML	4: Trees
5: Rules	6: PDP	7: ALE	8: Interactions
9: Decomposition	10: Permutations	11: Surrogates	12: Prototypes
13: ICE	14: Counterfactuals	15: Shapley Values	16: Learned Features
17: Saliency Maps	18: Concepts	19: Adversarials	20: Influential Instances