Dienstag, 16. November 2021

Exercise 03.02: Reading/Discussion/Summary

Part 2: Think about the following questions:

- What is Explainable Machine Learning?
    - Explainable ML are ML techniques, that have a black box for the evaluation of the Problem. „Explanations“ are done post hoc.
- What are the features of Interpretable Machine Learning?
    - The features of Interpretable ML are that they inherently Interpretable, meaning you can exactly follow their decisions.
- Why and when do we need both?
    - We need both, if we want to find the most accurate Model, wich can be either.
- Are there any disadvantages?
    - They are often computationally harder to solve, take more time to develop and need more resources like money, experts and so on.
- What are specific challenges in their application?
    - strongly domain specific
    - constructing optimal logical models
    - construct optimal sparse scoring systems
    - define interpretability for specific domains and create methods accordingly, including computer vision
- Can, and if so, how, explainable and interpretable ML be implemented efficiently?
    - Explainable ML are easier to implement than interpretable ML. Interpretable ML can be easy to, but it is domain specific.
- What are some exemplary techniques to apply?
    - CORELS algorithm
    - RiskSLIM
    - append a special prototype layer to the end of the network