Karlsruher Institut für Technologie

Institute for Program Structures and Data Organization (IPD)
Information Management Systems
Prof. Dr.-Ing. Klemens Böhm

Explainability and Data Privacy – Alleviating the Antagonism

Project Group "Praxis der Forschung"
Summer Term 2022

1 Topic

While Machine Learning (ML) has become ubiquitous, models also tend to be more and more intricate. In many settings however, it is necessary to understand why a certain automated decision has been taken. For instance, you might want to know why you did not get admitted to a certain university, or why you just got to see a specific ad. So-called explainers, which one can apply to a certain model and a decision it has generated, have been touted to help with this. However, explanations may contain sensitive information. Think of the explanation "You did not get the loan, because the person Bernhard Beckert who earns more than you did not get it either.". This antagonism between explainability and data privacy not only exists in theory, we have also observed it in systematic experiments with real data with various combinations of conventional ML approaches and off-the-shelf explainers.

2 Assignment

Your assignment is to design explainers that are better than the state of the art in this respect and demonstrate their superiority. The experiments just mentioned provide indications where you can start with the design of these explainers.

3 Contact

- Clemens Müssener <clemens.muessener@partner.kit.edu>
- Prof. Dr.-Ing. Klemens Böhm <klemens.boehm@kit.edu>