UMEÅ UNIVERSITET

Context-Based Explanations for Machine Learning Predictions

Sule Anjomshoae

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Hörsal NAT.D.320, Naturvetarhuset,
den 26:e september, kl. 08:30.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor Maria Riveiro,
Department of Computer Science, Jönköping University, Jönköping, Sweden.

Department of Computing Science

Organization
Umeå University
Dept. of Computing Science

Document type
Doctoral thesis

Date of publication
5 September 2022

Author
Sule Anjomshoae

Title
Context-Based Explanations for Machine Learning Predictions

Abstract
In recent years, growing concern regarding trust in algorithmic decision-making has drawn attention to more transparent and interpretable models. Laws and regulations are moving towards requiring this functionality from information systems to prevent unintended side effects. Such as the European Union's General Data Protection Regulations (GDPR) set out the right to be informed regarding machine-generated decisions. Individuals affected by these decisions can question, confront and challenge the inferences automatically produced by machine learning models. Consequently, such matters necessitate AI systems to be transparent and explainable for various practical applications.

Furthermore, explanations help evaluate these systems' strengths and limitations, thereby fostering trustworthiness. As important as it is, existing studies mainly focus on creating mathematically interpretable models or explaining black-box algorithms with intrinsically interpretable surrogate models. In general, these explanations are intended for technical users to evaluate the correctness of a model and are often hard to interpret by general users.

Given a critical need for methods that consider end-user requirements, this thesis focuses on generating intelligible explanations for predictions made by machine learning algorithms. As a starting point, we present the outcome of a systematic literature review of the existing research on generating and communicating explanations in goal-driven eXplainable AI (XAI), such as agents and robots. These are known for their ability to communicate their decisions in human understandable terms. Influenced by that, we discuss the design and evaluation of our proposed explanation methods for black-box algorithms in different machine learning applications, including image recognition, scene classification, and disease prediction.

Taken together, the methods and tools presented in this thesis could be used to explain machine learning predictions or as a baseline to compare to other explanation techniques, enabling interpretation indicators for experts and non-technical users. The findings would also be of interest to domains using machine learning models for high-stake decision-making to investigate the practical utility of proposed explanation methods.

Keywords
Explainable AI, explainability, interpretability, black-box models, deep learning, neural networks, contextual importance

Language
English

ISBN
print: 978-91-7855-859-9
PDF: 978-91-7855-860-5

ISSN
0348-0542

Number of pages
48 + 5 papers