UNIVERSITÄT
BIELEFELD
Faculty of Technology

Bachelor or Master Thesis

Constrained Counterfactual Explanations

Semantic Computing Group
Christoph Düsing
cduesing@techfak.uni-bielefeld.de

Today, we wittiness the prevalence of machine learning and deep learning in a vast variety of everyday tasks as well as highly specialized domains. Due to their black-box nature, most such models are inherently intransparent, hence hindering humans to make sense of the underlying decision-making. Recent efforts toward increasing the interpretability of these systems exploit explainable artificial intelligence (XAI) techniques in order to increase their reliability and trustworthiness. Among the wide range of XAI approaches, counterfactual explanations seem particularly suitable for certain domains such as healthcare, due to their more human-like approach of explaining the model outcome. To explain a certain prediction, counterfactual explanations provide a hypothetical, yet similar counterexample that would result in a different model prediction.

Numerous works address the application of counterfactual explanations in different domains, speed up the process of explanation generation, or improve the quality of explanations. However, they have in common that they provide only a single layer of explanation, i.e., one or many counterfactual explanations for the prediction to be explained. In certain domains, e.g., in healthcare, providing additional layers of explanations could further increase the trustworthiness but also perceived usefulness of counterfactual explanations, as they resemble the familiar concept of differential diagnoses.

In this thesis, the use of constrained counterfactual explanations as additional layer of explanations should be evaluated. This could either be empirically by refining existing approaches which are tested using automated metrics and human judgement. Alternatively, a taxonomy could be developed relating different layers of explanations to known concepts from computer science and cognitive science.

Related literature

1. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

The Semantic Computing Group researches and develops methods that enable machines to acquire relevant knowledge as well as linguistic capabilities. Using methods from *natural language understanding* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading unstructured textual data. In particular, the group focuses on methods for information extraction, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at: http://sc.cit-ec.uni-bielefeld.de.

Interested? @mail to `cduesing@techfak.uni-bielefeld.de`