MLKD 2024
The First International Conference on
Machine Learning and Knowledge Discovery
Amirkabir University of Technology, December 18-19, 2024

ICSE: Interpretable Counterfactual/SemiFactual Explanations

Sooroush Riazi* Mohammad Akbari† Zahed Rahmati‡

Abstract

The lack of explainability and interpretability is one of the biggest barriers in real world integration of machine learning. Many researchers tried to improve the interpretability of black box models by post hoc explanations, which try to explain the model by methods like model simplifications, local explanations, explanations by example and many more. Recent user studies showed that Grad-CAM and Lime were less understandable than simple nearest neighbors from the training set. Counterfactual explanations, which are one of the example based explanations have emerged as one of the main methods that can unravel the causal relationship learned in the black box models. to tackle these challenges we propose a novel method to create counterfactual explanations with desired probability in desired class which makes this method more user friendly. In particular, we delve into the concept of semi-factual explanations and define near-bound counterfactuals as points with two dominant class probabilities, which makes them closer to the decision boundary. We used Variational AutoEncoders (VAEs) to create latent space and utilize this latent space to find the minimum semantic (not adversarial) change that can change the prediction of instance to any probability in the desired class. However one of the main challenges in creating counterfactuals is the trade-off between the amount of change applied on the instance and the plausibility and interpretability of generated counterfactuals, we conducted experiments on two datasets demonstrating the effectiveness of our approach

Keywords: Explainable Ai, Counterfactual, Interpretablity

*Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran sooroushr@aut.ac.ir
†Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran akbari.ma@ut.ac.ir
‡Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran zrahmati@aut.ac.ir