Explanation Dropout: Practical Counterfactual Explanations for Machine Learning Models

Abstract

A significant part of responsible use of machine learning models, particularly in financial decisioning, is the ability to produce justifiable explanations for the primary reasons for a score, commonly interpreted as quantifying and ranking the input features which drive the score the strongest in this particular instance.

Two approaches dominate: restrict the design and architecture of the model to have a particularly clear and intrinsic explanatory algorithm associated, or, a model-agnostic algorithm which queries the model beyond the baseline score to yield explanatory statistics. This work focuses on a novel concept for the second task which offers advantages over common existing methods, which have some drawbacks.

Our approach does not use local perturbations or gradients, which may yield limited or less intuitively relevant explanations, and furthermore it does not ask the ML model to score inputs in conditions which have never been seen during the training process. This second problem is underappreciated, as complex ML models may behave in uncontrolled ways when queried “out of distribution” or “out of data manifold” as might be required in some other explanation techniques. Our Explanation Dropout method includes the explanatory conditions during model training so there are never “out of distribution” circumstances encountered during scoring. There is no restriction on model architecture, and run-time explanation computations do not require any Monte Carlo randomization or large databases of external data.

Explanation Dropout offers an intuitively attractive counterfactual notion of explainability driven by answering the hypothetical “How would an almost identical model perform if trained without using certain inputs?” in a computationally concrete way.