Robust Counterfactual Explanations for Tree-Based Ensembles

Speaker: Saumitra Mishra, J.P. Morgan XAI Centre of Excellence

Abstract:

Counterfactual explanations inform ways to achieve a desired outcome from a machine learning model. However, such explanations are not robust to certain real-world changes in the underlying model (e.g., retraining the model, changing hyperparameters, etc.), questioning their reliability in several applications, e.g., credit lending. The talk will introduce our recently proposed strategy – that we call RobX – to generate robust counterfactuals for tree-based ensembles, e.g., XGBoost. Tree-based ensembles pose additional challenges in robust counterfactual generation, e.g., they have a non-smooth and non-differentiable objective function, and they can change a lot in the parameter space under retraining on very similar data. The talk will first introduce a novel metric – that we call Counterfactual Stability – that attempts to quantify how robust a counterfactual is going to be to model changes under retraining, and comes with desirable theoretical properties. The proposed strategy RobX works with any counterfactual generation method (base method) and searches for robust counterfactuals by iteratively refining the counterfactual generated by the base method using the Counterfactual Stability metric. The talk will end with a discussion of some empirical results that demonstrate that the proposed strategy generates counterfactuals that are significantly more robust (nearly 100% validity after actual model changes) and also realistic (in terms of local outlier factor) over existing state-of-the-art methods.

Bio:

Saumitra is Vice President/AI Research Lead at J.P. Morgan where he is associated with the XAI Center of Excellence within the AI research team. Saumitra completed his PhD in Electronics Engineering from Queen Mary University of London, UK in 2020. Later, he was a research associate at the Alan Turing Institute, London. Prior to PhD, Saumitra was a technical manager at Samsung Research India, Bangalore developing novel technologies for consumer electronics products. Saumitra is broadly interested in research on explainable AI, fairness and robustness of machine learning models.