Inference for Interpretable Machine Learning: Feature Importance and Beyond

Genevera I. Allen

Department of Statistics, Columbia University, USA

ABSTRACT

Machine Learning (ML) systems are being used to make critical societal, scientific, and business decisions. To promote trust, transparency, and accountability in these systems, many advocate making them interpretable or explainable. In response, there has been dramatic growth in techniques to provide human understandable interpretations of black-box techniques. Yet we ask: Can we trust these ML interpretations? How do we know if they are correct? Unlike for prediction tasks, it is difficult to directly test the veracity of ML interpretations. In this talk, we focus on interpreting predictive models to understand important features and important feature patterns. We first present motivating results from a large-scale empirical stability study illustrating that feature interpretations are generally unreliable and far less reliable than predictions. Motivated by these issues, we present a new statistical inference framework for quantifying the uncertainty in feature importance and higher-order feature patterns. Based upon the Leave-One-Covariate-Out (LOCO) framework, we develop a computational and inferential approach that does not require data splitting or model refitting by utilizing minipatch ensembles, or ensembles generated by double random subsampling of observations and features. Even though our framework uses the same data for training and inference, we prove the asymptotic validity of our confidence intervals for LOCO feature importance under mild assumptions. Finally, we extend our approach to detect and test feature interactions via the iLOCO metric. Our approach allows one to test whether a feature significantly contributes to any ML model's predictive ability in a completely distribution free manner, thus promoting trust in ML feature interpretations. We highlight our inference for interpretable ML approaches via real scientific case studies and a fun illustrative example. This is joint work with Lili Zheng, Luqin Gan, Camille Little, Tarek Zikry, and Mariah Loehr.

Keywords: Conformal Inference, Selective Inference, Feature Importance Inference, Feature Interaction, Interpretable Machine Learning, Ensemble Learning