A CONSTRUCTIVE APPROACH TO EXPLAINABLE NEURAL NETWORKS

Aijun Zhang

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong

Email: ajzhang@mit.edu  Website: http://www.statsoft.org/

Prediction accuracy and model explainability are the two most important objectives when we develop machine learning algorithms to solve the real world problems. In this paper, we propose a constructive approach to the explainable neural network (xNN) by integrating additive, orthogonal, sparse and smooth constraints. Such xNN is built upon the classical projection pursuit regression (PPR) or additive index model (AIM), together with the following interpretable considerations. The orthogonal and sparse constraints on projection indexes make the model more identifiable, while the smooth constraint on the ridge functions prevents unnatural wiggles. For model estimation, we develop a modified backpropagation algorithm based on stochastic gradient descent and Cayley transformation. Both simulation and real data studies show that the new xNN model is an efficient and promising tool for interpretable machine learning.