NN2Poly: a polynomial representation for deep feed-forward artificial neural networks

P. Morala Miguélez; J. Cifuentes Quintero; R.E. Lillo Rodríguez; I. Ucar

Abstract-

Interpretability of neural networks (NNs) and their underlying theoretical behavior remain an open field of study even after the great success of their practical applications, particularly with the emergence of deep learning. In this work, NN2Poly is proposed: a theoretical approach to obtain an explicit polynomial model that provides an accurate representation of an already trained fully connected feed-forward artificial NN a multilayer perceptron (MLP). This approach extends a previous idea proposed in the literature, which was limited to single hidden layer networks, to work with arbitrarily deep MLPs in both regression and classification tasks. NN2Poly uses a Taylor expansion on the activation function, at each layer, and then applies several combinatorial properties to calculate the coefficients of the desired polynomials. Discussion is presented on the main computational challenges of this method, and the way to overcome them by imposing certain constraints during the training phase. Finally, simulation experiments as well as applications to real tabular datasets are presented to demonstrate the effectiveness of the proposed method.

Index Terms- Interpretability, machine learning, multilayer perceptron (MLP), multiset partitions, neural networks (NNs), polynomial representation.

Citation:
Morala Miguélez, P.; Cifuentes, J.; Lillo, R.E.; Ucar, I. "NN2Poly: a polynomial representation for deep feed-forward artificial neural networks", IEEE Transactions on Neural Networks and Learning Systems, , .