A Mathematical Certification for Positivity Conditions in Neural Networks With Applications to Partial Monotonicity and Trustworthy AI

A. Polo Molina; D. Alfaya Sánchez; J. Portela González

Abstract-

Artificial neural networks (ANNs) have become a powerful tool for modeling complex relationships in large-scale datasets. However, their closed box nature poses trustworthiness challenges. In certain situations, ensuring trust in predictions might require following specific partial monotonicity constraints. However, certifying if an already-trained ANN is partially monotonic is challenging. Therefore, ANNs are often disregarded in some critical applications, such as credit scoring, where partial monotonicity is required. To address this challenge, this article presents a novel algorithm (LipVor) that certifies if a closed box model, such as an ANN, is positive based on a finite number of evaluations. Consequently, since partial monotonicity can be expressed as a positivity condition on partial derivatives, LipVor can certify whether an ANN is partially monotonic. To do so, for every positively evaluated point, the Lipschitzianity of the closed box model is used to construct a specific neighborhood, where the function remains positive. Next, based on the Voronoi diagram of the evaluated points, a sufficient condition is stated to certify if the function is positive in the domain. Unlike prior methods, our approach certifies partial monotonicity without constrained architectures or piecewise linear activations. Therefore, LipVor could open up the possibility of using unconstrained ANN in some critical fields. Moreover, some other properties of an ANN, such as convexity, can be posed as positivity conditions, and therefore, LipVor could also be applied.

Index Terms- Artificial neural networks (ANNs), mathematical certification, partial monotonicity, trustworthy AI

Citation:

Polo-Molina, A.; Alfaya, D.; Portela, J. "A Mathematical Certification for Positivity Conditions in Neural Networks With Applications to Partial Monotonicity and Trustworthy AI", IEEE Transactions on Neural Networks and Learning Systems, , .