XVII Congreso Galego de Estatística e Investigación de Operacións
Ourense, 23 - 25 de outubro de 2025

Combining Neural Networks and Generalized Additive Models for Explainable AI

Inés Ortega-Fernández[1], Marta Sestelo[2,3] and Nora M. Villanueva[2,3]

[1] Galician Research and Development Center in Advanced Telecommunications (Gradiant), 36214 Vigo, Spain
[2] Galician Center for Mathematical Research and Technology (CITMAga), 15782 Santiago de Compostela, Spain
[3] Universidade de Vigo, Department of Statistics and O.R. & SiDOR Group, 36310 Vigo, Spain

ABSTRACT

Neural Networks are currently considered one of the most powerful tools for a variety of tasks, including anomaly detection, computer-aided disease diagnosis, and natural language processing. However, a major drawback of these models is their lack of interpretability, often referred to as the "black-box" problem, which makes it difficult to understand how predictions are made. To address this issue, a Neural Network architecture inspired by Generalized Additive Models (GAMs) was introduced. This approach involves fitting an independent Neural Network to estimate the contribution of each feature to the output variable, resulting in a highly accurate and interpretable Deep Learning model. The method provides a flexible framework for building Generalized Additive Neural Networks without imposing constraints on the network architecture. The proposed algorithm is applied to a real-world use case. Additionally, in order to make this methodology accessible to the research community, an R package is being developed, `neuralGAM`, providing robust, well-documented, and reproducible tools to facilitate its application and further exploration.

Keywords: Interpretable deep learning; Generalized additive models; Explainable Artificial Intelligence; neural network

REFERENCES

Hastie, T., Tibshirani, R. (1990). Generalized Additive Models, 1931 (11), 683–741. Chapman and Hall, London.

Ortega-Fernández, I., Sestelo, M. and Villanueva, N. M. (2024). Explainable generalized additive neural networks with independent neural network training. Statistics and Computing, 34(1):6, 1–12.

Ortega-Fernández, I. and Sestelo, M. (2025). neuralGAM: Interpretable Neural Network Based on Generalized Additive Models. v1.1.1. URL https://cran.r-project.org/web/packages/neuralGAM/.

R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.