Title: Statistics for a SAFE AI

Author(s): Paolo, Giudici

Affiliation(s): Statistical laboratory, University of Pavia

Abstract:

The current widespread use of AI motivates the need to develop advanced statistical methods that can measure its "trustworthiness", in line with the Artificial Intelligence Act recently proposed by the European Commission (European Commission, 2021).

To measure trustworthiness of AI, we propose statistical metrics that consist of a set of four integrated statistical measures of trustworthiness, all based on the extension of the Lorenz Curve (Lorenz, 1905): from the measurement of income concentration to the measurement of the concentration of machine learning predictions. The four statistical metrics can be summarised with the acronym S.A.F.E., which derives from the four considered variables: Sustainability, which refers to the resilience of the AI outputs under anomalous extreme events and/or cyber attacks; Accuracy, which refers to the predictive accuracy of the model outputs; Fairness, which refers to the absence of biases towards population groups, induced by the AI output; Explainability, which refers to the capability of the model output to be understood and oversight by humans, particularly in its driving causes. While the former two requirements are more technical, and "internal" to the AI process, the latter two are more ethical, and "external" to the AI process, involving the stakeholders of an AI system.

We remark that tbs proposed metrics consist of "agnostic" statistical tools, able to post-process the predictive output of a machine learning model in a general way, independently on the underlying data structure and statistical model.

Keywords: machine learning (1); Lorenz zonoids (2); explainable AI (3).

Author Profiles (s): https://scholar.google.com/citations?user=ogeVB1kAAAAJ&hl=en