Contribution ID: **9** Type: **not specified**

Global Importance Measures for Machine Learning Model Interpretability, an Overview

Tuesday, 12 September 2023 17:50 (20 minutes)

Machine learning (ML) algorithms, fitted on learning datasets, are often considered as black-box models, linking features (called inputs) to variables of interest (called outputs). Indeed, they provide predictions which turn out to be difficult to explain or interpret. To circumvent this issue, importance measures (also called sensitivity indices) are computed to provide a better interpretability of ML models, via the quantification of the influence of each input on the output predictions. These importance measures also provide diagnostics regarding the correct behavior of the ML model (by comparing them to importance measures directly evaluated on the data) and about the underlying complexity of the ML model. This communication provides a practical synthesis on post-hoc global importance measures that allow to interpret the model generic global behavior for any kind of ML model. A particular attention is paid to the constraints that are inherent to the training data and the considered ML model: linear vs. nonlinear phenomenon of interest, input dimension and strength of the statistical dependencies between inputs.

Keywords

Sensitivity analysis, Shapley, Sobol' indices, Relative weight analysis

Classification

Mainly methodology

Primary authors: IOOSS, Bertrand (EDF R&D); CHABRIDON, Vincent (EDF R&D); Dr PELAMATTI, Vincent (EDF R&D)

Presenter: IOOSS, Bertrand (EDF R&D)

Session Classification: CONTRIBUTED Machine Learning 3

Track Classification: Machine learning