EGU2020-21281
https://doi.org/10.5194/egusphere-egu2020-21281
EGU General Assembly 2020

Uncertainty Quantification and Explainable Artificial Intelligence

Maria Moreno de Castro
DKRZ German Climate Computer Center, Germany

The presence of automated decision making continuously increases in today's society. Algorithms based in machine and deep learning decide how much we pay for insurance, translate our thoughts to speech, and shape our consumption of goods (via e-marketing) and knowledge (via search engines). Machine and deep learning models are ubiquitous in science too, in particular, many promising examples are being developed to prove their feasibility for earth sciences applications, like finding temporal trends or spatial patterns in data or improving parameterization schemes for climate simulations.

However, most machine and deep learning applications aim to optimise performance metrics (for instance, accuracy, which stands for the times the model prediction was right), which are rarely good indicators of trust (i.e., why these predictions were right?). In fact, with the increase of data volume and model complexity, machine learning and deep learning predictions can be very accurate but also prone to rely on spurious correlations, encode and magnify bias, and draw conclusions that do not incorporate the underlying dynamics governing the system. Because of that, the uncertainty of the predictions and our confidence in the model are difficult to estimate and the relation between inputs and outputs becomes hard to interpret.

Since it is challenging to shift a community from “black” to “glass” boxes, it is more useful to implement Explainable Artificial Intelligence (XAI) techniques right at the beginning of the machine learning and deep learning adoption rather than trying to fix fundamental problems later. The good news is that most of the popular XAI techniques basically are sensitivity analyses because they consist of a systematic perturbation of some model components in order to observe how it affects the model predictions. The techniques comprise random sampling, Monte-Carlo simulations, and ensemble runs, which are common methods in geosciences. Moreover, many XAI techniques are reusable because they are model-agnostic and must be applied after the model has been fitted. In addition, interpretability provides robust arguments when communicating machine and deep learning predictions to scientists and decision-makers.

In order to assist not only the practitioners but also the end-users in the evaluation of machine and deep learning results, we will explain the intuition behind some popular techniques of XAI and aleatory and epistemic Uncertainty Quantification: (1) the Permutation Importance and Gaussian processes on the inputs (i.e., the perturbation of the model inputs), (2) the Monte-Carlo Dropout, Deep ensembles, Quantile Regression, and Gaussian processes on the weights (i.e, the perturbation of the model architecture), (3) the Conformal Predictors (useful to estimate the

confidence interval on the outputs), and (4) the Layerwise Relevance Propagation (LRP), Shapley values, and Local Interpretable Model-Agnostic Explanations (LIME) (designed to visualize how each feature in the data affected a particular prediction). We will also introduce some best-practises, like the detection of anomalies in the training data before the training, the implementation of fallbacks when the prediction is not reliable, and physics-guided learning by including constraints in the loss function to avoid physical inconsistencies, like the violation of conservation laws.