Contribution ID: **77** Type: **not specified**

Selection of a validation basis from a given dataset for supervised statistical learning

With the development of automatic diagnostics based on statistical predictive models, coming from any supervised machine learning (ML) algorithms, new issues about model validation have been raised. For example in the non-destructive testing field, generalized automated inspection (that will allow large gain in terms of efficiency and economy) has to provide high guarantees in terms of performance. In this case, it is necessary to be able to select a validation data basis that will not be used for the training nor the test of the ML model. This validation data basis has not to be communicated to the ML builder (which is often a supplier of the main company) because it will serve to realize an independent evaluation of the provided ML model. In this communication, we address the important question about the way to select a "good"validation basis from the dataset that is used for the problem. In the cases of small-size dataset or unbalanced dataset, simple selection algorithms (as a random choice) are insufficient to ensure the representativity of the validation basis, and a supervised selection based on statistical criteria is necessary. In large dataset cases, an appropriate selection will lead to more robust ML model predictive capabilities evaluations than a random choice. We adopt a "design of experiments"point of view, which seems mathematically natural. The particularity of our validation basis selection problem is that the dataset already exists (so the problem turns to selecting a certain number of points in a finite collection of points) and the output is also known. Several methods, adapted from the literature of computer experiments, are studied and tested in a machine laerning classification purpose. We show that the approach based on the "support points"concept is particularly relevant. An industrial test case from the company EDF illustrates the practical interest of the methodology.

Primary author: Dr IOOSS, Bertrand (EDF R&D)

Presenter: Dr IOOSS, Bertrand (EDF R&D)

Track Classification: Data Science in Process Industries