An overview of the role of Machine Learning in hydraulic and hydrological modeling

J. P. Carbajal
Swiss Federal Institute of Aquatic Science and Technology, Eawag, Dübendorf, Switzerland
V. Bellos
National Technical University of Athens, Greece

ABSTRACT

We provide an overview of Machine learning (ML) and its role in hydrology and hydraulic. The aim is to ease the access of researchers in the latter fields to the techniques in ML.

1 INTRODUCTION

Machine Learning (ML) has received increasing attention in the later years. It promises to ease the problem of modeling from observations. It is a heavily mathematized field, with strong statistical jargon. Therefore, it can be difficult to access for researchers from the fields of hydrology and hydraulics (hydro-research). A bird's-eye view of ML and its role, can be beneficial for the hydro-research community. Herein we address the question of what is ML in general terms and what are its role in hydro-research. We discuss some important features in each of these roles

Machine learning at a glance ML could be described as "the use of a set of observations to uncover an underlying process". The process is usually stated as a mathematical relation between the observations. Herein we consider only the case in which observations include the inputs and outputs of the process, i.e. supervised learning. The sought functional pattern, called *unknown target function* establishes the relation between the inputs and the outputs, i.e. is a model of the process. The *training examples*, are input-output samples from the underlying process and they represent all the direct information we have about it. Fig. 1 illustrates the structure of ML in this case [1].

To search for the target function we choose an extensive set of functions, which we call the *hypothesis set*. For example, we could choose all the linear functions between inputs and outputs (i.e. linear models), or all the functions generated by a given neural network. The key point is that the hypothesis set is built so as to contain the unknown

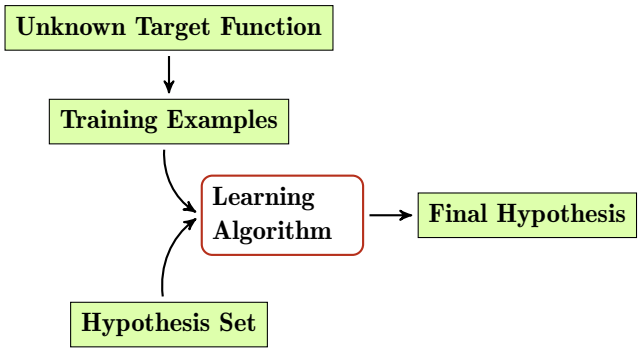

Figure 1: Bird's eye view of ML. Taken from "Learning from data" by Yaser S. Abu-Mostafa et al.

target function or at least a very good approximation for it. This choice is based on our previous experience and the available expert knowledge about the underlying process, e.g. mathematical or phenomenological models. The *learning algorithm* uses the training examples to select the best candidate function from the hypothesis set, i.e. the *final hypothesis*. Since all prior information about the unknown target function is encoded in the hypothesis set, the quality of the best candidate heavily depends on the elements in the set.

All fundamental research in supervised ML consist in developing new learning algorithms (mainly optimization algorithms), novel or concise descriptions of different hypothesis sets, and useful representations for input-output data (encodings).

2 ML IN HYDRO-RESEARCH

ML as described before can be useful for hydro-research in at least three situations: i) (artificial science) learning new models from measured data; ii) (scientific numerical modeling) using data to find the value of parameters of known models; iii) (emulation) replacing a model with a simpler version while maintaining the quality of the predictions.

The relation between these situations is summarized in Fig 2.

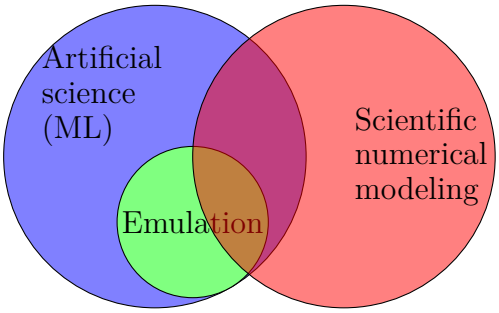

Figure 2: Three uses of ML in hydro-research. Emulation is a subset of pure ML (artificial science), and it intersect with numerical modeling when prior knowledge is exploited.

Artificial science This is the realm of pure ML in which a relation is learned using only data. We use ML to discover new natural laws based only on experimental data. Many popular ML techniques, specially artificial neural networks, have been seen as the grail to perform artificial science, but none have been up to the task. There is a long standing discussion about the topic but the main issue is that these tools fail to create abstractions the way human researchers do. [2, 3] The quest is a valid scientific endeavor, but the question itself still eludes a scientific formulation and scientist should be aware that we are far from obtaining such technology. Nevertheless, it was very popular in hydro-research during the last two decades. Besides the mentioned issue, three other difficulties can be identified: (i) there is a lack of measured data in the majorities of case studies; (ii) there are significant uncertainties in measuring variables related to hydro-research; (iii) even in gauged case studies, observed data exist usually in regular conditions, because either extreme conditions are rare or the extreme conditions have as consequence the failure of the measurement system. Therefore, learning scientific models under these circumstances is at least problematic.

Scientific numerical modeling When learning models from data all scientist will come with their own bag of beliefs. In this case we move from artificial science to scientific modeling, in which our beliefs are informed by data. We discard models and hypothesis based on their ability to predict observations.

Scientific numerical modeling is the current dominant use of ML in hydro-research applications. Modern numerical models can resolve fine spatial and temporal scales, mainly thanks to the exponential increase of computational power. Concomitantly, although open challenges still exist, efforts to improve our fundamental understanding of hydro-sciences are decreasing, e.g. improving theories, collecting data, validating models, etc.

The most common use of ML is for model calibration (a.k.a. system identification), in which model's parameters are determined using measured data. These parameters often enjoy a physical interpretation and this forms the basis for selecting or rejecting models. Although ML offers a large variety of optimization and calibration tools, these are not frequently used in applied hydro-research, even in the cases in which observed data exist. This might be partly due to the fact that the level of detail on fine-scaled models impose large runtime for each model evaluation, and optimization based on sampling becomes unfeasible.

Emulation In this component, instead of dealing with measured data, we use a simulator or model to sample input-output examples. With the data sampled from a known model we learn a new model which is numerically simpler than the original one.

Emulation is a subfield of ML and the intersection with scientific numerical modeling is given by emulators that use prior scientific knowledge. Not all methods of ML permit easy inclusion of this sort of prior knowledge, the best know to the authors are kernel methods (which includes many known ML methods).

Emulation is related to model order reduction (MOR) [4] , both produce fast surrogate models. We distinguish the two based on the dimensionality of the output of the surrogate. A MOR surrogate will generally provide as many states as the original simulator. Emulators on the other hand provide outputs of smaller dimension, [5] e.g. water depth at certain locations. This makes emulators less general than reduced models, but this specification of the former generally renders them faster than the latter. In practice the selection between reduced models or emulators is dictated by the application at hand and its engineering constraints (memory, processing power, accuracy, time budget, etc.)

The large runtime needed to sample fine-scaled models was already mentioned in the previous component. Sampling for emulation needs to generate good enough emulators with sparse datasets, here is were a reduced output dimension and prior knowledge play a crucial role. Once an emulator is learned, taking new samples becomes very cheap and optimizations technique for model calibration can be applied exhaustively, allowing not only model calibration but also real-time control, optimal design, uncertainty quantification, etc.

Emulation merges artificial science and scientific numerical modeling, with it we are able to discover hidden input-output relations while exploiting all available information about the phenomena.

Acknowledgements We thank Dr. Jörg Rieckermann for his continuous support. This research has received funding from Eawag's discretionary funding program (project EmuMore

kakila.bitbucket.io/emumore).

REFERENCES

[1] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012. ISBN 1600490069, 9781600490064. URL https://work.caltech.edu/telecourse.html.

[2] Gary Marcus. Deep Learning: A Critical Appraisal. jan 2018. URL http://arxiv.org/abs/1801.00631.

[3] James Somers. Is AI Riding a One-Trick Pony?, 2017. URL https://www.technologyreview.com/s/608911/is-ai-riding-a-one-trick-pony/.

[4] Model Order Reduction Wiki. URL https://morwiki.mpi-magdeburg.mpg.de/morwiki/index.php/Main{_}Page.

[5] James P. Crutchfield, Ryan G. James, Sarah Marzen, and Dowman P. Varn. Understanding and designing complex systems. dec 2014. URL http://arxiv.org/abs/1412.8520.