UCL CDT DIS Note

UCLCDTDIS-2019-XX

20th April 2020

Draft version 0

Results from TBR Group Project

Petr Mánek[a] and Graham Van Goffrier[a]

[a]University College London

The abstract of my report.

Contents

1 Introduction

The analysis of massive datasets has become a necessary component of virtually all technical fields, as well as the social and humanistic sciences, in recent years. Given that rapid improvements in sensing and processing hardware have gone hand in hand with the data explosion, it is unsurprising that software for the generation and interpretation of this data has also attained a new frontier in complexity. In particular, simulation procedures such as Monte Carlo (MC) event generation can perform physics predictions even for theoretical regimes which are not analytically soluble. The bottleneck for such procedures, as is often the case, lies in the computational time and power which they necessitate.

Surrogate models, or metamodels, can resolve this limitation by replacing a resource-expensive procedure with a much cheaper approximation [1]. They are especially useful in applications where numerous evaluations of an expensive procedure are required over the same or similar domains, e.g. in the parameter optimisation of a theoretical model. The term ”metamodel” proves especially meaningful in this case, when the surrogate model approximates a computational process which is itself a model for a (perhaps unknown) physical process [2]. There exists a spectrum between ”physical” surrogates which are constructed with some contextual knowledge in hand, and ”empirical” surrogates which are derived purely from the underlying expensive model.

In this internship project, in coordination with the UK Atomic Energy Authority (UKAEA) and Culham Centre for Fusion Energy (CCFE), we sought to develop a surrogate model for the tritium breeding ratio (TBR) in a Tokamak-class nuclear fusion reactor. Our expensive model was a MC-based neutronics simulation [3], itself a spherical approximation of the Joint European Torus (JET) at CCFE, which returns a prediction of the TBR for a given reactor configuration. We took an empirical approach to the construction of this surrogate, and no results described here are explicitly dependent on prior physics knowledge.

For the remainder of Section 1, we will define the TBR and set the context of this work within the goals of the UKAEA and CCFE. In Section 2 we will describe our datasets generated from the expensive model for training and validation purposes, and the dimensionality reduction methods employed to develop our understanding of the parameter domain. In Section 3 we will present our methodologies for the comparison testing of a wide variety of surrogate modelling techniques, as well as a novel adaptive sampling procedure suited to this application. After delivering the results of these approaches in Section 4, we will give our final conclusions and recommendations for further work.

1.1 Problem Description

- Introduce nuclear fusion via tokamak [4][5]

- details of JET / CCFE [6]

——————————————————————————-

- Describe tritium breeding [4] - Modelling of reactor geometries [7]

[1] Sondergaard 2003 [2] Myers and Montgomery 2002 [3] Collaboration with Jonathan 2020

[4] Hernandez 2018 [5] Tokamak wikipedia [6] Keilhacker 1999 [7] Coleman 2019

2 Data Preprocessing

Data

2.1 Data Description and Initial Sampling

2.2 Dimensionality Reduction

2.2.1 Principal Component Analysis

2.2.2 Variogram Computations

2.2.3 Autoencoders

3 Methodology

Assuming that input has been appropriately treated to eliminate redundant features, we may turn to characterise proposed surrogate models and the criteria used for their evaluation. The task all presented surrogates strive to solve can be formulated using the language of conventional regression problems. In the scope of this work, we explore various possible choices available to us in the scheme of supervised and unsupervised learning.

Labeling the expensive Monte Carlo simulation $f(x)$, a surrogate is a mapping $\tilde{f}(x)$ that yields similar images as $f(x)$. In other words, $f(x)$ and $\tilde{f}(x)$ minimise a selected similarity metric. Furthermore, in order to be considered *viable*, surrogates are required to achieve expected evaluation time that does not exceed the expected evaluation time of $f(x)$.

In the supervised learning setting, we first gather a sufficiently large training set of samples $\mathcal{T} = \left\{ \left(x^{(i)}, f\left(x^{(i)} \right) \right) \right\}_{i=1}^{N}$ to describe the behaviour of $f(x)$ across its domain. Depending on specific model class and appropriate choice of its hyperparameters, surrogate models $\tilde{f}(x)$ are trained to minimise empirical risk with respect to $\mathcal{T}$ and a model-specific loss function $\mathcal{L}$, where empirical risk is defined as

$$R_{\text{emp.}}(\tilde{f} \mid \mathcal{T}, \mathcal{L}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\tilde{f}(x^{(i)}), f(x^{(i)}) \right). \tag{1}$$

The unsupervised setting can be viewed as an extension of this procedure. Rather than fixing the training set $\mathcal{T}$ for the entire duration of training, the points of evaluation $\{x^{(i)}\}_{i=1}^{N}$ that determine the set are first initialised randomly, and continuously extended throughout training. This permits the learning algorithm to motivate the choice of new points following the evaluation of surrogates trained thus far by appropriately biasing the proposal distribution, in order to better focus on problematic regions within the domain.

3.1 Metrics

Aiming to provide objective comparison of a diverse set of surrogate model classes, we define a multitude of metrics to be tracked during experiments. Following the motivation of this work, we are primarily interested in two properties of surrogates: (i) their capability to approximate the expensive model well and (ii) their time of evaluation. We strive to maximise the former while minimising the latter.

To prevent undesirable bias in results due to training set selection, both properties are collected using k-fold cross-validation with a standard choice of $k = 5$. Herein, a sample set is subdivided into 5 disjoint folds which are repeatedly interpreted as training and testing sets, maintaining a constant ratio of samples. In each such interpretation experiments are repeated, and the overall value of each metric of interest is given by the mean across all folds.

The remainder of this section provides exhaustive list and description of regression performance and evaluation time metrics recorded in the experiments.

3.1.1 Regression Performance Metrics

TODO: provide formal definition of each of the following metrics

Mean absolute error

R^2 ratio

Adjusted R^2 ratio

Standard error of regression

3.1.2 Evaluation Time Metrics

TODO: provide formal definition of each of the following metrics

Training time per sample

Prediction time per sample

3.2 Model Comparisons

3.3 Adaptive Sampling

4 Results

Results

4.1 Results of Model Comparisons

4.2 Results of Adaptive Sampling

5 Conclusion

Conclusion