Benchmarking in the context of data analysis
DRAFT

Felix Strnad

September 8, 2020

Abstract

Contents

1 Introduction

Due to higher computational power and significant advances in algorithm development machine learning (ML) has become a powerful tool for data analysis and model simulations. In particular, deep learning finds it way to more and more applications. Much of the success of deep learning is based on the ability of neural networks to recognize patterns in high-dimensional spaces. This development comes with consequences. On the one hand one has to analyze the new solution methods itself, on the other hand one has to think about benchmark datasets with key questions on which new methods should be applied to. Hence, benchmarking becomes an issue in the context of either **new methods and models** or in the context of **datasets**. Both domains use the term *benchmark*, however they understand something different by it. In this text we briefly summarize both domains by outlining how typically benchmarks are set up in both domains, what challenges are included and provide links to relevant repositories and data collections.

2 Benchmarks for ML algorithms

First one has to define what we mean with a *benchmark* in the context of the usage of ML algorithms. Generally speaking, here we understand the term benchmark as a measure for a method or model which is standard against which you compare against your own the solutions, i.e. to get a feeling for the quality of your solution.

First of all one has to clarify whether the benchmark should be between *different* models to motivate the choice of a specific model or *within* a model to motivate the specific choice of hyperparameters.

2.1 Overview of ML tools

Classical statistical or machine learning methods that are relevant for climate analysis are (list uncompleted):

(i) (Multidimensional-) Linear Regression

(ii) Logistic Regression

(iii) Gaussian Processes (GP)

(iv) (Deep) Artificial Neural networks (ANNs), such as

- Feed Foreward NNs
- CNNs
- Recurrent NNs (e.g. Long-Short Term Memory models LSTMs)

(v) Principal Component Analysis (PCA)

(vi) K-means algorithm

(vii) K-nearest neighbors (KNN)

(viii) Support Vector Machines (SVM)

(ix) Random Forests (RF)

(x) ...

Different python packages implement these algorithms. Most famous python implementations are SCIKITS.LEARN, MLPY, PYBRAIN, PYMVPA, MDP. For deep learning the mainly used libraries are tensorflow and pytorch. When applying a ML tools onto a specific problem, one has to justify his choice by either substantial new insights that other algorithms were not able to achieve or significantly better prediction scores or lower uncertainties.

2.2 Benchmark for single ML algorithm

After having chosen a specific ML algorithm one still needs to figure out if the algorithm performs well, i.e. if the hyperparameters are chosen properly. Hyperparameters can be arbitrarily set by the user before starting training (not to be confused with model parameters that are learned during the model training eg. weights in Neural Networks, Linear Regression,...). Here one benchmark is to think about the Bias-Variance-Tradeoff as a benchmark [1]:

(i) **Overfitting:** Does the model generalize on unseen the data?

(ii) **Underfitting:** Does the model capture the true signal from the data. Underfitted models have bad accuracy in training *as well as* on the test data.

Keeping in mind these key questions can be regarded as a lower baseline for the applicability of a ML algorithm.

2.3 Comparing different ML algorithms

If we want to apply a new advanced ML tool, a benchmark is here meant in the sense of the application of a standard solution on a problem which already performs well and to how much your new model performs the old one or to what degree an established model already succeeds to reproduce your results (or fails at doing so).

These benchmarks can be roughly divided into:

(i) Time (how much computation time and/or wall-clock time is needed).

[1] see e.g. https://towardsdatascience.com/guide-to-choosing-hyperparameters-for-your-neural-networks-38244e87dafe

(ii) Validation score (how "accurate" are predicted results in accordance test data). There are many different validation methods. These are more detailed described and compared in the chapter .

If one wants to compare both, the combination of time and validation score, this is often referred to as *model performance*. Many time and validation scores comparison on multiple machines and OS and different model performance analyses can be found at https://openbenchmarking.org/tests.

3 Benchmark standard datasets for application

Instead of evaluating different ML algorithms, one can think as well about different research fields on which ML methods might bring new insights. To formulate the problem to different communities, in this understanding benchmarks can be regarded as an extension of classical data repositories, called *benchmark standard datasets*, which is often shortened just as *benchmarks* as well. Here, instead of focusing on the ML algorithm part, a benchmark is to be understood as a minimum requirement of skills a new proposed tool should be capable to provide at least.

Benchmark datasets can serve here as summary descriptions of problem areas, providing a simple interface between disciplines without requiring extensive background knowledge. Such benchmark datasets can be analyzed in terms of computational costs, accuracy, utility and other measurable scores (see section 2.3) to address a particular question in climate or geoscience. Therefore, benchmark datasets are a simple way to a domain specific problem (e.g. a Geoscience problem) accessible for data scientists.

3.1 Available benchmark datasets for ML algorithms

More and more code, classification algorithms and datasets are shared among other scientists in the field of ML [Vanschoren et al., 2014]. There is a big data base for open data from any domain available under https://www.openml.org/ and even a python wrapper to directly access these datasets has been implemented.

Detailed analysis for detailed comparison of single methods exists as well. For example there is a database of 165 supervised classification datasets stored under the Penn Machine Learning Benchmark (PMLB) [Olson et al., 2017b], available[2]. Different ML types perform differently on different datasets. In [Olson et al., 2017a] a good overview of the comparison between different ML type algorithms applied on the PMLB benchmark dataset can be found.

3.2 Benchmark datasets for Geoscience

The development of benchmarks is necessary to reinforce the connection between Geoscience and ML which can be achieved through the use of specially-designed benchmark datasets.[Ebert-Uphoff et al., 2017]. Geoscientists can preselect and preprocess interesting data, couple them with interesting and unsolved science questions, and add data documentation and background explanations suitable for non-domain scientists who can consequently process the data further, often by application of ML tools. There are increasing efforts in making benchmark datasets and key questions publicly available. Here are some benchmark datasets:

- For more geoscience related questions, some first datasets are uploaded under the label IS-GEO, described in [Ebert-Uphoff et al., 2017].
- For weather forecasting by purely-data driven prediction of the global atmospheric flow is such a benchmark dataset provided by [Rasp et al., 2020].

3.3 Baselines for Geoscience

To start with a benchmark dataset it is important, for what purpose the dataset is and what already has been done.

[2]The PMLB benchmark dataset for detailed comparison of different types of ML algorithms applied on the same problem can be found https://github.com/EpistasisLab/penn-ml-benchmarks.

In case of predictions or weather forecast methods there are two baselines one has to keep in mind [Rasp et al., 2020]:

i **Persistence**: Input variable=Output variable (in weather forecasts:"tomorrow's weather is today's weather")

ii **Climatology**: Compute daily/weekly/monthly/yearly mean values over past times and compute forecasts based on these mean values.

In the area of climate predictions (not to be confused with weather forecasts), which often are evaluated by comparing long-term statistics to observations, e.g. SST) some baseline rules are [Scher and Messori, 2019]:

- Are seasonal cycles reproduced correctly?
- Are long term mean statistics between ML prediction and comparison data in accordance?

References

I. Ebert-Uphoff, D. R. Thompson, I. Demir, Y. R. Gel, M. C. Hill, A. Karpatne, M. Guereque, V. Kumar, E. Cabral-Cano, and P. Smyth. A VISION FOR THE DEVELOPMENT OF BENCHMARKS TO BRIDGE GEOSCIENCE AND DATA SCIENCE. Technical report, 2017. URL https://is-geo.org/.

R. S. Olson, W. La Cava, Z. Mustahsan, A. Varik, and J. H. Moore. Data-driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070*, 2017a.

R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz, and J. H. Moore. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10(1):1–13, 2017b.

S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. Weatherbench: A benchmark dataset for data-driven weather forecasting. 2020.

S. Scher and G. Messori. Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, 12(7):2797–2809, jul 2019. ISSN 1991-9603. doi: 10.5194/gmd-12-2797-2019. URL https://gmd.copernicus.org/articles/12/2797/2019/.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. Openml: Networked science in machine learning. *SIGKDD Explor. Newsl.*, 15(2):49–60, June 2014. ISSN 1931-0145. doi: 10.1145/2641190.2641198. URL https://doi.org/10.1145/2641190.2641198.