Test your objectives!

Timothy O. Hodson[1], Thomas M. Over[1], Tyler J. Smith[2], and Lucy M. Marshall[3]

[1]U.S. Geological Survey Central Midwest Water Science Center, Urbana, Illinois
[2]Clarkson University, Potsdam, New York
[3]Macquarie University, Sydney

Correspondence: Timothy O. Hodson (thodson@usgs.gov)

Abstract. Choosing the wrong objective function leads to suboptimal calibrations (information loss), unexpected biases, and misrepresented uncertainty. So, stop assuming objectives and start testing them instead. Here, we demonstrate the "classical" method for doing so.

1 Introduction

In machine learning or scientific computing, model performance is measured by an objective function. But why choose one function over another? According to information theory, you should select the objective that maximizes the information in the model, and the most likely choice is whichever objective encodes the error in the fewest bits. To evaluate that encoding, transform the objectives into log-likelihoods ℓ, which normalizes them, then divide ℓ by $\log 2$ to yield the conditional entropy $\hat{H}$, which is the expected number of bits required to encode the error.

A classic example is the mean squared error (MSE), which corresponds to the log-likelihood of the normal distribution. MSE is the optimal objective when errors are normal, independent and identically distributed (*iid*). However, for many problems, the true error distribution is more complex. Rather than assuming a de facto objective function like MSE, information theory can guide that choice. Here, we demonstrate how on a toy runoff-prediction problem. The process is similar to traditional model selection, where the model is varied while the data and objective are held fixed. In objective selection, the experiment is flipped such that the objective is varied while the model and data are fixed. The objective yielding the highest likelihood (shortest encoding) is typically best for that particular problem.

2 Demonstration

Using simple techniques like changing variables, we derived the log-likelihoods ℓ of ten objective functions (Table 1), computed them on a dataset of streamflow predictions, converted those ℓ to conditional entropies $\hat{H}$, which is in bits, then weighted and ranked the results. The best objective encodes the error in the fewest bits.

In the experiment, the data and model were fixed, only the objective varied. Relative to ZMALE, the excess bits in the other objective functions are noise. So, MSE measures at least 40 percent noise, and NSE at least 38 percent. In general, noisier objectives convey less information, so they require more iterations during calibration, yield suboptimal calibrations,

Table 1. Similarity (measured as conditional entropy $\hat{H}$), weights, and ranks of ten objective functions for the test data and model.

Objective	Description	$\hat{H}$, as bit rate	Weight	Rank
MSPE	mean squared percent error	23.54	0.00	10
U	uniformly distributed error	18.17	0.00	9
MSE	mean squared error	11.62	0.01	8
NSE	normalized squared error*	11.20	0.01	7
MAE	mean absolute error	9.49	0.04	6
MSLE	mean squared log error**	7.47	0.15	5
MARE	mean absolute square root error	7.34	0.17	4
ZMSLE	zero-inflated MSLE	7.18	0.19	3
MALE	mean absolute log error**	7.04	0.21	2
ZMALE	zero-inflated MALE	6.95	0.22	1

*Also known as Nash–Sutcliffe efficiency. **Undefined for zero or negative flows but included for context.

and produce models that require more storage space (better model, better data compression). A well-known example of that point is stochastic gradient descent, where noise in the objective causes slower convergence (Bottou and Bousquet, 2007). In that case, each iteration completes faster, so the solution may be reached quicker overall, but in general, a poorly chosen objective incurs a similar penalty but potentially without benefit.

This process of selecting an objective is more-or-less what "universal likelihoods" do automatically, which is a more advanced topic. However, those methods are ultimately limited because the best objective is not computable, so there will always be a role for human intuition and intelligence (Rissanen, 2007) . . . at least until machines can reason as well as humans. This demonstration is pedagogical but also useful in that simple tweaks to "classic" objective functions can have a large effect on model error. We do not advocate for one objective over another —the choice varies from problem to problem— only that benchmarking objectives is good practice that will yield better models.

The authors thank H.V. Gupta for a thought-provoking discussion at HydroML 2022 and the U.S. Geological Survey HyTEST project.

References

Bottou, L. and Bousquet, O.: The Tradeoffs of Large Scale Learning, in: Advances in Neural Information Processing Systems, edited by Platt, J., Koller, D., Singer, Y., and Roweis, S., vol. 20, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2007/file/0d3180d672e08b4c5312dcdafdf6ef36-Paper.pdf, 2007.

Rissanen, J.: Information and complexity in statistical modeling, vol. 152, Springer, 2007.