Abstract: Should Algorithm Evaluation Extend to Testing? We Think So.

Lien Michiels[1,2,†], Robin Verachtert[1,2,†], Kim Falk[3,†] and Bart Goethals[1,2,4]

[1]*Froomle N.V., Belgium*
[2]*University of Antwerp, Antwerp, Belgium*
[3]*Shopify, Canada*
[4]*Monash University, Melbourne, Australia*

Abstract

Software engineers test virtually all of their code through unit, regression and integration tests. In contrast, data scientists and machine learning engineers often evaluate models based solely on their training or evaluation loss and task performance metrics such as accuracy, precision or recall. When 'code' becomes 'algorithms', software best practices are often neglected. In our research, we found that most publicly available algorithm implementations indeed are not tested beyond ranking performance metrics, such as recall and normalized discounted cumulative gain. Applying software testing best practices to algorithms can seem daunting (and unnecessary). However, software packages like scikit-learn and SpaCy have demonstrated that it definitely is possible to test (at least some aspects of) algorithms. We believe that algorithms should be tested. Without tests, you may just end up with dead code paths, gradients that do not update, or logical errors you failed to detect. The question then becomes: How should we test algorithms? During the workshop, we would like to open up this discussion. We start with an overview of software testing paradigms: from black-box to white-box testing, unit to regression testing and more. We then present some examples of testing patterns we have applied to our recommendation algorithm implementations. At the end of the discussion, we hope to have answered some of the following questions: (1) Should recommendation algorithms be tested? (2) What aspects of the recommendation algorithm would benefit most from testing? (3) How can we translate these software testing paradigms to recommendation algorithms? (4) What sorts of tests can we design? (5) Which of these tests should be part of a researcher's standard experimentation pipeline?

We plan to summarize the conclusions of this discussion in a future publication, accompanied by a testing toolkit.

Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2022), September 22nd, 2022, co-located with the 16th ACM Conference on Recommender Systems, Seattle, WA, USA.

[†]These authors contributed equally.

lien.michiels@froomle.com (L. Michiels); robin.verachtert@froomle.com (R. Verachtert); Kim.falk.jorgensen@gmail.com (K. Falk); bart.goethals@uantwerpen.be (B. Goethals)
0000-0003-0152-2460 (L. Michiels); 0000-0003-0345-7770 (R. Verachtert); 0000-0002-3573-9257 (K. Falk); 0000-0001-9327-9554 (B. Goethals)