Interpretable Meta-Score for Model Performance: Extended Abstract

Alicja Gosiewska ALICJAGOSIEWSKA@GMAIL.COM
Why R? Foundation
Katarzyna Woźnica* KATARZYNA.WOZNICA.DOKT@PW.EDU.PL
Faculty of Mathematics and Information Science
Warsaw University of Technology
Przemysław Biecek PRZEMYSLAW.BIECEK@PW.EDU.PL
Faculty of Mathematics and Information Science
Warsaw University of Technology,
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw

Editors: P. Brazdil, J. N. van Rijn, H. Gouk and F. Mohr

This extended abstract is based on a published article with the same title and authors, which appeared in Nature Machine Intelligence 4(9):792–800, 2022.

In benchmarking different machine learning models, we are seeing a real boom—in many applications such as computer vision or neural language processing, more and more challenges are being created (Wang et al., 2018, 2019; Zhai et al., 2020). We are therefore collecting more and more data on the performance of individual models or architectures, but the question remains as to how to decide which model gives the best results and whether one model is significantly better than another. The most commonly used measures, such as AUC, accuracy, or RMSE, return a numerical assessment of how well the predictions of the selected model satisfy specific properties: they correctly assign the probability of belonging to the chosen class, they are not wrong in assigning the predicted class, or the difference between the predictions and the true values is not large. From an application point of view, however, we lack information:

- what is the probability that a given model gets a better performance model than another;
- whether the differences we observe between models are statistically significant;
- in most cases, the values of the selected model performance metrics are incomparable between different datasets, i.e., how to compare a model's AUC improvement by 0.01 if for one dataset the best achieved AUC is of the order of 0.9 and for the other 0.7.

To address these shortcomings, in Gosiewska et al. (2022) we introduce a new meta-measure of model performance—EPP. It is inspired by the Elo ranking used in chess and other sports games. By comparing the rankings of two players and transforming them accordingly, we obtain information on the probability that one player is better than the other. EPP adapts this property to the specific conditions of benchmarks in machine learning but allows for universal application in many benchmarking schemes. We emphasize this by introducing a unified terminology, the Unified Benchmark Ontology, and the description of the new measure is given in these terms. Hence, models are referred to as players and model performance to score.

Definition 1 *The odds(i,j) are odds that Player M_i has a better Score than Player M_j, and are expressed as*

$$odds(i,j) \;=\; \frac{p_{i,j}}{1-p_{i,j}},$$

where $p_{i,j}$ is the probability that Player M_i has a better Score than Player M_j in a random Round R.

Definition 2 *The β_{M_i} and β_{M_j} are EPP Meta-Scores for Players $M_i, M_j \in \mathcal{M}$ respectively if they satisfy the following property*

$$\log \frac{p_{i,j}}{1-p_{i,j}} = \beta_{M_i} - \beta_{M_j},$$

where $p_{i,j}$ can be estimated $\hat{p}_{i,j}$ in two exploratory variables logistic regression of the form

$$\log \frac{\hat{p}_{i,j}}{1-\hat{p}_{i,j}} = \hat{\beta}_{M_i} x_{M_i} + \hat{\beta}_{M_j} x_{M_j}, \qquad where \;\; x_{M_i} = 1 \;\; and \;\; x_{M_j} = -1,$$

where $\hat{\beta}_{M_i}$ and $\hat{\beta}_{M_j}$ are estimated EPP Meta-Scores. For brevity, in the following sections, we refer to them simply as EPP Meta-Scores.

EPP does not introduce a new definition under which aspect we are comparing the performance of models, but for a selected metric, e.g., AUC EPP determines how often a model has a better AUC metric than another. Hence, we prefix EPP as a meta-measure. It can be used and specified for any model performance measure.

From the probabilistic interpretation of EPP and the estimation of EPP coefficients in logistic regression, the new measure has significant advantages as a method for aggregating benchmarks composed of repeated measures, e.g. in cross-validation. To summarise the performance model obtained by a given algorithm on successive folds in cross-validation or on different datasets in NLP benchmarks, an average performance model was most often used to introduce a ranking of algorithms or architectures. EPP is an alternative aggregation method to the mean. Its main advantages are that it takes into account the stability of the performance model obtained - because EPP looks at how often a model has performed better than another, it automatically has more uncertainty about models with unstable, jittery results. This uncertainty is also considered in statistical tests for the significance of differences.

An essential feature of the EPP is the ability to validate how closely the adjusted values of match-winning probabilities between two selected models are close to the observed outcome. Using deviance for a logistic model, we can apply statistical tests to assess the quality of the fit. This is a novel property not available with other aggregation methods.

To fully demonstrate the capabilities of the EPP application, in this paper we presented the use of this measure on two benchmarks—one for the OpenML (Bischl et al., 2021) datasets and five different classification algorithms (logistic regression, knn, two implementations of random forests and gradient boosting) and also for the VTAB (Zhai et al., 2020) benchmark.

Acknowledgements

Work on this project is financially supported by the NCN Opus grant 2017/27/B/ST6/01307

References

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan van Rijn, and Joaquin Vanschoren. OpenML Benchmarking Suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Alicja Gosiewska, Katarzyna Woźnica, and Przemysław Biecek. Interpretable meta-score for model performance. *Nature Machine Intelligence*, 4(9):792–800, 2022.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275, 2019.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. *arXiv preprint arXiv:1910.04867*, 2020.