LCDB 1.0: An Extensive Learning Curves Database for Classification Tasks

Felix Mohr[1], Tom J. Viering[2], Marco Loog[2,3], and Jan N. van Rijn[4]

[1] Universidad de La Sabana, Colombia
felix.mohr@unisabana.edu.co
[2] Delft University of Technology, The Netherlands
{t.j.viering,m.loog}@tudelft.nl
[3] University of Copenhagen, Denmark
[4] Leiden University, The Netherlands
j.n.van.rijn@liacs.leidenuniv.nl

Abstract. Learning curves can be used for for model selection, speeding up model training, and to determine the value of more training data. Yet our understanding of their behavior is rather limited. To facilitate a deepening of our knowledge, we introduce the Learning Curve Database (LCDB), which contains empirical learning curves of 20 classification algorithms on 246 datasets. It unifies the properties of similar high quality databases in that it (i) defines clean splits between training, validation, and test data, (ii) provides training times, and (iii) provides an API for convenient access (pip install lcdb). We demonstrate the utility of LCDB by analyzing some learning curve phenomena. Improving our understanding of these matters is essential for efficient use of learning curves.

Keywords: Learning Curves · AutoML · Meta-learning

1 Motivation

A learning curve plots the performance of a learning algorithm versus the training set size. They can be used for various machine learning tasks, such as: speeding up training and faster model selection [3, 6]. Most of these tasks rely on learning curve extrapolation: for example, training can be done on a smaller training set size, if the curve extrapolation seems to flatten. For accurate extrapolation, we should have a good understanding of typical learning curve shapes, e.g. exponential, power law, etc. However, recent work has illustrated that learning curves can have varying surprising shapes, such as local maxima [6, 1], where learners do not perform better when they receive more data. By publishing this big database, we are aiming to make a first step towards better characterizing learning curve shapes in practice. Previous databases either were not easily accessible, of a smaller scale, or only used a single test set.

2 The Learning Curve Database

In [7] we publish the learning curve database. In this initial version 246 datasets were used. Our main criterion for the source of the data is API-based reproducibility. To this end, we chose OpenML.org [5] as a source, which unlike the UCI, offers an official Python API. As preprocessing steps, missing values are replaced by the median, and categorical features were binarized using Bernoulli encodings. We first make five (possibly overlapping) 90/10 splits. We then further split this largest part again in five 90/10 splits. Stratified splits are made. The outer split serves as test data, the inner smaller split serves as validation data. The larger inner split serves as training set. It is subsampled with a geometric schedule to come to sets of varying sizes to make the learning curve. We use 20 classifiers with default hyperparameter settings from Scikit [4]. All predictions are saved so various evaluation metrics can be used, but we already precompute the error rates, F1, AUC and log loss and computation times. Since we have $5 \times 5 = 25$ training sets, we can make 25 learning curves per learner per dataset. The database can be installed in Python using `pip install lcdb`.

3 Some Preliminary Insights

We have looked at whether error rate curves are monotonically decreasing, convex, whether they show local maxima (peaking) [2], whether curves cross for different learners, and which parametric models provide the best fit.

Preliminary results indicate that most curves seem to be monotone and convex. Local maxima seem to occur for specific learners: LDA, QDA, and the Perceptron. The larger the training set size, the rarer the peaking, non-monotonicity and non-convexity become. We find that Gradient Boosting and Random Forests can start off weak, but when giving sufficient amount of data surpass most other learners, indicating that their curves often cross.

We perform one of the largest curve fitting studies, where we compare 16 parametric models for learning curve fitting. In contrast to other learning curve studies [6], we find that when given sufficient amount of learning curve points for fitting, models with 4 parameters are most competitive. Specifically, we find `mmf4` to be the best overall, with `wbl4` a close second. We used the Levenberg-Marquadt for fitting curves, as standard in the learning curve literature, but we ran into various issues when fitting parametric models, such as not fits that do not converge. In the end, we discarded 2% learning curve fits due to bad quality, indicating the potential of better fitting approaches.

4 A Next Version?

In a next version, we plan to include regression datasets and pipelines for learners. Then features can be scaled or other preprocessing steps can be applied before learning. Finally, in this version of the database, training sets of one size did not overlap with the next. We plan to address this in a next version, so that the database can be used to estimate the value of gathering more data.

References

1. Loog, M., Viering, T., Mey, A.: Minimizers of the empirical risk and risk monotonicity. In: Advances in Neural Information Processing Systems 32. pp. 7478–7487 (2019)
2. Loog, M., Viering, T.J., Mey, A., Krijthe, J.H., Tax, D.M.J.: A brief prehistory of double descent. Proceedings of the National Academy of Sciences **117**(20), 10625–10626 (2020)
3. Mohr, F., van Rijn, J.N.: Learning curves for decision making in supervised machine learning - A survey. CoRR **abs/2201.12150** (2022)
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
5. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: Networked science in machine learning. SIGKDD Explorations **15**(2), 49–60 (2014)
6. Viering, T.J., Loog, M.: The shape of learning curves: a review. CoRR **abs/2103.10948** (2021)
7. Mohr, F., Viering, T.J., Loog, M., van Rijn, J.N: LCDB 1.0: An Extensive Learning Curves Database for Classification Tasks. Accepted at ECML, 2022.