How does scaling a learning curve influence the curve fitting process?

TUDelft

Author: Chaan van den Oudenhoven
C.vandenOudenho@student.nl

Supervisor: Tom Viering, Cheng Yan, Taylan Turan

1. Background

Learning Curves

- A curve showcasing the error rate of a classification model versus the quantity of data used to train it.

Curve Fitting

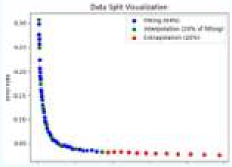
- Trying to fit a given parametric model to learning curve data by tuning the parameters to produce the optimal Mean Squared Error (MSE).

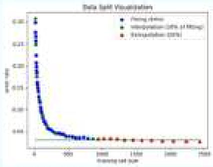
LCDB Database

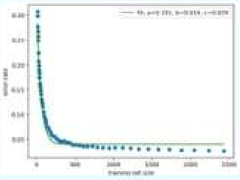
- LCDB [1] database contains learning curves for various learners and datasets. We use these for our experiments

Interpolation and extrapolation

- Interpolation: predicting unseen data points between seen ones
- Extrapolation: predicting unseen data points after seen ones







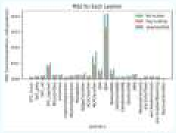
2. Research goal

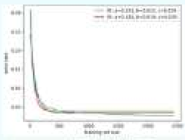
- How does scaling a learning curve influence the curve fitting process in terms of MSE and in terms of successful fit frequency?

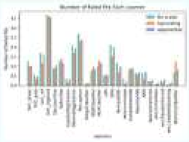
3. Methodology

- We take the learning curves of each learner and dataset pair.
- We take the mean learning curve of the curves that were created using different dataset splits.
- We scale the data and our parametric model using the natural logarithm and the exponential function, fit on the scaled data, and use the parameters found in the scaled space directly in the original space
- We analyse the MSE of this new curve using using the sum of interpolation and extrapolation, and the frequency of succesfull fits

4. Results







- MSE: Log scaling consistently produced better fits while the effects of exponential scaling are inconclusive.
- Successful fit frequency: Very mixed and thus difficult to draw any definitive conclusions

5. Discussion

Limitations:

- Currently only considering exponential and log scaling, should analyse more
- We sample from the uniform distribution for the initial parameter guesses. This could have a large impact
- Change in found parameters could be caused due to the change in loss in the scaled space. Should be investigated by using weights to circumvent this possibility

References

[1]: FELIX MOHR, TOM J VIERING, MARCO LOOG, AND JAN N VAN RIJN. LCDB 1.0: AN EXTENSIVE LEARNING CURVES DATABASE FOR CLASSIFICATION TASKS. IN MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES (ECML PKDD), VOLUME 13717 OF LECTURE NOTES IN COMPUTER SCIENCE, PAGES 3–19. SPRINGER, 2022