Learning Curve Extrapolation using Machine Learning

Authors

Pratham Johari
p.johari@student.tudelft.nl

Supervisor

Mr. Tyalen Dr.Tom Viering

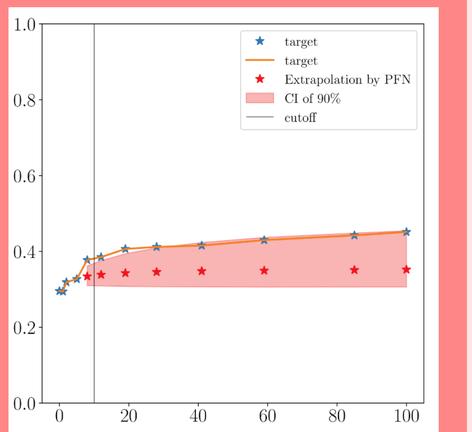
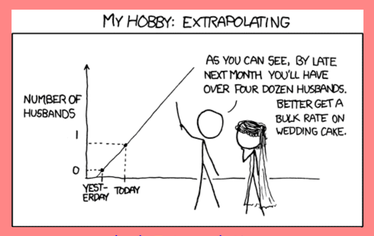


Figure 1 : Visualization of curve extrapolation at 10% Dataset ID is 188, the learner is SVC_rbf



xkcd: Extrapolating

References

1. Mohr, F., T. J. Viering, M. Loog and J. N. Van Rijn (2023). LCDB 1.0: An Extensive Learning Curves Database for Classification Tasks, Springer Nature Switzerland: 3-19.
2. Adriaensen, S., H. Rakotoarison, S. Muller and F. Hutter (2023). "Efficient Bayesian Learning Curve Extrapolation using Prior-Data Fitted Networks."
3. Github Repo url : https://github.com/pratham2442000/BachelorThesis

1 Background

- **Learning curves** show **model performance** with varying **data sizes**.
- Determining the optimal amount of data for a model is beneficial.
- Existing **Parametric models** may not capture **real-world curve complexity**.
- Research focuses on **Learning Curve Prior-Data Fitted Networks.**
- Idea is to use LCDB[1] tests LC-PFN's[2] ability to extrapolate learning curves.

2 Research question

What are the benefits and limitations of using LC-PFN for learning curve extrapolation and how does it compare to other methods?

3 Methodology

1 Database

- Utilization of **Learning Curve Database** (LCDB).
- Focus on the validation curve set (**4367 Curves**).
- Preprocessing involves aggregating curves per dataset and learner pair and converting dataset size into percentages.

2 Machine Learning Model

- Employment of the **Learning Curve Prior Fitted Network** (LC-PFN), a transformer pre-trained on synthetic data.
- LC-PFN trained on right-censored curves from a parametric prior showcasing model accuracy vs epoch; different from the learning curves examined we look at.

3 Model Evaluation Parameters

- Evaluation metric: **Mean Squared Error** (MSE).
- Cutoff percentages for extrapolation: **10%, 20%, 40%, and 80%.**
- Incorporation of baseline models (**last1** and **mmf4**) for comparison.

4 Model Breakdown

- Model Breakdown is defined by verifying **40% target accuracies outside** the LC-PFN's **90% confidence interval.**
- Investigation of breakdown cases by **grouping** curves by **learner** and **dataset**, assessing whether the model tends to **overestimate** or **underestimate** accuracies for points outside the expected CI.

4 Results

LC-PFN Model Performance Assessment

Table 1 : Average MSE of each model at different cutoffs

Cutoff	10%	20%	40%	80%
$last1$	0.0053	0.0032	0.0016	**0.0007**
$mmf4$	0.0091	0.0048	0.0029	0.0022
LC-PFN	**0.0032**	**0.0021**	**0.0013**	0.0009

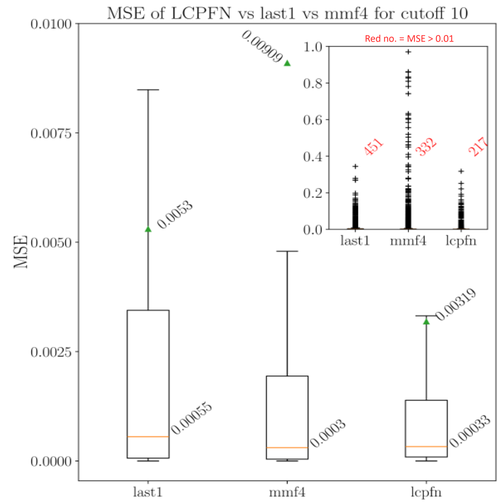


Figure 2 : Box plot of each model at 10% cutoff

LC-PFN Model Shortcomings

Table 2 : Count of curves that causes model breakdown

Cutoff	Count	% Out of total
10%	425	9.73%
20%	572	13.10%
40%	986	22.58%
80%	1255	28.74%

Table 3 : Count of curves by over & under estimating tendencies of the model at different cutoffs

	10%	20%	40%	80%
Under Estimate	266	272	406	543
Over Estimate	158	296	557	703
Neither	1	4	23	9
Total	425	572	986	1255

Table 4 : Count of curves that causes model breakdown grouped by Dataset (OpenML ID)

10		20		40		80	
Dataset id	Count	Dataset id	Count	Dataset id	Count	Dataset id	Count
1457	18	**346**	13	41157	19	**346**	17
1515	10	41157	11	380	19	18	16
446	10	**1465**	11	**1465**	17	299	16
1083	10	1083	11	55	15	1083	15
188	8	1084	11	61	15	1465	15

Table 5 : Count of curve that cause model breakdown grouped by Learner

	10%		20%	
Top 2	SVC_sigmoid	68	SVC_sigmoid	99
	sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis	54	sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis	54
Bottom 2	sklearn.ensemble.RandomForestClassifier	9	sklearn.tree.DecisionTreeClassifier	17
	sklearn.linear_model.LogisticRegression	9	sklearn.ensemble.RandomForestClassifier	15
	40%		80%	
Top 2	SVC_sigmoid	104	SVC_sigmoid	114
	sklearn.linear_model.PassiveAggressiveClassifier	71	sklearn.linear_model.Perceptron	101
Bottom 2	sklearn.ensemble.RandomForestClassifier	40	SVC_poly	47
	sklearn.naive_bayes.BernoulliNB	36	sklearn.ensemble.ExtraTreesClassifier	45

5 Conclusion

Performance Evaluation: Compared LC-PFN to baselines (**last1** and **mmf4**) across different cutoff values.

- LC-PFN demonstrated superior performance at lower cutoffs and remained competitive at higher values.

Analysis of Model Shortcomings: Identified instances of model breakdown due to overconfidence, increasing with higher cutoffs.

- Observed consistent underperformance on specific learners, notably SVC_sigmoid.
- No clear pattern in failing curves when grouped by datasets, though datasets ID 346 and 1465 exhibited worse performance in some cases

6 Future Directions

- Refine LC-PFN training with LCDB. Preliminary attempts in the repository[3].
- Address failures caused by mmf4 and last1 in curve fitting.
- Refine the scaling approach to prevent model breakdowns.
- Explore different percentages of points required outside the Confidence Interval of LC-PFN.
- Determine confidence intervals for mmf4 and last1 and extend to other parametric models for better comparison