The influence of dimensionality on the parameters of the learning curve model

Author: Andrei Mereuta a.mereuta@student.tudelft.nl
Supervisors: dr. Jesse Krijthe, dr. Tom Viering

1 Background

The machine learning curve is useful for many purposes including:

- comparing different algorithms
- choosing model parameters during design
- adjusting optimization to improve convergence
- determining the amount of data used for training

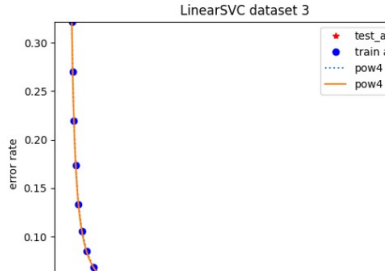


Figure 1: Example of a learning curve

2 Research Question

"How does dimensionality influence the parameters of the learning curve model?"

7 References

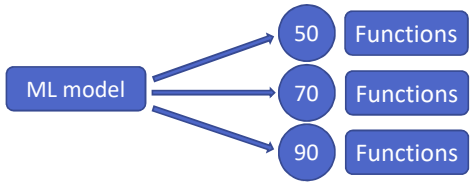
[1] Tom Viering et. al. LCDB 1.0: An Extensive Learning Curves Database for Classification Tasks. 2022
[2] Herve Abdi and Dominique Valentin. Multiple correspondence analysis. volume 2, pages 651–657, 2007

3 Methodology

1. Reproduce experimental setup from LCDB[1]
2. Use **Principal Component Analysis** (PCA) to reduce dimensionality to preserve 50%, 70% and 90% of original information
3. Train all machine learning models from LCDB on the datasets with reduced dimensions
4. Fit suitable functions to describe behaviour of the obtained learning curves
5. Analyze results from two perspectives:

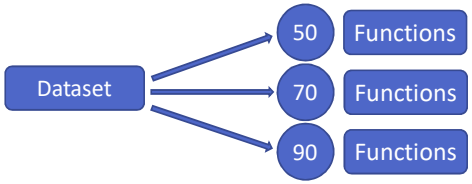
3.1 Machine learning model view

Map **ML model** to the combination of PCA variation and all unique fitted functions with smallest mean squared error (MSE)



3.2 Dataset view

Map **Dataset** to the combination of PCA variation and all unique fitted functions with smallest MSE



4 Results

The table on the right serves as a legend. It shows formulas to all functions, which will be presented further.

Reference Name	Formula
EXPP3	$c - e^{(x-b)^a}$
LAST1	$(a+x) - x$
EXP4	$c - e^{-ax^d+b}$
VAP3	$e^{\frac{a+b}{x+c\log_{10}x}}$
WBL4	$c - be^{-ax^d}$

Table 1:Formula and corresponding reference name

4.1 Machine learning model view

1) Parameters ***a*** & ***c*** decrease as PCA percentage increases

PCA %	*a*	*b*	*c*
50	-0.56	-2.09	-0.006
70	-0.54	-2.45	-0.005
90	-0.55	-2.85	0.002
100	-0.98	-0.05	0.07

Table 2: Results for *ExtraTreeClassifier*, function *expp3*

2) Parameter ***c*** decreases as PCA percentage decreases

PCA %	*a*	*b*	*c*
50	-0.1	-6183.96	3.06
70	-0.15	-6201.53	2.8
90	-0.26	-529.22	2.23
100	-0.44	-7.1	2.1

Table 3: Results for *SVC sigmoid*, function *vap3*

4.2 Dataset view

1) Parameter ***c*** decreases as PCA percentage increases

PCA %	*a*	*b*	*c*
50	-0.26	-5511.14	3.05
70	-0.48	-11878.4	2.47
90	-0.31	-193.84	2.25
100	-0.4	-8.8	2.1

Table 4: Results for openmlid *3*, function *expp3*

2) Almost no change in parameters is identified

PCA %	*a*
50	0.68
70	0.68
90	0.67
100	0.67

Table 5: Results for openmlid *41141*, function *last1*

3) No pattern identified at all

PCA %	*a*	*b*	*c*	*d*
50	280.62	263.12	0.63	0.94
70	505.88	264.47	1.13	0.22
90	6510.91	65.16	0.60	1.23
100	276.56	274.71	0.67	1.26

Table 6: Results for openmlid *41145*, function *exp4*

5 Conclusion

5.1 Machine learning model view gives impression that there might be a dependency between PCA variation and parameters of the fitted function.

5.2 Dataset view does not show any recurring pattern.

Overall conclusion: the research is inconclusive due to limitations primarily concerned with the volume of analyzable data.

6 Limitations & Future Work

1. Not enough data was processed to make more solid conclusions
2. No hyper parameters optimizations
3. Alternatively, choose another dimensionality reduction algorithm, for example: Multiple Correspondence Analysis [2]