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

Andrei Mereuta

Supervisor(s): dr. Jesse Krijthe, dr. Tom Viering

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Andrei Mereuta
Final project course: CSE3000 Research Project
Thesis committee: dr. Jesse Krijthe, dr. Tom Viering, dr. Zhengjun Yue

Abstract

Learning curves in machine learning are graphical representations that depict the relationship between a model's performance and the amount of training data it has been exposed to. They play a fundamental role in obtaining the knowledge and skills across a range of domains. Although there are already quite some researches studying machine learning curves, explaining the importance and practical application of learning curves, we still know very little about the factors that influence the parameters of the learning curve. The aim of this research is to give a better understanding of different factors affecting the parameters of the learning curve. Specifically, we are interested in how the dimensionality of a dataset can influence the parameters of the learning curve. Since learning curves are useful and have several applications, such as estimation of the time required to complete production runs [2], we would like to know if the dimensionality has any effect on the shapes of learning curves. To conduct the research I applied principal component analysis (PCA) three times with different amount of information preserved to reduce number of dimensions on several datasets and analysed the changes in the parameters of the obtained learning curves. The research showed that potentially there might be some relation between dimensionality and shape of the curve, but only in cases of specific machine learning model. The amount of experiments conducted is not sufficient to make solid conclusions and it is advised to continue with proposed experimental setup, but train machine learning models on increased number of datasets.

1 Introduction

A learning curve is a plot used to show performance of a model as the training set size increases. The reasons for researching machine learning curves are important from multiple perspectives. Studying learning curves gives insights into the behaviour of an algorithm, what data and what types of data are facilitating best performance of an algorithm. Vast amounts of researches have been conducted to model the shapes of learning curves, [4], [12]. In their study Viering and Loog conclude that there does not exist the universal shape of the learning curve [12].

The aim of this research is to give some insight into factors that influence machine learning curves. Specifically, we are going to look into how the dimensionality of the dataset affects the parameters of the learning curve. Answering this question can decrease time spent on training machine learning algorithm. If we find any evidence that shows, that training a specific machine learning with decreased dimensionaliy leads to the same or almost the same performance, that would mean that we can save some time training the algorithm. Looking at the parameters of a learning curve we can determine whether a machine learning model learns relatively fast or slow. The idea will be more thoroughly explained further in the section 5.

To summarize this research will answer the question:

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

2 Related Work

In this section I analyse previous works that are relevant to my study. I start with Bui [3], who concludes that there should always exist the ideal number of features such that the curve behaves exponentially, it relates to my research as I am also interested in the influence of dimensionality factor on the shape of the learning curves. Additionally, the analysis indicates that as the number of discretized features increases, the behavior of the learning curve becomes increasingly unpredictable. This work is very valuable and provides a solid basis for my research. Bui argues that dimensionality experiment can be conducted using other dimensionality reduction technique rather than PCA, the other techniques will be discussed in section 3.1.

It is important to emphasize the reason we are interested in researching the effect of dimensionality on the shapes of the learning curves. One of the most known problems to all researchers in machine learning field is *curse of dimensionality*. Lei Chen states that *curse of dimensionality* means that the number of objects in the data set that need to be accessed grows exponentially with the underlying dimensionality [6]. For my research it means, that understanding how dimensionality influences the shape of the learning curve we can potentially avoid or at least partially compensate curse of dimesionality by wisely choosing the PCA variation and/or machine learning model.

3 Methodology

To conduct empirical study of the influence of dimensionality on the shapes of the learning curves, I created the setup, which consists of the following steps:

1. Prepare experimental setup
2. Conduct experiments and collect data
3. Group the obtained data and analyze

Further, each step is described in detail.

3.1 Prepare Experimental Setup

To analyse the influence of dimensionality on the parameters of the learning curve, I decided to apply PCA on the initial datasets. Principal component analysis is a technique used to reduce dimensionality of a dataset preserving the maximum amount of information.

An alternative to PCA is Linear Disciminant Analysis (LDA). Traditionally LDA is used as a classifier, but can be also used to select most important features. In the [11] paper authors state that LDA can be applied for feature selection. However, Martinez and Kak conclude in their work [7] that PCA is more efficient than LDA.

Another alternative to PCA is using Genetic Algorithms to reduce dimensionality. The paper [10] presents how Genetic

Algorithms are used to select features from the feature space. Zamalloa et al. in their paper [13] discover that Genetic Algorithms outperform PCA and LDA on some datasets, but on other datasets PCA shows better results.

Taking into consideration all of the above, I decided to choose PCA as dimensionality reduction algorithm as it shows best performance on average.

3.2 Conduct Experiments and Collect Data

The experimental setup from previous step is quite flexible and universal, applying it to all datasets and all machine learning algorithms outlined in LCDB paper [9] would be very time consuming, therefore it was decided to take only random subset of datasets with all machine learning algorithms trained on these datasets. For each combination of dataset and machine learning algorithm I apply three variations of PCA, they are:

- 90% or more of variance is explained
- 70% or more of variance is explained
- 50% or more of variance is explained

The reason I chose the outlined above percentages comes from the paper [5]. It states that total variance explained by all components should be around 70%. I decided to agree with the results of the paper [5] as I found other articles coming to similar conclusions[1]. However to better analyse the change in parameters of the fitted functions I decided to take two more variations of PCA, which are 50% and 90%. For my research it should be enough to select these three variations as they are most demonstrative.

The data obtained at this point only illustrates the performance of machine learning algorithms over selected datasets, the last step is to apply fitting procedure from LCDB repository[2] and the useful and comparable data is obtained.

3.3 Group the Obtained Data and Analyze

The data obtained in the previous step is grouped by the variance explained and then further grouped by the fitted functions. The parameters for fitted functions are averaged and compared with parameters of the same function, but with different PCA variation.

4 Experimental Setup

To conduct experiments I broke down the whole pipeline into three isolated atomic parts, which are: training machine learning models on different datasets and using different PCA variations, fitting the results of the trained models, analyzing the change in fitted learning curves. Analysing each part individually will help in understanding the whole setup and the idea I used while setting up and conducting experiments.

[1]https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6

[2]https://github.com/fmohr/lcdb

4.1 Training of Machine Learning Models

To imitate training as closely as possible to the procedure described in LCDB paper[9], I used file **database-accuracy.csv**, which already contains seeds for splitting data into training, validation and test sets. I have adjusted existing code from LCDB repository to use the above mentioned seeds and to use same sizes for training and testing. What is also important to mention is that while repeating training procedures all random seeds (both for the machine learning models and for splitting the data) were the same for consistency reasons.

4.2 Fitting Procedure

In the fitting procedure I almost fully use code from the LCDB repository to fit the most suitable function to describe behaviour of the corresponding learning curves. As an output of this step I have fitted functions, their corresponding parameters and useful metrics such as mean squared errors for training and test datasets.

4.3 Analysis Setup

I analysed the obtained fitted functions and parameters from two different angles: machine learning model view and dataset view.

The machine learning model view maps learning model to the combination of PCA variation and all unique fitted functions with smallest mean squared error (MSE) across all datasets. Afterwards, for each unique fitted function I average the parameters. This perspective allows me to see if there is any dependency between parameters of fitted functions and machine learning model.

The dataset view uses similar approach as the previous one. The difference is that I average parameters across machine learning models trained on specific dataset. The idea was to spot any tendency in parameters for the datasets.

In the table, I illustrate all parameters and fitted functions for all three PCA variations. The tables with results are presented in the next section.

The whole setup described in this section is generic and flexible allowing to look at the results from multiple perspectives. It increases chances of discovering interesting dependencies or phenomenons, but it also means that the number of questions and hypothesises potentially increases.

5 Results

This section shows the results of the conducted experiments. The tables presented in this section are divided into two categories:

1. Tables that show best function for the dataset 5.1.
2. Tables that show best function for the machine learning model 5.2.

The best fitting function is defined by the averaged MSE across test sets. The tables show the change in parameters for different PCA variations and different *openmlids*. For the reference, I also include all below mentioned formulas.

Table 1: Formula and corresponding reference name

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	$c - be^{-ax^d}$

5.1 Function vs. Dataset

Looking at the first table, we can notice the continuous decrease in parameter *c*, while increasing PCA %. However, it is important to note that *expp3* does not have the lowest MSE in all four cases, still in all four cases the MSE is lower than **0.001**.

Table 2: Results for openmlid *3*, function *expp3*

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

In this case, we barely see any change in parameter change. Very important to note, that the fitted function is *last1*, which is the least sophisticated function.

Table 3: Results for openmlid *41142*, function *last1*

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

In this example, we see completely different situation, compared to previous two tables. There is no pattern visible.

Table 4: Results for openmlid *41145*, function *exp4*

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

5.2 Function vs. Machine Learning Model

In the next table, I present the *ExtraTreesClassifier* and immediately notice the dependencies in the parameters *a* and *c*. As we can see the bigger the PCA is, the smaller are the two parameters.

Table 5: Results for ML model *ExtraTreesClassifier*, function *expp3*

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

In this table we see a tendency for parameter *c* to decrease with the decrease in dimensionality, specifically for *SVC sigmoid* machine learning model. Important to note that funciton *vap3* does not always have the smallest MSE, however it is always smaller than **0.001**.

Table 6: Results for ML model *SVC sigmoid*, function *vap3*

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

6 Conclusions

To summarize, I tried to understand whether dimensionality of the dataset influences the parameters of the learning curve. To find answers I showed a comparison of the original learning curves from LCDB between the learning curves of the datasets transformed by three PCA variations.

The results for the *function vs. dataset* 5.1 do not show any recurring pattern. It means that even if there is a pattern or dependency between the amount of information preserved by PCA and the parameters of the fitted function, they should be individual and are related to the nature or individual characteristics of a dataset.

The results for the *function vs. machine learning model* 5.2 tend to have a pattern or dependency between the amount of information preserved by PCA and the parameters of the fitted function more often. However, I have trained machine learning models on at most five different datasets, which in my opinion is not enough to make a solid conclusion.

In my opinion, the research is inconclusive due to several limitations, which will be discussed in detail in section 7. Briefly, there is not enough data to make solid conclusions. To gain more data we need to conduct more experiments on more datasets from LCDB, following the outlined setup from section 4, this will be more thoroughly discussed in the section 7.3.

7 Discussion, Future Work and Limitations

In this section we will discuss what else could have been done, which limitations I encountered while researching and what is still left unanswered.

7.1 Experimental Setup Limitations

The research allowed me to create the code to generate learning curves for multiple *openmlids*. The code generates learning curves for all machine learning models, which are used in the LCDB paper[9]. The first limitation I found was that I have not optimized hyper parameters of the machine learning models, due to time limitations. In the future it is strongly advised to redo all the experiments, to see if the outcome will change. Since, not all datasets were used and the setup still had small deviations from the original setup I expect to obtain more precise numbers for averaged parameters.

Second setup limitation concerns the nature of Principal Component Ananlysis and the details of the library responsible for reducing dimensionality. While conducting the exper-

iments I noticed that for some datasets the results for fitted functions and parameters were the same for different PCA variations. After taking a closer look I concluded that it happened, because PCA algorithm from *sklearn.decomposition* selects the number of components such that the amount of variance that needs to be explained is greater or equal to the percentage specified. Therefore it might happen that number of components for PCA=70% is the same as for PCA=50%. As a potential outcome, we might see that parameters of fitted functions will be the same for PCA=70% and PCA=50%, which will not give us any meaningful information. To illustrate the consequences of the above mentioned limitation, I demonstrate a table, which is an actual result of one of the conducted experiments:

Table 7: Results for openmlid *44*, function *wbl4*

PCA %	*a*	*b*	*c*	*d*
50	37.79	1.63965980e+08	3.16	2.37
70	37.79	1.63965980e+08	3.16	2.37
90	37.79	1.63965980e+08	3.16	2.37

Closer investigation showed that when PCA was applied it narrowed down initial dataset to one dimension in all cases.

7.2 Processing the Curves Limitations

For processing the learning curves I chose two approaches, described in section 4. However, the data I generate for learning curves is extensive and can be analysed from different perspectives. I looked at the data and analysed behaviour of the fitted functions from the dataset point of view and machine learning model point of view. In essence it means that I analyse all data associated with the chosen dataset or with the chosen machine learning model. Alternatively, in the cases where I was averaging parameters we can only choose the best set of parameters, based on the smallest MSE, or we can choose another metric alternative to MSE, for example mean absolute error (MAE).

Another limitation I found concerns the parameters of the fitted functions. Analysing each function for specific variation of PCA shows that in some rare cases the beta for specific fitted function can be too high or too low, indicating that it might be the outlier. In the future, I advise to filter such results and not include outliers into calculation of average parameters.

7.3 Future Work

As it was already mentioned in section 6, there are several aspects of the research that can be improved to obtain reliable results.

First of all, the I trained machine learning algorithms on five datasets only, ideally we should use all 246 datasets from LCDB. This will increase the accuracy of the results.

Secondly, the setup itself is not ideal as well, it can be improved by setting hyper parameters more wisely.

Thirdly, we used PCA as dimensionality reduction technique, but we can also try and use other techniques and compare results. Other tehcniques are LDA, Genetic Algorithms or Multiple Correspondence Analysis (MCA)[1].

Finally, instead of using MSE as a metric to choose best performing learning curve, we can also use MAE.

8 Responsible Research

In this section, I will reflect on the reproducibility of my experiments and the scientific integrity of the report.

All the experiments were conducted locally on my machine. I use HP Probook 450 G5 with Windows 10 operating system installed on my machine. All the code used for experimental setup is made open source and posted on GitHub[3]. I do not expect any code changes to be introduced, therefore no versioning recommendations should be followed. All the datasets were taken from the public online dataset library called *OpenML*[4].

Throughout the research process, I thoroughly examined the potential ethical implications stemming from the identified results. The focus of the study was to investigate how the dimensionality affects the shape of the learning curve, with the aim of uncovering methods to enhance the performance of machine learning algorithms, see section 1. It is important to acknowledge that the findings could potentially be exploited by malicious individuals who may misuse the information to develop algorithms for nefarious purposes. Miller states that machine learning algorithms can be used for evil purposes, for example to influence voters' opinion by creating precise voters' profiles [8]. Therefore ethical aspects should be always considered in any research revolving around machine learning.

References

[1] Hervé Abdi and Dominique Valentin. Multiple correspondence analysis. volume 2, pages 651–657, 2007.

[2] Michel Jose Anzanello and Flavio Sanson Fogliatto. Learning curve models and applications: Literature review and research directions. volume 41, pages 573–583, 2011.

[3] NAM THANG Bui. Factors related to dataset that influence the shape of learning curves. 2022.

[4] David Cohn and Gerald Tesauro. Can neural networks do better than the vapnik-chervonenkis bounds? 1990.

[5] Statistical Consulting IDRE. Principal components (pca) and exploratory factor analysis (efa) with spss. 2020.

[6] Chen Lei. Curse of dimensionality. In *Encyclopedia of Database Systems*, pages 545–546, Boston, MA, 2009. Springer US.

[7] A.M. Martinez and A.C. Kak. Pca versus lda. volume 23, pages 228–233, 2001.

[8] Seumas Miller. Machine learning, ethics and law. volume 23, May 2019.

[9] 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. Research Track -*

[3] https://github.com/Andrew-Mereuta/learning-curve-experiments

[4] https://openml.org/

European Conference, ECML PKDD 2022, Grenoble, France, September 19-24, 2022, 2022.

[10] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain. Dimensionality reduction using genetic algorithms. volume 4, pages 164–171, 2000.

[11] Fengxi Song, Dayong Mei, and Hongfeng Li. Feature selection based on linear discriminant analysis. In *2010 International Conference on Intelligent System Design and Engineering Application*, volume 1, pages 746–749, 2010.

[12] Tom Viering and Marco Loog. The shape of learning curves: a review. 2022.

[13] Maider Zamalloa, Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, German Bordel, and Juan Uribe. Comparing genetic algorithms to principal component analysis and linear discriminant analysis in reducing feature dimensionality for speaker recognition. pages 1153–1154, 07 2008.