Empirical Investigation of Learning Curves: Assessing Convexity Characteristics

AUTHORS

Student: Kristian Gogora (K.Gogora@student.tudelft.nl)
Supervisor: Dr. Tom Viering (t.j.viering@tudelft.nl)
Responsible Professor: Dr. Jesse Krijthe (j.h.krijthe@tudelft.nl)

AFFILIATIONS

EEMCS, Delft University of Technology, The Netherlands

1 INTRODUCTION

- Learning curves: the rate at which a machine learning model improves with respect to the number of data samples
- Anchors: They are the data points on the learning curve graph
- Convexity: Generally researchers regard learning curves as convex functions [1]
- LCDB: Learning curve database [2], providing learning curves for different ML models and datasets

2 OBJECTIVE

> Estimate the convexity of the learning curves and decide whether they are convex or nonconvex.

3 METHODOLOGY

Second Derivative Estimation

- *Discrete function*
- *Linear regression of neighbouring anchors*
- *Use Linear regression again on the first derivative*

Measures of convexity violation:

- *Fraction of convex vs nonconvex anchors*
- *Confidence interval of convexity*

Rank learning curves based on the above-specified measures.

4 RESULTS

- 8,36% of learning curves in LCDB are nonconvex
  - The most nonconvex learner was *sigmoid SVC*, with 28.34% of its curves being nonconvex
- Identified 2 types of nonconvex curves that were the most common, also see Figures 1 and 2 for an example
  - Type I: An initial increase in error-rate, after which the error-rate stagnates
  - Type II: The error-rate does not change until a larger sample size, where it experiences a sudden drop

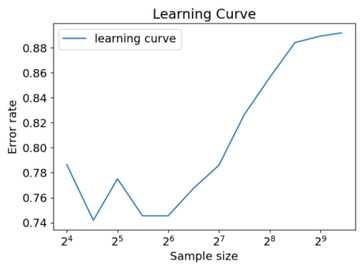

Figure 1: Type I Nonconvex Learning Curve

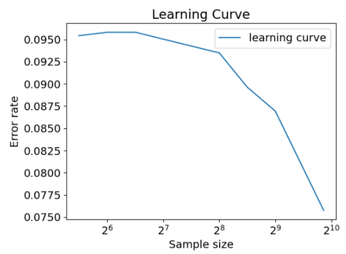

Figure 2: Type II Nonconvex Learning Curve

- On average 58.36% of the anchors per Learning Curve had confidence interval violation
- LCDB violation ranked high double peaking learning curves, Figure 3 shows one such learning curve

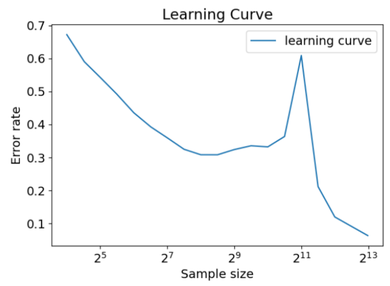

Figure 3: Double peaking Learning Curve

5 CONCLUSION

- The majority of the learning curves are convex
- Nonconvex learning curves are usually undesirable
- More precise results can be obtained by adding more data
- LCDB ranks double peaking learning curve high in its convexity violation metric

6 FUTURE LIMITATIONS

- Use Quadratic Regression instead of Linear Regression
- Define a convex curve as a convex hull or convex set
- We can only estimate the Second Derivative on the limited domain of the learning curve

RELATED LITERATURE

[1] Georgi Li and S. Rajagopalan. A learning curve model with knowledge depreciation. European Journal of Operational Research, 105(1):143–154, 1998.
[2] LCDB. https://github.com/fmohr/lcdb. Accessed: 2023-04-30.

TU Delft