A Comparative Analysis of Learning Curve Models and their Applicability in Different Scenarios

Anna Kalandadze
a.g.kalandadze@student.tudelft.nl

Responsible Professor: Dr. Jesse Krijthe
Email: J.H.Krijthe@tudelft.nl
Supervisor: Dr. Tom Viering
Email: t.j.viering@tudelft.nl

1 Background

Learning curve in Machine Learning - a plot that shows performance of the algorithm that is trained versus the dataset size.

Motivation for studying learning curves

Learning curves show when error rate stops to reduce significantly depending on dataset size

↓

Time and cost of data collection can be reduced

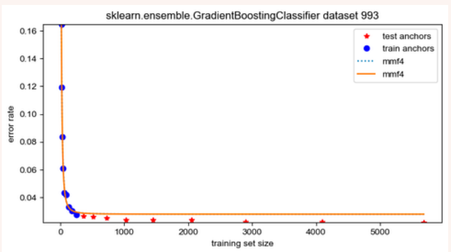

Figure 1: Example of learning curve

2 Research question

Which learning curve model provides the best fit in what case?

Hypothesis: there exist characteristics of dataset and model chosen that lead to identical shape of learning curve.

3 Methodology

1. Collect fitting results from learning curve database [1]
   - predictions, scores, metrics to measure performance
2. Analyse measure metrics: mean squared error and mean absolute error
3. Using chosen metric, find if there are patterns when certain learning curve gives best performance
   - number of features
   - number of outliers
   - number of classes
   - machine learning model

Use **statistical tests** to find out if one curve outperforms the other or works better given certain characteristic

4 Results

1 Machine learning model

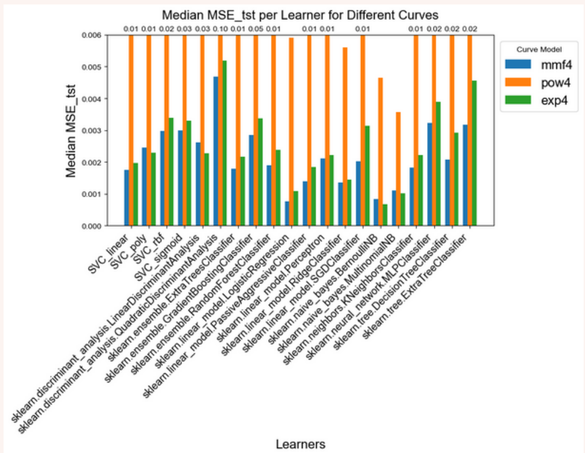

Figure 2: Comparison of mmf4, pow4 and exp4 learning curves using MSE based on machine learning model

mmf4 and exp4 outperform pow4

Best performance:
- Logistic Regression, Multinomial Naive Bayes, Bernoulli Naive Bayes

Worst performance:
- Quadratic Discriminant Analysis

2 Number of features

Table 1: Pairwise comparisons for MSE and MAE based on number of features. X denotes no significant difference

Bucket (Features)	MSE			MAE		
	mmf4 vs pow4	mmf4 vs exp4	exp4 vs pow4	mmf4 vs pow4	mmf4 vs exp4	exp4 vs pow4
[0, 10)	mmf4	mmf4	exp4	mmf4	mmf4	exp4
[10, 20)	mmf4	mmf4	exp4	mmf4	mmf4	exp4
[20, 30)	mmf4	x	x	mmf4	x	x
[30, 40)	mmf4	mmf4	x	mmf4	x	x
[40, 50)	mmf4	x	exp4	mmf4	x	exp4
[50, 60)	mmf4	x	exp4	mmf4	x	exp4
[60, 70)	mmf4	x	exp4	mmf4	x	x
[80, 280)	mmf4	x	x	mmf4	x	x
[1880, 100001)	mmf4	x	x	mmf4	x	x

3 Percentage of outliers

Table 2: Pairwise comparisons for MSE and MAE based on percentage of outliers. X denotes no significant difference

Bucket	MSE			MAE		
	mmf4 vs pow4	mmf4 vs exp4	exp4 vs pow4	mmf4 vs pow4	mmf4 vs exp4	exp4 vs pow4
[0, 0.5)	mmf4	exp4	exp4	mmf4	mmf4	exp4
[0.5, 1)	mmf4	x	exp4	mmf4	x	exp4
[1, 3.5)	mmf4	x	x	mmf4	x	x
[3.5, 6)	mmf4	x	x	mmf4	mmf4	x
[6, 18.61)	mmf4	x	exp4	mmf4	x	exp4

4 Number of classes

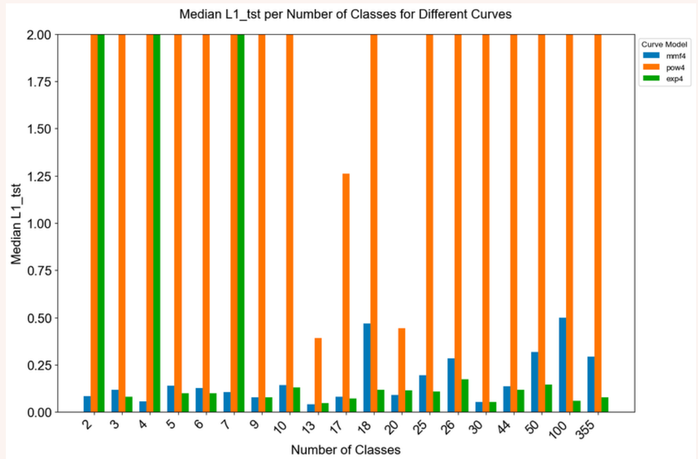

Figure 3: Comparison of mmf4, pow4 and exp4 learning curves using MAE based on the number of classes

1. mmf4 outperforms pow4 for n < 20
2. mmf4 outperforms exp4 for n < 5
3. exp4 outperforms pow4 for n < 10

5 Pearson correlation between outliers in features and predictions *r = 0.101*

5 Conclusion & Limitations

- Proved that mmf4 performs best on average
- Parametric models rank machine learning models similarly
- Exponential curve shows better results than power law when there exists significant difference

- Not enough data for number of classes and features
- Averaged performance for all training set sizes might be not optimal
- All characteristics are inspected separately
- Use MSE/MAE as a measure not considering noise

Future work

✓ combine characteristics

✓ add more data

✓ find optimal training/test size

[1] Mohr, Felix, et al. "LCDB 1.0: An extensive learning curves database for classification tasks." Machine Learning and Knowledge Discovery in Databases, ECMLPKDD. (2022).