Factors related to dataset that influence learning curve shapes

N. T. Bui

Delft University of Technology, The Netherlands.

Introduction

- A learning curve is a tool for plotting a machine learning model's generalization performance against incremental subsets of training data.
- Applications: model selections, reducing complexity
- However, no universal model for learning curve shapes [2]
- Lower dimensionality, worse performance?
- Can noise make the problem more complex?
- Can discrete problem [1] be learned faster (exponentially)?
- Research question:

"How do the inherent factors related to the datasets such as noise, types of numerical input, and dimensionality influence the shapes of the learning curves?"

Methodology

Learning curve generating

- *Kfold cross-validation*
- Anchor (training size) $s_i = \lceil 2^{\frac{7+i}{2}} \rceil$ [3]
- Anchor $S_i \subset S_{i+1}$
- For each anchor:
  - Preprocessing
  - Tuning
- Average over k to get the mean learning curve

Feature noise model

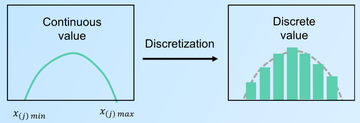
Every entry x_{ij} in the matrix $X \in R^{N \times D}$ will be replaced by x'_{ij} which can be calculated as follows:

$$x'_{ij} = x_{ij} + n\sigma_{xj} z_j \; if \; n > 0$$

Where:

- n is the noise level
- σ_{xj} is the standard deviation of feature x_j
- z_j is random variable: $z_j \sim N(\mu, \sigma^2)$ for normally distributed noise and $z_j \sim U(a, b)$ for uniformly distributed noise

Equal width interval binning



$$k \text{ equally sized bins where bin width} = \frac{x_{(j)\,max} - x_{(j)\,min}}{k}$$

Dimensionality

Using Principal Component Analysis (PCA) to transform the original dataset into different datasets with different dimensions

Results

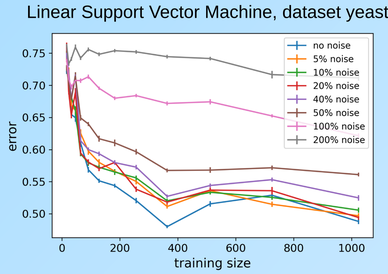


Figure 1: The learning curves of the dataset injected normally distributed noise

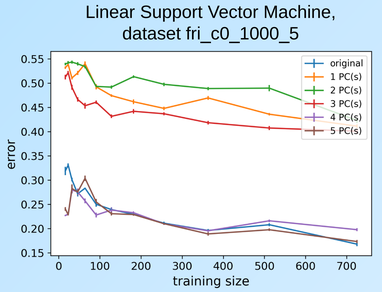
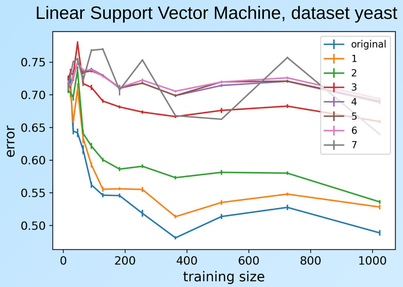
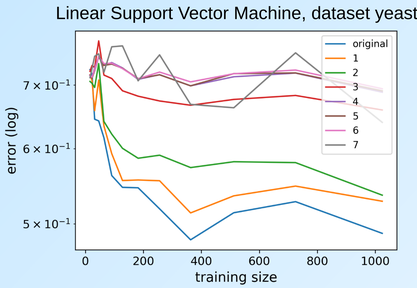


Figure 2: The learning curves of the dataset with different principal components



(a) Unscaled version



(b) Log scale error

Figure 3: Learning curves of dataset with different discretized features. The number of bin $k = 10$

Discussion & Conclusion

- All the experimented curves seem not be well-behaved and not monotonically decreasing.
- Regardless noise distribution types, the higher the noise level, the more complex the problem and the more varying the learning curve shapes.
- Some noisier curves cross less noisy curves.
- The discretized curves are considered strange curves and do not behave exponentially, thus not belonging to the discrete problem class in [1].
- The curves cross each other when many features is discretized
- The lower numbers of principal components, the worse performance of the learners and the more unpredictably the shapes of the curves change
- Ideal number of dimensions for different anchors.

References

[1] D. Cohn and G. Tesauro, "Can neural networks do better than the vapnikchervonenkis bounds?" in Advances in Neural Information Processing Systems, R. Lippmann, J. Moody, and D. Touretzky, Eds., vol. 3. MorganKaufmann, 1990. [Online]. Available: https://proceedings.neurips.cc/paper/1990/file/816b112c6105b3ebd537828a39af4818-Paper.pdf

[2] T. J. Viering and M. Loog, "The shape of learning curves: a review," CoRR, vol. abs/2103.10948, 2021. [Online]. Available: https://arxiv.org/abs/2103.10948

[3] F. Mohr, T. J. Viering, M. Loog, and J. N. van Rijn, "Lcdb 1.0: An extensive learning curves database for classification tasks," unpublished.