Tradeoff of generalization error in unsupervised learning

Gilhan Kim[1], Hojun Lee[1], Junghyo Jo[2], Yongjoo Baek[1]
[1]*Department of Physics and Astronomy & Center for Theoretical Physics, Seoul National University, Seoul, South Korea,*
[2]*Department of Physics Education & Center for Theoretical Physics, Seoul National University, Seoul, South Korea*

Finding the optimal model complexity that minimizes the generalization error (GE) is a key issue in many machine learning tasks. For conventional examples of supervised learning, it is well known that the GE can be decomposed into two parts, namely the bias and the variance. A too simple model may not be able to capture many features of the given data set, yielding a highly biased training result with low variance (underfitting). In contrast, a too complex model may learn even the sampling noise of the data set, which results in a highly variable training result with low bias (overfitting). In this manner, lowering the bias by making the model more complex typically entails an increase in the variance, inducing the U-shape behavior of the GE with the model capacity. This phenomenon is called the 'bias-variance tradeoff.' In this study, we investigate whether unsupervised learning also exhibits the same tradeoff behaviors by training the restricted Boltzmann machine to generate the steady-state configurations of the two-dimensional Ising model at a given temperature and the totally asymmetric simple exclusion process with given entry and exit rates. In fact, we find that a straightforward generalization of the bias-variance decomposition to unsupervised learning leads to the non-monotonic dependence of the bias on the model complexity. This indicates that the bias–variance decomposition is not useful for describing the generalization behaviors of unsupervised learning. Instead, we propose a different decomposition of the GE, which consists of the model error (ME) and the data error (DE)—the former quantifies the minimum GE achievable by the model, and the latter quantifies the excess error that mainly stems from the sampling error of the training data set. We find that using a more complex model reduces the ME at the cost of the data error, with the data error playing a more significant role for a smaller training dataset. Based on these findings, we claim that the optimal model complexity of unsupervised learning is determined by the 'ME-DE tradeoff.' We also find that the optimal model tends to be more complex when the data to be learned involve more correlations.