Question_9

Given a data set, it is possible to have different models that are trained on the same task and give predictions with approximately-equally accuracy, which is called the Rashomon effect in machine learning. Accordingly, a model set containing all these accurate predictive models is named the Rashomon set. Ideally, as long as the Rashomon set is large enough, there must exist one interpretable model. In particular, there is often a simpler model within the Rashomon set that is both accurate and interpretable.

Before trying to find the simpler models within the Rashomon set, it would be wiser to first prove the existence of such models. Lesia and et al. propose some assumptions where one or multiple simpler models can be proved to exist in the Rashomon set once these assumptions are satisfied[1]. The researchers also mention that for practical application, it is possible to find a larger Rashomon set on a complex space first, and then locate interpretable simpler models from a simpler space within the Rashomon set by posing constraints. From this perspective, finding interpretable models with the Rashomon set seems to be practicable. But proving the existence of such models alone can be difficult in reality, not to mention solving the models. For instance, although a deep neural network can fit highly complex data, it could be challenging to solve it for it has a large number of parameters. Lesia and et al. use finding optimal sparse, accurate models as an example to show that the process could be Non-deterministic Polynomial hard.

However, there may be problems with finding interpretable models using the Rashomon set. On one hand, although models in the Rashomon set have similar accuracy or error rate, they achieve that by using different algorithms or parameterizations, or focusing on different features of the data set. This means that even for the same training task and the same prediction result, these models could give quite different interpretations which needs more careful consideration. For another, Breiman points out that during the construction of the model, even a small perturbation of the data could lead to the skip of one model to another[2]. And these two models could be very distinctive in spite of their similar accuracy. A case in point would be logistic regression where a common practice is the deletion of less important variables. By deleting some variables, different models can lead to totally distinct prediction conclusion.

In conclusion, it is feasible to use Rashomon set to capture interpretable models theoretically, but there may be some aspects that need more thoughtful consideration in practical application.

References:

1. Semenova, L. and C. Rudin, *A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning*. 2019.
2. Breiman, L., *Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author).* Statist. Sci., 2001. **16**(3): p. 199-231.