Week 12 Assignment

1. **Using the data synthesis R script provided by the instructor as part of the week 11 assignment instructions, produce datasets of the following sizes, and fit deep learning models with the configurations shown below. Associated with each model, record the following performance characteristics: training error, validation (i.e., holdout set) error, time of execution. Use an appropriate activation function**.

Data size	Configuration	Training error	Validation error	Time of execution
1000	1 hidden layer 4 nodes	0.4602	0.4520	5.87 (s)
10000	1 hidden layer 4 nodes	0.0334	0.281	15.70
100000	1 hidden layer 4 nodes	0.0062	0.0067	156.31
1000	2 hidden layers of 4 nodes each	0.3261	0.2487	4.94
10000	2 hidden layers of 4 nodes each	0.0183	0.151	22.07
100000	2 hidden layers of 4 nodes each	0.0037	0.0039	160.87

2. **Based on the results, which model do you consider as superior, among the deep learning models fit?**

   Based on my results, the superior model is the 2-hidden-layer model with data size 100000. Its training error is 0.0037, which indicates better generalization, and the lowest validation error, which is 0.0039. It represents strong predictive accuracy.

   It has execution time is 160.87 seconds, which is acceptable for high accuracy.

   The second superior layer is a 2-hidden-layer model with a data size of 10,000. Its execution time is less when compared to the superior model. It also shows high accuracy and efficiency. If we want to attain speed and accuracy balance, this is the best choice.

3. **Next, report the results (for the particular numbers of observations) from applying xgboost (week 11 – provide the relevant results here in a table).**

Comparing the results from XGBoost and deep learning models fit, which model would you say is superior to others? What is the basis for your judgment?

Method used	Size	Predictive accuracy	Time
XGBoost with 5-fold CV (scikit-learn)	**1000**	**0.9500**	**0.267**
XGBoost with 5-fold CV (scikit-learn)	**10,000**	**0.9772**	**0.825**
XGBoost with 5-fold CV (scikit-learn)	**100,000**	**0.9874**	**4.031**

From the results, XG boost is superior. At size of 100000, XG boost achieves 98.74% predictive performance. While MLP has (2layer,100000) has only 0.0039.Coming to time, XG boost takes only 4.031 seconds, while MLP took 160.87 seconds. XG boost uses gradient boosting for better regularization and prevents overfitting more effectively.

From the results, we can also say that, as the data size increases XGBoost achieves highest predictive accuracy and MLP reduces validation error. Large datasets increase processing time, with MLP being significantly slower than XG boost.

Naga Jahnavi Kancharla