Mentor Visit Assessment 18

Mentor: Hao Tian
Profession: Graduate Research Assistant
Location: Heroy Hall at SMU
Date: February 7, 2022
Time: 1 p.m.

Assessment:

For this mentor visit, I updated my mentor on my model progress, discussed machine learning techniques, and met Professor Tao. I was very happy with my latest results: averaging the results of an xgboost model and a graph convolutional network produced a mean absolute error (MAE) of .618. This benchmark result would put my model in first place.

I knew the model had the potential to do even better, but I wasn't sure how. One method I found out about was a more complex ensemble learning process. So far, I had only been taking the mean of two models, but some of the best performing Kaggle competitors combine up to 20 models to fine tune their results, each multiplied by their own weights. This raises the question: how do I determine the optimal model weights? I will need to search for a package that handles this or research a technique to do it efficiently. Before my next visit, I will experiment with more featurizer/model combinations and see if any of them can help me get a lower MAE. I expect this will be challenging because I will be putting together different functions, and packages have not been working well together for me lately, whether due to data format or package version incompatibility. Even though the next couple of weeks will be busy, I hope to have some success.

One thing I wanted to find out more about from my mentor was k-fold cross validation. I learned that the method is useful for smaller datasets, typically containing fewer than ten thousand data points. Essentially, instead of using a 60/20/20 train/validate/test split, cross validation sets aside the test data (to avoid overfitting) and iterates through multiple splits on the combined training and validation data. This means that for 5-fold cross validation, the program starts by setting apart the first fifth of data for validation (optimizing hyperparameters) and trains on the rest. For the next run, it uses the second fifth of data for validation and the rest for training. As a result, all of the data is used for training while avoiding overfitting by keeping the test data independent. Though my dataset is relatively small, my mentor recommended sticking with splits based on random seeds for now to avoid unnecessarily complicating things.

This visit, I also spoke to Dr. Tao in-person for the first time. I was happy to share with him what I was working on, and I was grateful that he suggested I might get access to the SMU

supercomputer cluster. Until now, I have been training and evaluating the models on my laptop because Google Colab is even slower. Sometimes this means I don't get results for half an hour or even several hours. With cloud computing, however, it's likely that all of my models can finish training in under a minute. Dr. Tao even suggested that I could help with his research group's work next year or once I was finished with my final product. That is an exciting prospect for me because of the opportunity to do more hands-on work.

My next mentor visit is just before Research Showcase—until then, I'm looking forward to continuing to work with code and thinking about how to explain my topic to visitors.