COMP5434 - Big Data Computing

Hong Kong Polytechnic University

Department of Applied Mathematics

Bonus 2

Authors:
ZHANG Shimin (ID: 22049616G)
QIU Ruipeng (ID: 22068387G)
QUAN Hongli (ID: 22045145G)
HUANG Qiuyi (ID: 22057918G)
TANG Ka Shun (ID: 22028269G)

Date: April 8, 2023

1 Non-IID data influence

1.1 Experiment

Start by our dataset and based on the IID model we obtained from question3, we firstly turn the data into Non-IID form and conduct experiment on the network with same settings.

1.1.1 Non-IID data generation

First, randomly pick 25% of samples as testing data. Then split the remaining 75% of samples as training data into 4 Non-IID subsets for 4 real estate companies. That is we split samples of each category in training dataset into 4 parts according to dirichlet distribution (with alpha 0.5) for 4 companies. The distribution of new data can be seen from the figure below. Note that we set random seeds on the generation of Non-IID data, so that replication of subsequent experiments is plausible.

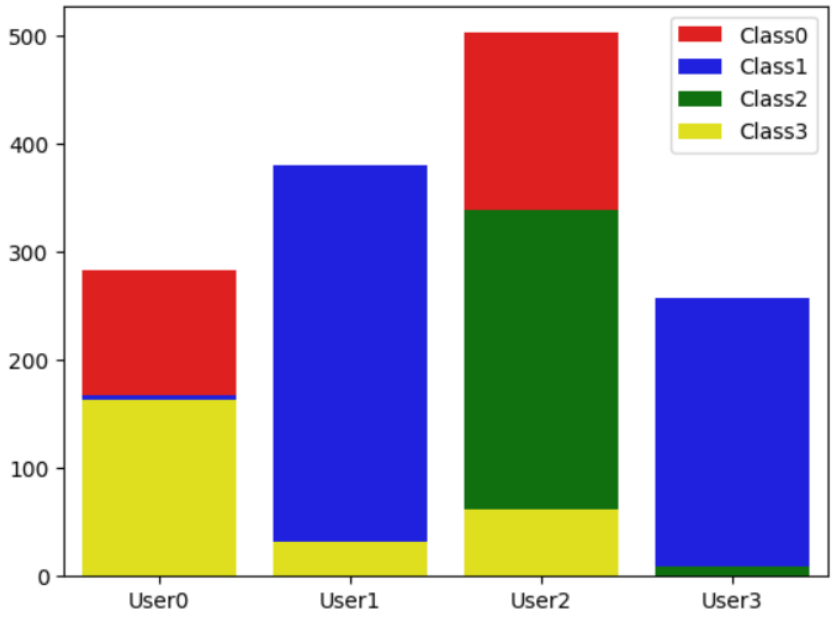

Figure 1: Non-IID data illustration.

1.1.2 Model performance

With Figure 2 we can see that the performance on the test set suffers a significant drop when the trained data becomes Non-IID. One of the explainable parts is the test accuracy. A maximum accuracy of 69.53% was obtained at 80 rounds of training, a

result that is approximately five percent smaller than the accuracy of the federated learning model obtained on the IID data.
For the subsequent training rounds where the accuracy decreases and the test loss increases significantly, we believe that this is caused by the phenomenon of overfitting. This is because the original network was relatively too complex after the data was distributed to four clients. On the other hand, the Non-IID data makes it worse.

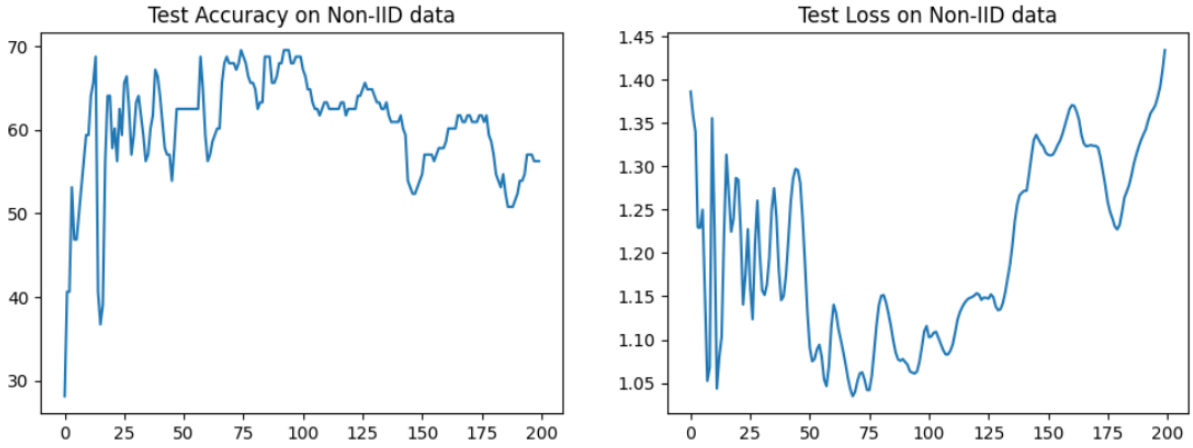

Figure 2: Test accuracy and loss on Non-IID data.

1.2 Analysis on Average Weight Local Update Delta

In fact, it is not sufficient to state directly that model performance is negatively affected by Non-IID in terms of a decrease in test accuracy. Taking this into consideration, we refer to some of the existing literature for an analysis of the impact of Non-IID. One of them is the paper **The Non-IID Data Quagmire of Decentralized Machine Learning** by researchers from Carnegie Mellon University, which provides both theoretical and experimental analyses of Non-IID data quagmire.
The mathematical derivation of the theoretical analysis section is complex and cumbersome, so we do not intend to include that part of the analysis. The authors propose an average weight local update delta, which is calculated as follows. That is, for each epoch, the clients network completes the update of the parameters, and the rate of change of the parameters in the network is recorded after changing the server network parameters through FedAvg. Based on this, it is possible to analyze the role of two different distributions of data for the update of model parameters.

$$\delta_t = \frac{\|\omega_{\mathbf{t}}\| - \|\omega_{\mathbf{t-1}}\|}{\|\omega_{\mathbf{t-1}}\|} \tag{1}$$

We added this analysis to our task by recording the norm of the weights in the three hidden layers in the server network each time. the rate of change of the norm was

calculated for each round and compared to the model obtained with the IID data as the training set. The results are shown in Figure 3.

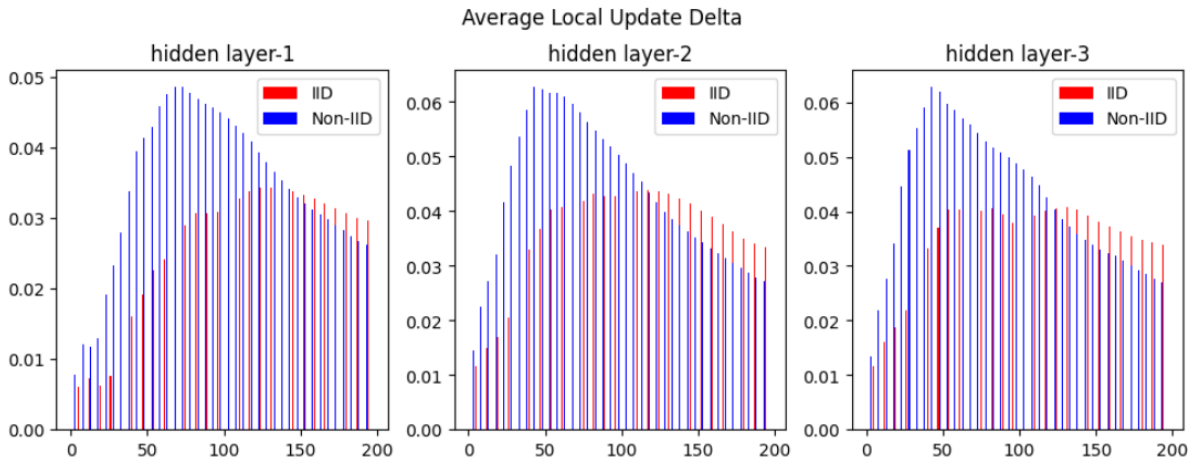

Figure 3: Average weight local update delta for IID and Non-IID scenario.

By comparing the changes in the norm values of the hidden layer weights of the server network in the Non-IID (blue) and IID (red) data scenarios, we can first find that the network trained by the Non-IID data has a significantly larger change in weights than the IID data. At the same time, the IID-trained network parameters are relatively smooth throughout the training process. This suggests that the Non-IID data caused a more unstable training process, which resulted in the network having difficulty converging on the training data and subsequently affecting the model performance. In the last few dozen epochs, we found that the Non-IID-trained network weights norm lower than the IID-trained. We believe this is due to the more severe overfitting caused by Non-IID, resulting in a lower rate of change of the model parameters. In fact, we can observe that the IID-trained network also suffers from overfitting in the later stages of training, but it is not significantly severe, compared to the Non-IID-trained network.
In conclusion, the non-IID data results in a more unstable training process and is more likely to cause overfitting.

2 Potential solutions

Non-IID data quagmire is arguably a pressing problem in the field of federal learning. There have been many discussions and proposed solutions based on the Non-IID problem in the academic community. We have reviewed these literatures and found that the paper **Federated Learning with Non-IID Data** may be applicable to our task.
We assume that part of the iid data can be shared globally. Therefore, we first randomly take part of samples (try the ratio from 0% to 25% to find the best ratio) as global data. Then randomly divide the remaning samples into training set (75% of the remaining

data) and testing set (25% of the remaining data). After that, we divide the training set into 4 Non-IID subsets for 4 clients.
Instead of distribute a randomly initialized global model to clients at the first round, we train a global model on shared IID data and distribute this global model to clients. At each training round, part of shared data (experiments needed to determine a good ratio) are distributed to clients and ask clients to combine both local data and shared data to train the local models (to reduce the earth mover's distance), and then aggregate local models by FedAvg.

3 Conclusion

To explore the impact of the Non-IID data scenario on the federation learning model, we first experimented on the network defined in question3. Approximately five percent of the test accuracy was degraded due to Non-IID. We then refer to the literature for the network based on the average weight local update delta to compare the network weight update magnitude for different data scenarios. It is found that the model under Non-IID data is difficult to converge because of the uneven distribution resulting in a larger update magnitude of the model parameters. Finally, based on the paper **Federated Learning with Non-IID Data**, a potential solution for our house price prediction task is proposed. We believe that the plausibility of the solution can be confirmed by increasing the amount of data.

Python code for Bonus1 please refer to COMP5434_bonus2.ipynb .