Federated Learning – Grade Predictor

Dr. Borcea, Ronak Pasricha

Abstract

Practical Research on Federated Learning creating a regression model to predict grades of students. Public GitHub repository https://github.com/rpasricha45/Federated-Learning-Research-NJIT *has source code from experiment ,and results*

1 What is our Problem

Performing well in academics is highly important and studying resources and behaviors can have a major impact on a student's grade. How can students optimize their studying patterns to result in the best grades? Students can try trial and error, and learn from their past graded exams, and adapt their strategy. However, this method requires students to get a bad grade and then learn from that result. What if there was something that could predict what grade they would get before they take an exam? One potential solution is to create a centralized machine learning regression model that can predict grades based on the student's attributes. This however requires that sensitive student data be held on the centralized server. Sensitive Data such as the student's family's income , lifestyle , to be stored would be a clear violation in privacy . Another major flaw with a Centralized approach is that it does not improve and learn from current students who are using the application. [1]For traditional Mobil machine learning on Android, it uses a pretrained model. Academic classes often change, and so does the studying behavior of each year. By using a pretrained regression model, the results would not be accurate in the future and would be unable to adapt to change.

2 Proposed Solution

The solution is to use Federated Learning to learn from other student's data in the class. Federated learning is a decentralized learning method that ensures privacy and allows for learning to be done from other users devices. All training data stays on the local device and only the updated model is sent to server via encryption. On every round of training, federated learning schedules training when the device is only in idle, and is plugged in. [2]Even when the federated model takes the model updates from the users, federated learning uses Secure Aggregation, which makes the model updates completely anonymous.. Federated training requires Non-IID distributions are samples from different 'users' and are mutually independent. This is perfect for the current problem where student's grades for an upcoming midterm is dependent on the results of the previous midterm, and the data is coming from multiple students .Creating a Federated application has the benefits of using an centralized model, but it also protects the privacy of students information , and has a distributive way of learning which can potentially outperform the centralized model.

[1] https://www.tensorflow.org/lite/guide

[2] *Towards Federated Learning at Scale*

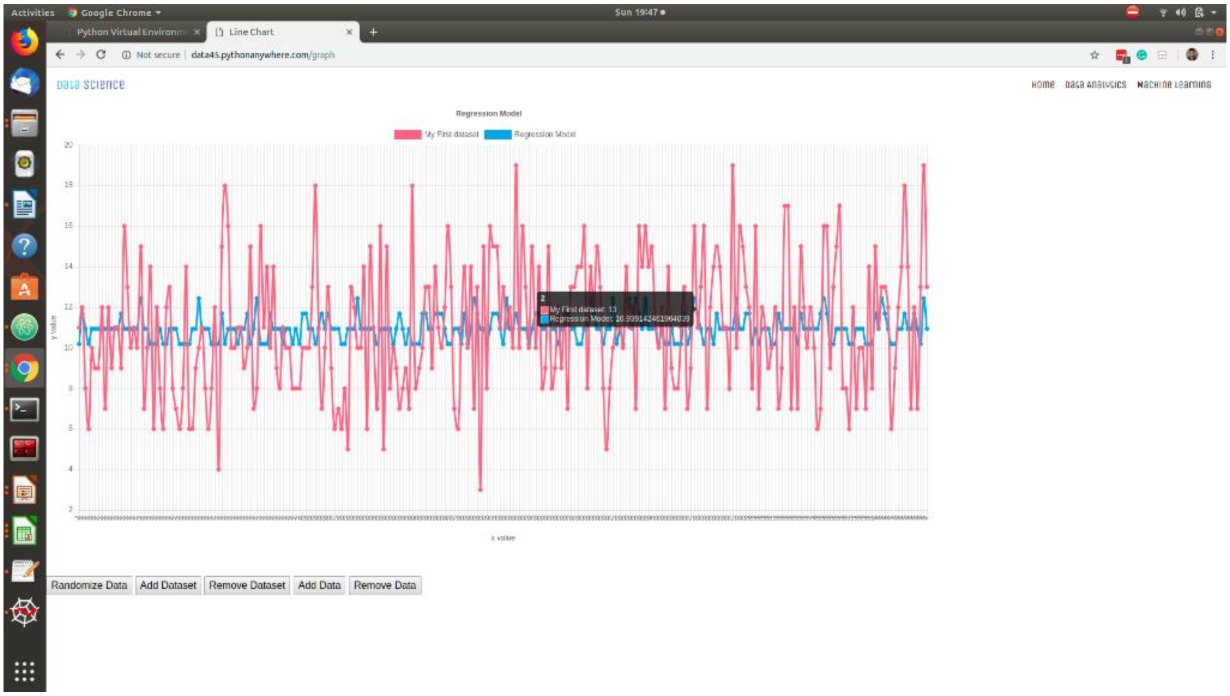


Figure 1: *The above picture shows how close the regression model is to the actual model, the red points are the actual data, and the blue is the regression points Grade 1 and Grade 2*

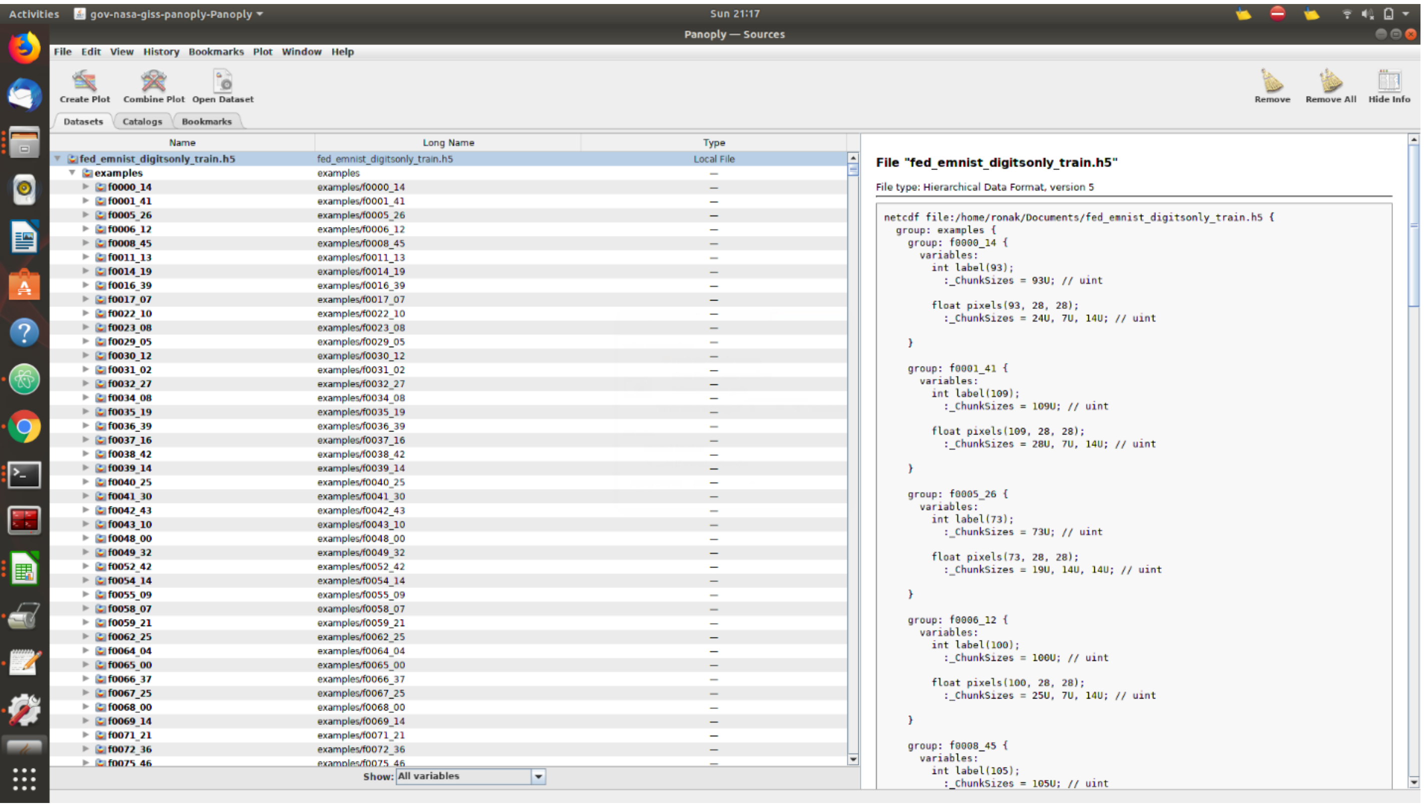
3 Deployment

While most of the training has been done via simulations using federated learning, in deployment it will use multiple Android phones from users to learn from students. During the experiments I collected data from a group of friends, and I asked them after a major assignment or grade what their studying behaviors were. This was clearly not a scalable solution, and since the results were not anonymous there was biases when my sample population would answer questions of how hard they worked. One of the main criticisms of Federated learning is that for small samples, there will not be enough devices being trained per round. To conduct this experiment, I would increase my sample population to all my fellow classmates.

4 About Data Set

To conduct federated learning, it is possible to start with no centralized data and to start only with federated data .I decided to train a centralized data first due to the volume of data entries that I had. [3]The data that I trained with for my experiments were from two high schools in Portugal with 649 entries from different students. There was a total of 33 attributes being measured, with them ranging from age, sex, to whether they had a romantic relationship. Before testing, the results I created a model, to measure

[3] https://archive.ics.uci.edu/ml/datasets/student+performance



Name	Long Name	Type
fed_emnist_digitsonly_train.h5	fed_emnist_digitsonly_train.h5	Local File
examples	examples	—
f0000_14	examples/f0000_14	—
f0001_41	examples/f0001_41	—
f0005_26	examples/f0005_26	—
f0006_12	examples/f0006_12	—
f0008_45	examples/f0008_45	—
f0011_13	examples/f0011_13	—
f0014_19	examples/f0014_19	—
f0016_39	examples/f0016_39	—
f0017_07	examples/f0017_07	—
f0022_10	examples/f0022_10	—
f0023_08	examples/f0023_08	—
f0029_05	examples/f0029_05	—
f0030_12	examples/f0030_12	—
f0031_02	examples/f0031_02	—
f0032_27	examples/f0032_27	—
f0034_08	examples/f0034_08	—
f0035_19	examples/f0035_19	—
f0036_39	examples/f0036_39	—
f0037_16	examples/f0037_16	—
f0038_42	examples/f0038_42	—
f0039_14	examples/f0039_14	—
f0040_25	examples/f0040_25	—
f0041_30	examples/f0041_30	—
f0042_43	examples/f0042_43	—
f0043_10	examples/f0043_10	—
f0048_00	examples/f0048_00	—
f0049_32	examples/f0049_32	—
f0052_42	examples/f0052_42	—
f0054_14	examples/f0054_14	—
f0055_09	examples/f0055_09	—
f0058_07	examples/f0058_07	—
f0059_21	examples/f0059_21	—
f0062_25	examples/f0062_25	—
f0064_04	examples/f0064_04	—
f0065_00	examples/f0065_00	—
f0066_37	examples/f0066_37	—
f0067_25	examples/f0067_25	—
f0068_00	examples/f0068_00	—
f0069_14	examples/f0069_14	—
f0071_21	examples/f0071_21	—
f0072_36	examples/f0072_36	—
f0075_46	examples/f0075_46	—

Show: All variables

File "fed_emnist_digitsonly_train.h5"

File type: Hierarchical Data Format, version 5

```
netcdf file:/home/ronak/Documents/fed_emnist_digitsonly_train.h5 {
  group: examples {
    group: f0000_14 {
      variables:
        int label(93);
          :_ChunkSizes = 93U; // uint

        float pixels(93, 28, 28);
          :_ChunkSizes = 24U, 7U, 14U; // uint

    }

    group: f0001_41 {
      variables:
        int label(109);
          :_ChunkSizes = 109U; // uint

        float pixels(109, 28, 28);
          :_ChunkSizes = 28U, 7U, 14U; // uint

    }

    group: f0005_26 {
      variables:
        int label(73);
          :_ChunkSizes = 73U; // uint

        float pixels(73, 28, 28);
          :_ChunkSizes = 19U, 14U, 14U; // uint

    }

    group: f0006_12 {
      variables:
        int label(100);
          :_ChunkSizes = 100U; // uint

        float pixels(100, 28, 28);
          :_ChunkSizes = 25U, 7U, 14U; // uint

    }

    group: f0008_45 {
      variables:
        int label(105);
          :_ChunkSizes = 105U; // uint
```

Figure 2 *H5 file structure*

the r square score. [4]The r2 score indicates how close a model is fitted to the regression line. The r2 score that I got for the data was 0.68. The r2 metric ranges from 1- 0, with 1 indicating a strong correlation and 0 indicating no correlation. With an r2 score of 0.68, I wanted to narrow down what attributes cause a grade to occur. After more research, there was a stronger correlation between the variables of grade 1, and grade 2. [5]I tested the regression of the line using the project that I created a few summers ago a python regression website that shows the model's performance with a CSV file. While the data did not show the best correlation with the r2 score, due to the lack of data available, and that with the more federated student over time the model will improve I decided to use this data set for the federated experiments. The centralized data main purpose was to act as a starting point, for the federated model, because Since the data was from high schools in Portugal, there may be significant differences from the student's behavior at NJIT.

5 **Preparation of Federated Experiment**

Unlike a traditional centralized model, it is required to create federated data. Federated data is data separated by individual users. For federated experiments, I used the Tensor Flow federated library, to conduct the experiments. For all of Google's examples used to preloaded federated data sets. However, to represent different users it is necessary to use an H5 file format.[6] H5 file format shows data in a

[4] https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-

[5] http://data45.pythonanywhere.com/

[6] https://www.h5py.org/

hierarchical format which makes it a perfect fit for creating federated users.

Figure 5, shows the H5 file for the Image recognition tutorial example from TensorFlow federated. There is a unique client id for each user along with the x and y data points. Tensor Flow federated has a higher-level API that allows developers to wrap their existing deep learning Keras model, to create a federated model, or have the option to manually configure their model. When a developer implements a federated model from their existing model, it is forming a Federated Plan for the model. Every federated model has a Federated Plan, which is a detailed instruction on how to train the data. Each Federated Plan has two-part one for the device and one for the server. The federated plan for the server contains aggregation instructions. For my federated experiments, I elected to use a Dense layer regression model, for simplicity. In a real deployment scenario only a fraction of the total sample of users would be available for training. however for local testing, the developer manually determines the clients that will be tested.

6 Results

Running the federated results, I hypothesized that it would yield, the same results, in the initial simulation rounds and after putting more real users the federated model would outperform the centralized model. The metric of performance that I used was the mean square error when testing. Mean square error measure the squared result of the actual y value and the estimated y value. When running the test, the mean square error was 127. This indicated that the federated model was not performing and was unable to predict an accurate grade.

7 Results Analysis

Since I did not get a very high initial r2 score of 0.68 for my centralized data, I aimed to try to make my existing deep learning model perform better with the data. I modified the preexisting model by including more layers and increasing the learning rate of the model. However even after I modified the model the results were only marginally better. To make sure that I was not giving my federated model bad data I needed to test with better-correlated data to see if it was my model that was causing the issue. So, I found data that had an r squared value of close to 1 this data showed a correlation between years worked and the salary that the employee. received. When running this dataset on the federated model I got a mean square of 121. Once again that the mean square was not a good result for the federated learning model and this experiment that I conducted eliminated the possibility of a low correlated data set. The next thing that I did was I investigated the federated repository where they had a regression example by using the lower it lower-level API. In this example that the model was much more detailed, and it and it was more specific for a regression type of problem. However, in this example, they did not test it using federated data or run any federated tests to it. So, I tried to manually implement all the federated model but was unable to calculate the right metrics for the model. The current performance metric that I used for my manual implementation was accuracy which will not work for a regression problem. I am currently working on changing the metric to get a valid result. While implementing the federated model, by scratch, requires more knowledge and work, for specific problems it does offer a better method to get better results.

8 Learning Experience

Not all experiments yield favorable results, but they do offer valuable knowledge. During this semester I had the opportunity to learn one of the newest topics within Deep Learning. Federated Learning is a very new topic, and amongst undergraduates and even master’s students at NJIT are not familiar with Federated Learning. From this research not only am I one of the few students who know working with federated Learning, but there are also very few developers who know federated learning. While in undergraduate studies most computer science students copy from stack overflow, from this research I can contribute on stack overflow and help millions of other developers. Another valuable skill I learned was dealing with large amounts of source code. In the industry code repository often contain hundreds of thousands of lines of code, and new software engineers are often overwhelmed and are not comfortable understanding with and working with code is such a volume. During this research not only did I look at large codebases, but I need to dive deeper into the federated source code. Since there were little resources about the implementation of federated learning, the best way to solve a problem with the code, was to look and understand the code within the federated library. Going deeper into a problem is what makes Computer Science unique, as it is not focused on just implementing technology and knowing how to use libraries. True Computer Scientist not only know how to use code libraries but can help design them. In addition, I strengthened my knowledge of Machine Learning and aspects of data science, such as different regression models, and preprocessing methods.

9 Future Work

I have invested a large sum of time researching federated learning and overcoming many of the barriers of entries that many faces when entering federated learning. For my current problem with implementing the model using the lower-level API, I know that I am close to finding a solution, as the main thing that I need to change is the performance metric from accuracy to a loss function to fit my problem. In the upcoming months, I plan to fix and try to run the federated simulation, gather more results. While Federated is a new topic in Deep learning , it will continue to grow in popularity in time.

Acknowledgments

I would like thank Dr. Borcea for being a great mentor and giving me this wonderful opportunity! In addition, I would to thank Xiaoping Jeng for help conducting Federated experiments.
.

Works Cited

"Federated Learning: Collaborative Machine Learning without Centralized Training Data." 2017. *Google AI Blog*. April 6. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

"HDF5 For Python." 2019. *HDF5 For Python*. Accessed December 16. https://www.h5py.org/.

Rieuf, Emmanuelle. 2019. "How To Interpret R-Squared and Goodness-of-Fit in Regression Analysis." *Data Science Central*. Accessed December 16. https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the.

2019. *UCI Machine Learning Repository: Student Performance Data Set*. Accessed December 16. https://archive.ics.uci.edu/ml/datasets/Student Performance.