COMPUTER SCIENCE AND DATA ANALYTICS

Course: Guided Research I

Report 3

Title: Federated Machine Learning Implementation on Image Classification

Student: Ali Asgarov

Instructors: Prof. Dr. Stephen Kaisler, Assoc.Prof Jamaladdin Hasanov

Measurement and Scale

In my project, I aim to explore federated learning for image classification while ensuring privacy. The main objective is to develop a federated learning algorithm specifically for image classification tasks and evaluate its performance in comparison to centralized training approaches. In order to conduct a quantitative research analysis, it is crucial to measure the variation in the dependent variable (Y) based on the variation in the independent variables (Xi). Without this ability, conducting any significant testing would not be possible [1]. Therefore, it is essential to determine the meaningfulness of specific variables within my project context.

To accomplish this, I will/am/followed these steps:

Defined the dependent variable (Y): In my project, the dependent variable will be the classification metrics (accuracy, precision, Recall, etc) for image classification. These metrics will help me assess the performance of the federated learning algorithm.

Identified the independent variables (Xi): These are the variables that may impact the performance of the federated learning algorithm. In the context of image classification, some potential independent variables could include the number of participating nodes (devices), the number of training rounds, the learning rate, the size and diversity of local datasets.

Designing experiments: I create a series of experiments where I systematically vary the independent variables while keeping other factors constant. For example, I will conduct experiments with different numbers of participating devices, varying from 1 to 10, and record the corresponding performance of the federated learning algorithm for each case.

Collecting data: By executing the experiments, I will collect data on the dependent variable (Y) for each combination of independent variables. This will involve recording the metrics achieved by the federated learning algorithm in each experiment.

Analyzing the data: Once the data is collected, I will analyze it to understand the relationship between the independent variables and the dependent variable. Statistical techniques, such as regression analysis, will be employed to quantify the impact of each independent variable on the dependent variable.

Interpreting the results: Through the analysis, I will be able to determine the meaningfulness of specific variables and ascertain their significant effects on the performance of the federated learning algorithm. Variables showing a strong correlation or a statistically significant impact on the dependent variable will be considered meaningful and will be subject to further investigation.

By following these steps, I will be able to measure the variation in the dependent variable (Y) based on the variation in the independent variables (Xi). This analysis will enable me to determine the meaningfulness of certain variables in my federated learning project. The measurement I am conducting in my project can be classified as ratio scale. It is appropriate for my project because it has a meaningful zero point (e.g., zero accuracy or error rate) and allows for comparisons of magnitudes between different experiments. I can perform mathematical operations on the measured values, which helps in analyzing the data and drawing meaningful conclusions about the performance of my federated learning algorithm.

Statistical Analysis

Here I would like to start with **DMAIC** cycle [3]. In my research project:

Define:

The problem is improving the accuracy of image classification while preserving privacy through federated learning. I have identified the scope of the project, including the specific image classification task, the dataset, and constraints or limitations.

Measure

I have already described above the how to collect data on the current performance of my federated learning algorithm, which includes the accuracy or error rates achieved in image classification tasks.

Analyze

During the *Analyze* phase, we'll need to draw conclusion based on the given sample data. The main problem is to understand the parameters of the population distribution according to distribution of the sample data [3].

Here I will be doing hypothesis testing. First one could be giving NULL hypothesis H0 which says that there are not any relationships between number of trained devices and averaged accuracy. The second NULL hypothesis will be there are not any relationships between number of training rounds and central model's classification metrics.

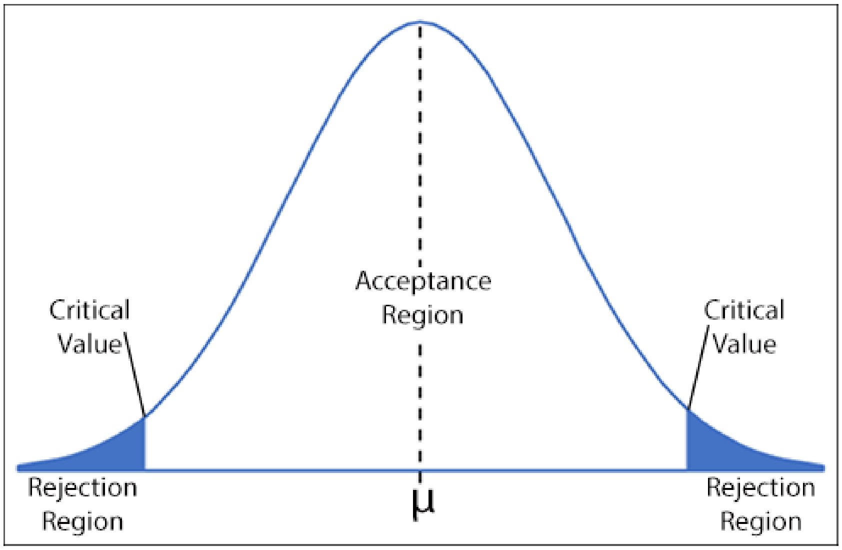

Figure 1 Hypothesis Testing

The level of significance will be between 5-10 %. And mostly will be conducting two-tailed tests.

Here correlation analysis is very crucial, since based on the correlation visualization I will be able to understand which independent factors are affecting my end results. Here as have been noted in lecture, correlation needs to taken into account that is not causation. I will be doing correlation analysis as well while analyzing the collected data.

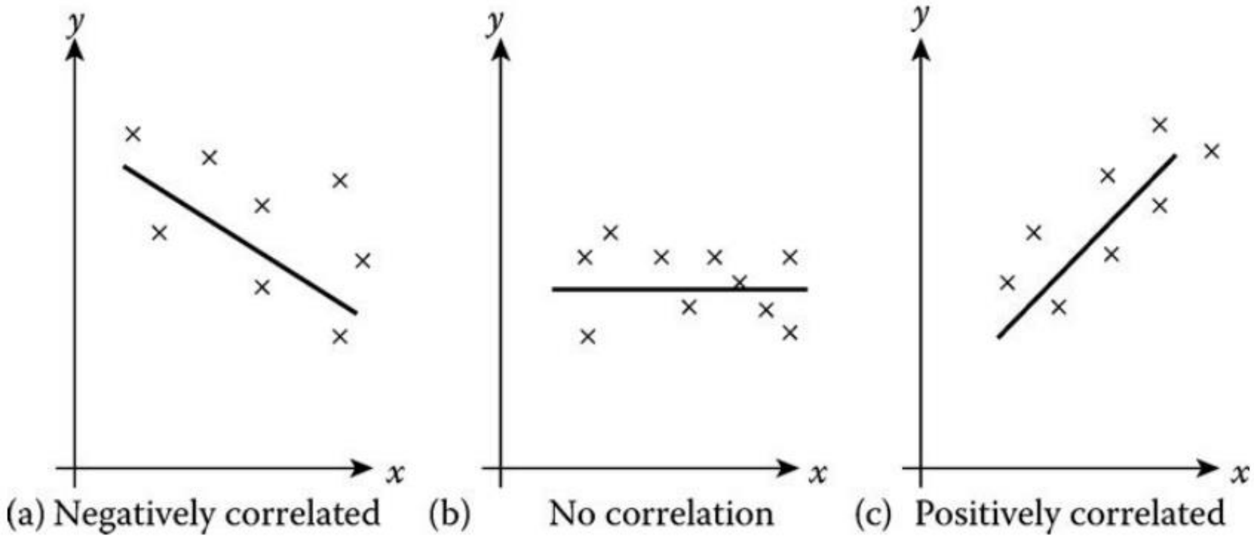

Figure 2 Correlation

Improve

Based on the analysis results the further actions will be taken, for example of the number of rounds are affecting in worse/better direction to the classification metrics, will be increasing/decreasing the rounds.

Control

Implementing processes for ongoing monitoring, tracking, and reporting of accuracy and privacy metrics.

I plan also to play with X-bar (X̄) and R-charts as part of statistical process control. These charts will help me to monitor the performance of my federated learning algorithm over time. By collecting data on accuracy or error rates, I can calculate the average (X-bar) and range (R) values. Plotting these values on the control charts allows me to visualize any variations or trends. If points go beyond the control limits, it indicates potential issues that require investigation. I can then take appropriate actions to improve the algorithm's performance.

Data Visualization.

First, have checked that whether we have the same number of samples for each one of our classes.

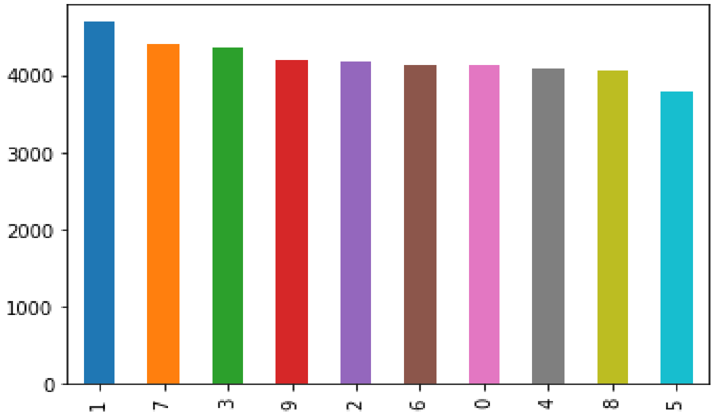

Figure 3 Data Distribution

The biggest class is 1 with about 4500 samples in our training data. The smallest one is 5 with about 3800 samples.

Visual Storytelling

The data set does not contain an equal number of each label. In order to distribute the data to the nodes as IID, an equal number of them must be taken. The function will be written which will group them as much as the amount given from each label and shuffles the order within itself.

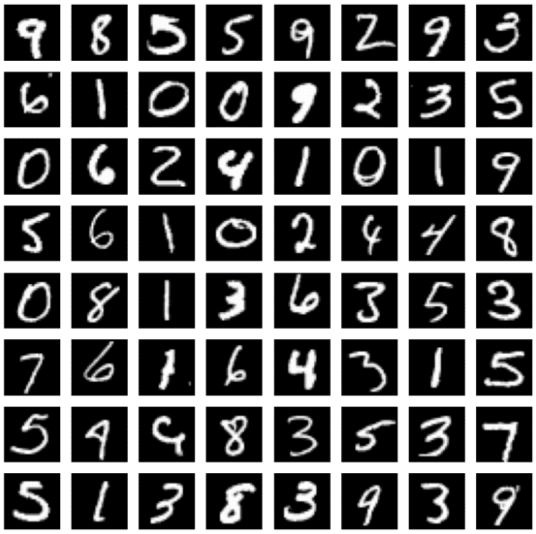

Figure 4 Sample Data

References

[1] CSCI6917 Lecture 8. Measurement and scale

[2] CSCI6917 Lecture 9. Data Visualization

[3] CSCI6917 Lecture 10. Statistical Analysis