BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

SECOND SEMESTER 2024-2025

BITS F464 – Machine Learning

Assignment #2

Weightage:15% (30 Marks)

Due Date: 27/04/2025

Federated Learning Simulation

Machine/Deep learning is at the forefront of technology and is solving problems related to practically all disciplines. A machine learning solution typically requires data from various sources/sites to be uploaded to a central server or a cloud where machine learning algorithms try to find patterns in the data. There are two major problems with this form of machine learning, called "Centralized Machine Learning" in literature. Firstly, the data needs to be communicated to the central server. In this era of Big Data, it can become a bottleneck. Second, and most important is the fact that your privacy is compromised when your data leaves your device, machine, or organization. Most of us are not comfortable with sharing our data with anyone, for any purpose. Privacy concerns become more prominent in domains like healthcare, e-business, finance etc. Federated Learning (FL) solves two major problems: reduces communication costs and preserves privacy of data. What you need to do as part of the assignment:

Build a ML/DL model for Animal image classification problem - using both Centralized Machine Learning and Federated Learning.

Steps involved are (see figure 1):

1. Simulate a thousand clients (mobiles)
2. Client selection is done based on battery power available on mobile device, type of communication (strength, free/paid), & computing capabilities
3. Each client has some labelled data and receives new data after every FL round
4. FL server has some data which is used to build a skeleton model
5. Skeleton model is sent to each client to begin with
6. Each client uses its data to update the model
7. Model updates are sent to the FL server (only the model parameters are communicated – for example, if you are using Linear SVM for a 2D problem, only the parameters of the line are communicated)
8. FL server updates the skeleton model with updates sent by each client as and when they arrive
9. Carry out multiple rounds of updates takes place till a desired accuracy level is reached

You need to compare results of Centralized Machine Learning and Federated Learning in terms of communication costs and accuracy.

Some resources for FL are uploaded on NALANDA. In Friday's (24th Nov.) class, Federated Learning will be discussed in detail.

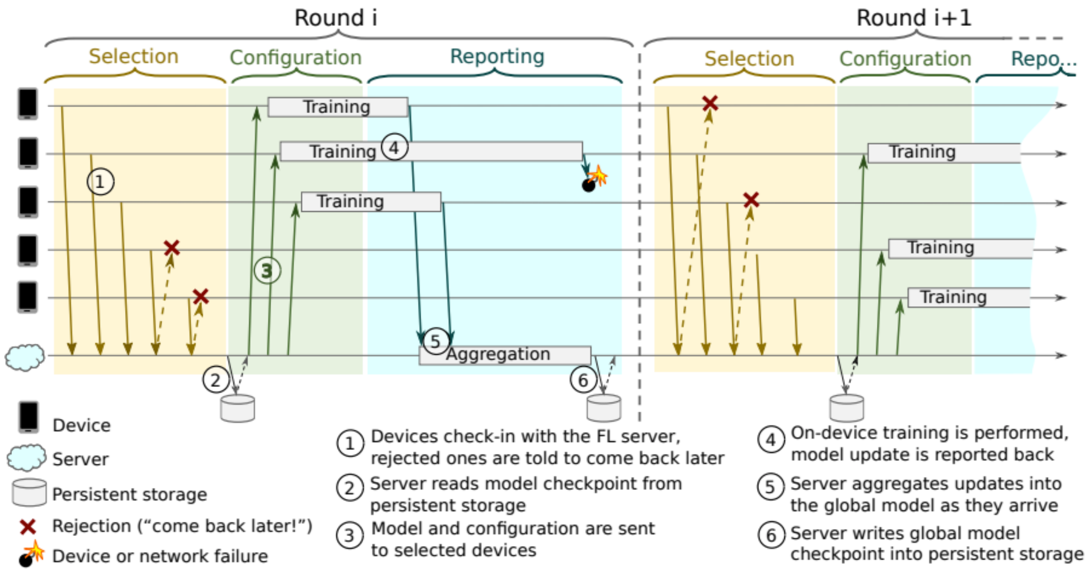
Round i
Round i+1
Selection
Configuration
Reporting
Selection
Configuration
Repo...
Training
Training
Training
Training
Training
Training
Aggregation
1
2
3
4
5
6
Device
Server
Persistent storage
Rejection ("come back later!")
Device or network failure
1 Devices check-in with the FL server, rejected ones are told to come back later
2 Server reads model checkpoint from persistent storage
3 Model and configuration are sent to selected devices
4 On-device training is performed, model update is reported back
5 Server aggregates updates into the global model as they arrive
6 Server writes global model checkpoint into persistent storage

Figure 1: Federated Learning Protocol

Figure source: Google Paper 2019 FL@Scale by Bonawitz et al

What you need to submit?
Submit a report in which you give an interoduction to FL. You also need to give details of the data used and the results obtained. You also need to submit code files! Each group will submit everything in a single ZIP file with name – GroupXX, where XX is your group number. Retain the same group as Assignment #1.

Grading: We will evaluate your submitted files and call you for a viva! You will be evaluated mainly based on what you have understood and not based on what you have submitted. Mere submitting the assignment (and not appearing for the viva) does not entitle you to any marks. All members of the group need to be present for the viva and there will be differential marking.

Navneet Goyal