Peer Grading Rules for CS-E4740 Federated Learning

Grading weights for student project
• Submission: 80%
• Review: 20%

Category 1. Title and Introduction
Q1.1 Is the title suitable? The title should (1) summarize the content of the project in a few words, (2) capture the reader's attention and (3) be specific.Some examples of good titles: "Comparing Logistic Regression and naive Bayes' classifier in spam detection"; "Using Support Vector Machines for analyses of histological samples and cancer prediction"
Some examples of bad titles: "Spam filtering with extras"; "CS-C3240 ML project, final report"
• 1p – Good title
• 0p – Bad title

Q1.2. Does the introduction clearly discuss the application domain?
Some examples of application domains are (1) medical diagnosis (classifying lung images into “Covid-19” vs. “No Covid-19”); (2) cross-country skiing (predict maximum daytime temperature to choose right ski wax).
• 1p – Yes, the application domain is very clearly explained.
• 0p – No discussion or explanation

Q1.3. Does the introduction section give a concise summary or overview of the report (E.g., “Section 2 discusses .., Section 3 then ... and in Section 4 ...”)
• 1p – Yes
• 0p – No

Category 2. Problem Definition.
Q2.1 Are the local datasets (data points, features, labels) clearly explained? Are the types of data also explicitly stated (e.g., integers, binary categories etc.)?
• 1p – Yes
• 0p – No, something is missing.

Q2.2 Does the report clearly describe the edges and their weights in the empirical graph?
• 1p – Yes
• 0p – No, something is missing.

Category 3. Methods

Q3.1 Does the report clearly state where the local datasets were collected from, the number of data points in each local dataset and give a brief description of the dataset?
• 1p – Yes, it is clearly stated where the dataset was obtained, and the description gives me a general understanding of the dataset.
• 0p – No, the source and the dataset are either not described at all or something is missing.

Q3.2 Does the report explain the process of feature selection? Note that theoretical justifications are not necessary, but instead we focus on the process of how the features were selected. It could be based on data visualization, domain knowledge and other strategies.
• 1p – Yes, I fully understand the process of how the final features were chosen.
• 0p – No, it is not mentioned at all OR it is still unclear to me how the features were chosen.

Q3.3 Does the report discuss at least two different FL methods that are based on GTV minimization? FL methods based on GTV Minimization differ in their design choices for local models, local loss functions, empirical graphs and optimization methods used to solve GTV Min.
• 2p – Yes, at least two FL methods are discussed.
• 0p – No.

Q3.4 Does the report clearly state and motivate the choice of local models (hypothesis spaces) for each local dataset (node in the empirical graph)? For example, “Linear models are used since we consider them interpretable and computationally attractive”
• 2p – The choice of local models is well described and motivated.
• 1p – The description or motivation is lacking in some aspects (explain!)
• 0p – The choice of local models is not explained or motivated at all.

Q3.5 Does the report clearly specify, for each local model, the loss function used to train and evaluate the model, i.e., to assess the quality of a hypothesis? For example, “For nodes 1 … 5, we use the logistic loss to train the linear model as these local datasets involve binary labels. For nodes 6 … 10 of the empirical graph, we used squared error loss to train their linear models as these nodes carry data points with numeric labels”; Note that it might be useful to use a different loss function for learning a hypothesis (e.g., logistic loss) than for computing the validation error (e.g., “accuracy” as the average 0/1 loss).
• 2p – The choice for the local loss functions is clearly described and motivated. .
• 1p – The description or motivation could have been more detailed.
• 0p – There is no description or motivation for the choice of local loss functions.

Q3.6 Does the report clearly specify and motivate how the variation of local models is measured. Examples for variation measures include the norm of the difference between local model parameters ([1, Sec. 7]) or the squared difference of their predictions on a common test set (see)
• 2p – The measure of local model variation is clearly described and motivated. .
• 1p – The description or motivation could have been more detailed.
• 0p – There is no description or motivation for the choice of variation measures.

Q3.7 Does the report explicitly discuss how the training and validation sets are constructed? Some examples are (1) using a single split into a training and validation set, (2) k –fold cross validation, etc. [2, Sec. 6.2]
• 2p – The construction of training and validation sets are discussed very clearly
• 1p – The construction of train and val sets could be explained better.
• 0p – The construction of training and validation sets is not discussed at all.

Category 4. Results and Conclusion

Q4.1 Does the report clearly (1) state and (2) compare the training and validation errors obtained for each FL method considered in the report, and thereby (3) decide which is the final chosen method? See [2, Sec. 6.2-6.3] for comparison between different methods.
• 2p – Yes, all aspects are well-discussed.
• 1p – Partially
• 0p – Almost no discussion.

Q4.2 Does the report (1) explain how the test set is constructed and (2) clearly state a test error of the final chosen ML method?A test set should consist of data points that have neither been used to train the local models (training set), nor for choosing between different local models (validation set). The test error is the average loss incurred on a test set.
• 2p – Yes,
• 1p – Somewhat
• 0p – Not at all.

Q4.3 Rate the quality of the discussion of the obtained results ("conclusion"). The conclusion should (1) briefly summarize the report and interpret the results; (2) discuss if the obtained results seem to be optimal or if there is room for improvement; (3) speculate about future directions on how to further improve the FL method. A brief example: "The training error at some nodes was much smaller than the validation error which hints at overfitting. (...) As promising directions for future work, we consider collecting more training data." Examples of other strategies to improve the FL method: use more/different features of data points, use different models, use a different empirical graph or use a different loss function for training, etc.
• 4p – The performance (average loss) of FL methods are discussed appropriately
• 1-3p –Results could have been discussed more
• 0p – No discussion at all.

Category 5. Overall Criteria.
Q5..1 Does the report follow the required outline as in the report template (click here).
• 2p – Yes.
• 0p – No.

Q5.2 Rate the use of language and clarity of presentation.
- 2p – Well-structured report using clear and concise language is clear and concise
- 1p – Well-written overall; could be improved in some respects
- 0p – Poor use of language; Lack of clarity.

Q5.3 Can you reproduce the numerical results of the report?
- 2p – Yes.
- 1p - Partially.
- 0p – No.

Q5.4 Does the report contain existing material without clearly indicating the source?
• 1p – No
• -100p – Yes [Please report this by email to cs-e4740 (at) …) ,

Q5.5 (Overall) This part is for any aspects of the report that are not included in other grading criteria. Do you find the project report worth some extra points, e.g., highly original problem formulation or did the excellent command of FL methods?
• 5p – Yes

References

[1] A. Jung, “Lecture Notes for CS-E4740 Federated Learning,”, available at https://github.com/alexjungaalto/FederatedLearning/blob/main/material/FL_LectureNotes.pdf , ,2023

[2] A. Jung, “Machine Learning: The Basics,” Springer, Singapore, 2022. (also available via Aalto library service https://primo.aalto.fi/)