COST INTERACT ML Challenge - PHY & NET

Marco Skocaj, Nicola di Cicco

28th - 30th August 2023

1 Introduction to the Challenge

Welcome to the COST INTERACT 20120 Machine Learning Challenge on PHY and NET layers. This document provides an overview of the competition's details and guidelines to assist you in completing the challenge.

2 Description

In this section, we offer a comprehensive overview of the two challenges. This will include an exploration of their underlying rationale, the available resources, the established baseline, related academic papers, and the employed datasets. Both challenges will be hosted on the Codalab platform, and will foresee two phases, as explained in the following.

2.1 PHY - Direct localization using MIMO CSI measurements

COST INTERACT ML Challenge on PHY layer focuses on ML-based direct indoor localization using MIMO CSI measurements. Indoor localization based on radio fingerprints is an important topic at the intersection between Communications and Networking, and Machine Learning. The dataset utilized for this challenge is the "Ultra-dense indoor Massive MIMO CSI dataset," which is part of the datasets available via the HA1 working group. For a preliminary inquiry, in preparation for the challenge, you can access this dataset from both https://interactca20120.org/wgs/datasets-2/ or https://ieee-dataport.org/open-access/ultra-dense-indoor-mamimo-csi-dataset. Additionally, the original paper in which the dataset was presented is available on IEEE Xplore. It's important to note that the dataset that will be used in the challenge has undergone downsampling from its original version. This downsampled dataset is provided to participants as part of their starting kit. Since the original dataset is publicly available, it has been standardized with concealed mean and variance. This standardization ensures that participants exclusively use the provided dataset for training purposes. In terms of specifics, the dataset used for the challenge encompasses 15,000 samples of Channel State Information (CSI) for training, along with 5,000 samples each for validation and testing. Each CSI measurement is a 64×100 matrix of complex channels and is associated with a ground-truth 2D position.

The objective of the challenge is to train a machine learning model yielding the minimum positioning error on the test set, as elaborated upon in the section "Evaluation Criteria". Participants are provided with a baseline script, implementing the model described in the original paper.

2.2 NET - Calibrated Predictive Quality of Service using on-field KPI measurements

COST INTERACT Machine Learning challenge on NET Layer focuses on Calibrated PQoS using KPI measurements. The scope of this challenge is to perform calibrated probabilistic regression of the Uplink Throughput (Mbps), based on the observation of a set of three distinct radio KPIs (DL SNR(dB), DL Throughput (Mbps), UL SINR (dB)). Calibrated probabilistic regression refers to the ability to infer predictive confidence intervals which align with empirical confidence intervals. In other words, the predicted confidence levels are representative of the true likelihood of observing the target variable within the provided intervals. For more comprehensive information regarding the evaluation metrics, participants are directed to the "Evaluation Criteria" section.

The dataset employed in this challenge has been collected by Huawei in the city of Munich, Germany, and is available at https://interactca20120.org/wgs/datasets-2/ or https://ieee-dataport.org/documents/huawei

The dataset gathers measurements conducted using a prototype 5G standalone system composed of one Huawei's base station and one user terminal in the city of Munich, Germany. Further information and numerical experiments can be found on IEEE Xplore. Similar to the adjustments made for the PHY Challenge, the dataset for this challenge has also undergone modifications compared to its original version. This has been undertaken to guarantee that participants exclusively employ the designated training data for model training, and the provided validation and test data for making predictions. Notably, certain features (such as LATITUDE and LONGITUDE of the measurements) have been excluded to add complexity to the learning process. Additionally, the dataset has been standardized using concealed mean and standard deviation values. Included in their starting kit, participants are provided with a baseline script that employs a shallow neural network trained via negative log-likelihood as an illustrative probabilistic regressor.

2.3 Phases of the Challenge

Both challenges adhere to a two-stage submission process, drawing inspiration from successful models such as Kaggle competitions.

1. **Competition**: Participants should submit their predictions based on the validation set provided in the starting kit to develop and refine models. Submissions are evaluated on a public leaderboard, providing real-time feedback. It's important to note that while this dataset varies from the final test set, both sets can be considered independently and identically distributed (i.i.d.) with respect to the training set.

2. **Final Leaderboard**: Upon the conclusion of the competition window, participants are expected to submit predictive results for the final test set, which is included in the starting kit but remains unassessed throughout the competition. This final submission will be used to evaluate and rank the models with respect to their performance on this unseen data. Rankings and winners will be determined based on the outcomes of this stage.

3 Registration

Participants reading this document have already completed the registration process using the provided Google form. Now, please proceed to the next step by submitting your registration via the CodaLab framework. To register, please create an account (one account per team) on the Codalab website. Once your account is registered, you can access the challenges via the following links:

1. PHY Challenge: *provided via e-mail*

2. NET Challenge: *provided via e-mail*

Kindly ensure to register for the challenge with one unique account per team.

4 Data Download

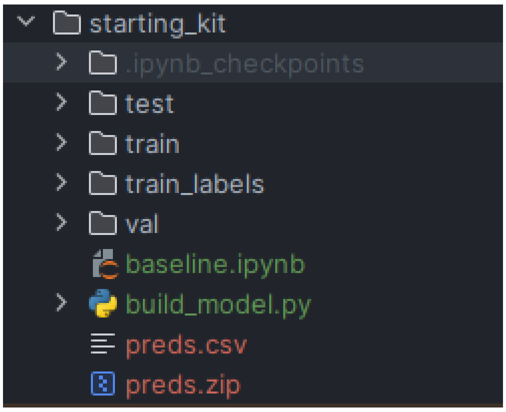

Figure 1: Starting kit - PHY layer

This section will guide you on how to download the necessary datasets for the challenge from the provided sources.

Each participant can download a "starting kit" containing training data (features and labels), validation, and test data (features only). Furthermore, participants will discover a file named "baseline.ipynb" that includes an illustrative baseline example for the challenge, along with essential code to generate a compressed .csv file for submission. The notebook also provides code to assess the scoring mechanism's logic on Codalab. The structure of the starting kit slightly differs for the PHY and NET challenges, and is represented in Fig. 1.

Starting kits with baseline code and data can be downloaded here:

1. PHY Challenge: Link
2. NET Challenge: Link

5 Submission of Entries

Participants willing to submit their models for public evaluation on Codalab must follow the following steps:

1. Utilize your machine learning model, which has been trained on the training set, to make predictions on either the validation set (phase 1) or the test set (phase 2).
2. Be sure to produce a compressed .csv file named "preds.csv" following the example of the baseline.
3. On Codalab, click on "Participate", then "Submit/View results". Add a description and upload your compressed file, then click "Submit". Codalab will take care of processing your file and compute a score based on your predictions.

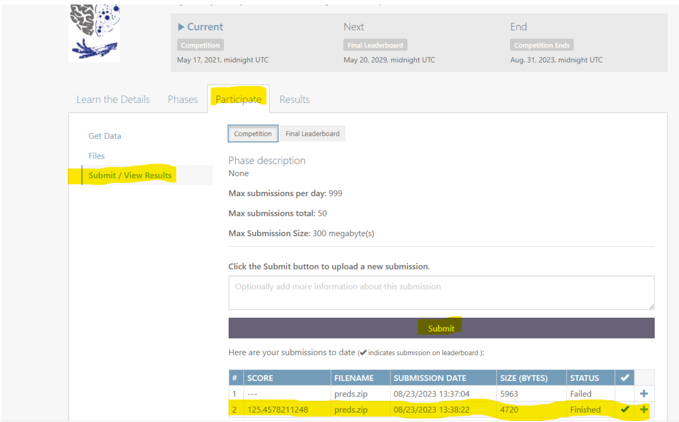

Figure 2: Submission of entries - example

4. Beware, you only have a maximum number of submissions, therefore be sure to evaluate locally your model on a held-out validation set from your available training set before submitting your score on Codalab.
5. Manually submit your score to the leaderboard by clicking on the relative button.
6. When phase 1 is concluded (as soon as the competition ends), participants are required to upload a pred.csv file computed on the test set.

6 Evaluation Criteria

In this section, we outline the criteria by which your submissions will scored, and we provide information on how the final leaderboards will be determined.

6.1 PHY Challenge

We will consider the average **Root Mean Squared Error** (RMSE) for the predicted $\hat{x}$ and $\hat{y}$ coordinates, as follows:

$$\text{RMSE}_x = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{x}_i - x_i)^2}$$
$$\text{RMSE}_y = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{y}_i - y_i)^2}$$
$$\text{RMSE}_{\text{total}} = \frac{\text{RMSE}_x + \text{RMSE}_y}{2}$$

The final rankings will be computed based on the test RMSE scores only.

6.2 NET Challenge

We will score both the **point predictions** and the **confidence intervals**, with the confidence intervals having a higher weight for the final leaderboard.

Point predictions: we will consider the Mean Absolute Error (MAE) as performance metric, computed as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

The rank in the MAE score will award **1 point** for computing the final rankings.

Confidence intervals: we will evaluate probabilistic predictions by measuring the **coverage** (also known as **validity**) and the **sharpness** of the produced confidence intervals. The coverage metric summarizes how "truthful" the confidence intervals are. For instance, a set of 95% confidence intervals is expected to contain approximately 95% of the ground truth points. When producing probabilistic predictions, ensuring the promised coverage is a paramount requirement. The sharpness metric refers to the width of the confidence intervals. For instance, given two sets of 95% confidence intervals delivering equally-good coverage, it is reasonable to prefer the less conservative ones (i.e., the ones with lower average width).

We will consider confidence intervals for 80%, 90% and 95% confidence. For evaluating coverage, we will compute the **calibration error** for each of the confidence levels. Note that, we expect confidence intervals to be **centered around the median**: therefore, e.g., a centered 95% confidence interval will require computing a lower 2.5% quantile and an upper 97.5% quantile. We compute the **Empirical Coverage** (EC) on the evaluation set for the X-th quantile as follows:

$$\text{EC}_X = \frac{|\{y_i \leq \hat{Q}_i^X\}|}{N}$$

Where $\hat{Q}_i^X$ is the predicted X-th quantile for the i-th sample. The empirical coverage represents the fraction of ground-truth points falling within the predicted quantile. We then compute a **Calibration Error** (CE) for each pair of quantiles in the centered confidence intervals (e.g., 2.5% and 97.5% for the centered 95% confidence intervals, etc.), as follows:

$$\text{CE}_X = \left|\text{EC}_X - \frac{X}{100}\right|$$

The total calibration error of the centered confidence interval with confidence X% is then computed as the average of the calibration errors for the lower and upper quantiles High_X and Low_X, respectively:

$$\text{CE}_{X\%} = \frac{\text{CE}_{\text{High}_X} + \text{CE}_{\text{Low}_X}}{2}$$

Finally, the sharpness for a set of X% confidence intervals is defined as the average width of the intervals, as follows:

$$\text{Sharpness}_{X\%} = \frac{1}{N}\sum_{i=1}^{N}|\text{High}_X - \text{Low}_X|$$

The rank in each calibration error and sharpness scores will award **1 point** (**3 points** in total) and **0.25 points** (**0.75 points** in total), respectively, for computing the final rankings. Therefore, ensuring close-to-optimal coverage is significantly more important than producing narrow confidence intervals.

7 Prizes

The winners of each challenge will be awarded two Machine Learning books (tbd) and free participation in the next COST INTERACT meeting.

8 Deadlines

The challenge is set to commence on the 28th, subsequent to an initial online meeting at 9 am, and will extend for a duration of 48 hours, upon which the participants are required to upload their final working solution (phase 2) on CodaLab for the final scoreboard. The official winners will be disclosed during the upcoming meeting in Poznan, following confirmation by the organizers that all rules have been adhered to.