ChAllenge Data
By MathA

ENS | PSL

COLLÈGE DE FRANCE 1530

INSTITUT Louis Bachelier

General documentation

What is a Challenge?

Challenges are **supervised learning problems** (classification, regression, prediction, ranking, …). Given x, the input data, the goal is to predict a value, or an array of values y. This is done by an algorithm which associates to x a prediction $f(x)$. This prediction is learned from training data (a subset of the entire set of data), where the input x and the output y are known. The algorithm is then evaluated on testing data, where y is not available for the participants (they are given by the data provider and stored on the platform but not available to be downloaded).

The data

Every challenge contains at least 4 files:

- **Training data**: the data on which participants will elaborate their algorithm and train their model, to learn the function f
  - x_train: the inputs
  - y_train: the outputs
- **Test data**: the data used for evaluation on unknown data
  - x_test: the inputs
  - y_test: the outputs, this is the data participants shall predict, they do not have access to it.

The score

Given an input x, the error between the prediction $f(x)$ and the true y is evaluated with a score function l which returns a real number $l(f(x), y)$.
The goal is to minimize (or maximize) the average score on the inputs x_i. The calculation of the score depends upon the challenge and is explained in the presentation of each challenge. It may be based on a Mean Square Error, a Cross Entropy, an R2 score, AUROC … A custom loss can be specified in some challenges.

Test set separation and overfitting

The test set is separated into two halves, the public and the private test sets. Every submission by a participant has two scores, one on each set.

- public score: the score computed on the public set. It is provided at each submission.
- private score: the score obtained on the private set. It is provided twice a year.

This separation prevents participants from overfitting the whole test set, as they cannot see how their solution behaves on the entire test set. To limit overfitting on the public set, the number of submissions is **limited to 2 per day** for each participant. Every time a participant submits a solution, the score is computed on the public set. Participants must train an algorithm that generalizes on both the public set and the private set. The private score is given every year, on the 15th of June and the 15th of December.

The leaderboards and their publications

Participants can choose which submission is used for the leaderboards. By default the submission with highest score on the public set is chosen. From submissions selected by each participant we establish the following leaderboards.

Public leaderboard

It ranks the submission scores of participants on the public test set. It is live and **permanently** available.

Private leaderboard

It ranks the submission scores on the private test set. It is **hidden** to participants. The private leaderboard is only revealed twice a year: on **June 15th** and **December 15th**.

About exact date time

Any date dd/mm/yyyy that is mentioned on the website, is understood as dd/mm/yyyy **11:59 pm GMT+00**.
For example, the private leaderboard is revealed every year on Dec 15 11:59 pm GMT+00 (which in France for example would correspond to Dec 16 00:59 am).

Schedule and key dates

Challenges are continuously running but special events are organized every year, to which anyone is welcome to attend.

New challenges

Every year in January, a new season of challenges is added to the platform, on top of the previous seasons. Sessions of presentations of these new challenges are organized at the Collège de France, as part of the course of Professor Stéphane Mallat. See examples here.

Award ceremony

We celebrate the winners of the challenges of the previous year, at the beginning of February. They are the participants who have obtained the best score on the December 15th **private leaderboard**. The prizes are thus awarded at the end of the first year of competition.

Getting started: creating an account

To create an account, click on the “login” tab, then on “register here”. There are three types of accounts:

- participant

- professor
- provider

When creating your account, make sure that your email address is not already used. You cannot have two accounts for the same email.
For professors or providers, your account must be validated by us. For professors, make sure to fill the “institution” field. Your validation will be easier if you use a professional address from your institution.

Supported by

CFM INSIGHT.DATA.CLARITY.

PR[AI]RIE PaRis Artificial Intelligence Research InstitutE

FSMP Fondation Sciences Mathématiques de Paris