ChallengeData By MathA

Season 2024

challengedata.ens.fr

Data science challenges are co-organized by the Data team at ENS and the Institut Louis Bachelier Data Lab in order to put companies in contact with excellent students, engineers and researchers in data science. The challenges proposed by companies or laboratories arise from practical problems they encounter in their activity. They are organized in the form of machine learning competitions, through the platform challengedata.ens.fr (in a Kaggle fashion). More than 10 000 students, researchers and engineers are registered, and we expect to still gain new participants in the future.
This initiative is in the spirit of scientific exchange, such that data and results are shared. The datasets made available by the companies and laboratories should be non-confidential. Algorithmic reports of students and researchers can be accessed by the companies. The web platform supports the data exchange and the automatic scoring of the results of the participants. The evolution of the scores and rankings can be monitored in real time.

Why participate?

These challenges offer companies an access to state-of-the-art algorithms, whatever the proposed problem may be, thanks to the live ranking of the results and the diversity of the participants. Companies can foster links with the best students and professionals participating in their challenge. The link can be established through the presentation of the challenges in the prestigious setting of Collège de France, through online interaction during the challenge, and through the closing ceremony, also at Collège de France. Over the last years, these links have given rise to numerous internships and hirings.

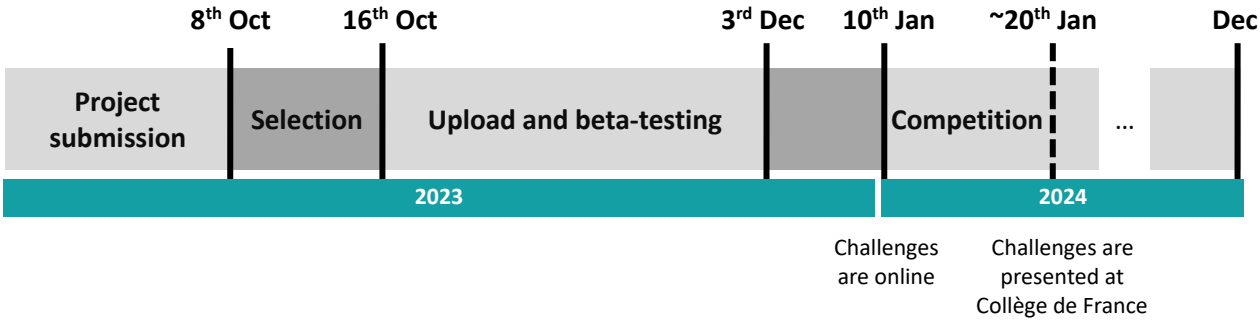
What types of projects?

Each company can propose a supervised learning problem (classification, regression, prediction, ranking...) by providing training and test datasets. All data types are accepted, be them medical signals, images, videos, physical measures, sounds, financial time series, texts, marketing questionnaires, web click data, etc. It is recommended to define a problem which can be easily understood by participants and the solution of which has a true impact, even if it can be difficult and involve sophisticated algorithms.

How to participate?

Interested companies are invited to submit a project as soon as possible by October 8th 2023 by email at challenge.data@ens.fr. The document should be no more than two pages long, in English only, and include the following items:
- the name of the company and a short description of its activities;
- the name and email address of the person responsible for the project;
- the definition of the machine learning project and its interest, specifying if it is a classification, regression, prediction, ranking, etc, task;
- a description of the input variables (*x*) and the output variables (*y*), specifying the number of examples, split between training and test;
- a metric allowing to quantify prediction errors on y;
- a proposition of a benchmark algorithm to obtain first prediction results on the dataset (which will be implemented).

Proposals shall follow the joined template.



An initiative supported by

FSMP Fondation Sciences Mathématiques de Paris · CFM INSIGHT.DATA.CLARITY. · ENS École Normale Supérieure · PSL Research University Paris · Collège de France 1530 · Institut Louis Bachelier

ChAllenge By MathA Data

Challenge proposal

I) Provider description

Quick description of your company, laboratory or organization

II) Contact

Name and email of the people responsible for the project

III) Problem definition

IV) Data description

A full description of the input, noted x, and the output, noted y

Please precise the number of samples and the approximate size of the final files

V) Metric

A function allowing to quantify error between a participant submission and the true output y

VI) Benchmark

A simple algorithm to obtain first prediction result

ChAllenge Data
By MathA

ENS | PSL

COLLÈGE DE FRANCE 1530

INSTITUT Louis Bachelier

Technical specifications

Introduction

This document aims at facilitating the submission of a challenge by precisely specifying the required pieces of information. After a sketch of how the challenge operates, this document details and provides remarks on some of the most important aspects of challenge creation. At the end, one can find an example of specifications.

General organization of a challenge

Challenge Data consists of statistical learning problems such as regression, classification, ranking, etc. In each case, it is about learning a function $f: x \to y$ from a set of observations (pairs (x, y)), called the train set. The quality of the learning is measured from predictions $\hat{f}(x)$ on a set of observations x such that the ground-truth solution $f(x)$ is not given to participants, called the test set. These submitted results are compared automatically based on a predefined metric with the ground-truth on the test set, which is stored on the platform and is not accessible to participants. The obtained score, which is the only thing revealed to participants, allows to classify the different contributions with no doubt on the probity of participants.

Problem definition

One of the most important elements when creating a challenge is the definition of the problem. This consists in statistically learning a function $\boldsymbol{f}$ from **observations**. For binary classification problem, we have for example $\boldsymbol{f}(\boldsymbol{x}) \in \{\boldsymbol{0}, \boldsymbol{1}\}$; for the problem of univariate regression, $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}$.
In other words, it is about constructing a function $\hat{\boldsymbol{f}}$ such that $\hat{\boldsymbol{f}}(\boldsymbol{x_i}) \approx \boldsymbol{f}(\boldsymbol{x_i})$ for all training samples $\boldsymbol{x_i}$, and which generalizes (**has similar performance**) over the test set. In general, the input $\boldsymbol{x}$ is constructed as a Cartesian product of different variables, called features, which can be either continuous-valued (intensity of a pixel in color, spatial position, etc), or discrete-valued (the gender of an individual, etc). It is important to note that the information contained in $\boldsymbol{x}$ needs to be sufficient to obtain a prediction $\hat{f}(x)$.

One obstacle to avoid in the problem construction is that $\boldsymbol{x}$ contains all or part of $\boldsymbol{f}(\boldsymbol{x})$ (without any linear or non-linear transformation), or eventually of another $\boldsymbol{f}(\boldsymbol{z})$. This occurs frequently in time-series. In fact, in order to predict the future values in a time-series $(s_t)_t$, a simple approach is to define the vectors in the intervals of a certain length $x_t = (s_{t-p}, \dots, s_t)$ and to fix $f(x_t) = s_{t+1}$. In this case, $f(x_t)$ is contained in the vector x_{t+1}, without any need to learn anything. Even when assuming that the order of vectors has been changed by a random permutation, whenever $p \geq 1$, it is possible to find out the consecutive (in-time) vector by comparing the last p coordinates with the first p coordinates of the other vectors. Such kind of challenge is ill-posed and needs to be avoided.

Data

The data are divided under two criteria, forming four distinct sets: train or test, input or output. All the training data $(\boldsymbol{x_i}, \boldsymbol{f}(\boldsymbol{x_i}))_i$ and the test inputs $(\boldsymbol{x_k})_k$ are provided to participants,

while the test outputs are only stored on the platform, which permits to evaluate the submissions on separated unknown data.

Every pair $(x, f(x))$ needs to be associated with a unique identification number, called index and written ID. In particular, there should be no intersection between the indices of the training and test sets. As the inputs x and outputs $f(x)$ are split between two files, it is compulsory to guarantee the equality of the IDs between these files. It is recommended to use increasing and consecutive indices across the training and test sets, for instances with 0, . . . , N_{tr} in the train set of size N_{tr} and with $N_{tr}, \ldots, N_{tr} + N_{te} - 1$ in the test set of size N_{te}.

For statistical reasons, the number of test data should not be too small, at least larger than 1000 preferably. In fact, within the platform, the test set is randomly divided into two parts of same size, one called public, the other called private. Only the results obtained from the public partition will be communicated to participants during the competition; one thus needs to ensure that the measure does not fluctuate too much on half of the test set.

In case of numerical, categorical or text data, the CSV format shall be favored for its simplicity and its interoperability. Each file needs to contain a header line indicating the title of each column (possibly some description), the first column containing the index ID. The index column must be identical between corresponding input and output files. Moreover, the header line should be identical for input files on the train and test sets.
Even if the input data is not adapted to the CSV format (for example in case of images), this format shall necessarily be used for the output data.

While missing input data can make the game more interesting, as long as their overall quantity remains reasonable, it is imperative that no output data are missing so as to permit a correct evaluation of the score.

The separator should be commas (,) for technical reasons. An input CSV file shall thus have the following format:

ID, Feature1, Feature2, ...
0, 1.215, 'Joe', ...
1, -785.5, 'Jack', ...
...
987, 0.343, 'William', ...

Finally, four files are to be provided:
Training data :
1. Input in format .csv or .zip: input_training.csv/.zip
2. Output in format .csv: output_training.csv

Test data:
3. Input in file .csv or .zip: input_test.csv/.zip
4. Output in format.csv: output_test.csv

A .zip file containing additional information (such as a Python notebook) can optionally be uploaded.

In total, it is preferable that the data size be smaller than several GBs, so that a participant can be able to run the challenge with a standard computer. In all cases, it is necessary to make the size of the output test set smaller than 10 MBs.

Metric

The metric associates to each submission a numerical score permitting to judge its quality. The simplest example for a classification problem is the classification accuracy (the ratio between the number of entries assigned to the good class and the number of entries of the test set), and the quadratic error for regression problems. As always, a metric is not necessarily a raw performance measure and it can integrate more refined criteria, such as the robustness measured by the AUC curve.

The metric needs to be chosen carefully according to the nature of the data and the scientific and technical objectives. It cannot be modified once the challenge has started.
Most of the standard metrics are already incorporated on the platform. It is nevertheless possible to add some new metrics by providing an appropriate Python script.

In order to be compatible with the automatic and random separation of the test set into a private part and a public part, the metric needs to be separable with respect to the individual observations. More precisely, if $y, \hat{y} \in \mathbb{R}^{N_{te} \times p}$ are the matrices respectively containing the true outputs and the outputs provided by the users for the test set, where each line $\boldsymbol{y_i}$ contains an example and each coordinate a value to predict, then the metric needs to be written as:

$$L(y, \hat{y}) = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} l(\boldsymbol{y_i}, \hat{y}_i)$$

In this way, this formula can be restricted without any ambiguity to the private or the public part.

Calibration

Before opening the challenge to the competition, it is important to calibrate it by testing standard algorithms. The score obtained from a company's submission is called the benchmark on the platform and appears as such in the ranking. This benchmark permits to verify that the challenge does not contain any obvious flaw and to guide students to the type of performance that they should attain to be in the game.
For this benchmark to be visible in the ranking of the challenge, it needs to be submitted from the challenge provider's admin interface. The submission process can be repeated as often as necessary: the last benchmark submission then becomes the public one.

Overfitting

In order to avoid overfitting training data, it is suggested to verify the absence of duplicates in the training dataset. At the level of the platform, the number of submissions is limited to two per participant (or per team when applicable) within 24 hours. AT the end, only the score calculated over the public part of the test set, called public score, is revealed to participants

when they submit. The score computed on the private part of the test set, called private score, is only revealed to participants twice: at an intermediate academic date, and at the end of the challenge.

Challenge example

Problem description

The goal of the challenge is to predict for multiple American stocks which volume will be exchanged during a certain period. Since the American stock market opens at 9:30am and closes at 4pm, the competitors will have access to the transactions between 9:30am and 2pm and need to predict the transactions between 2pm and 4pm.

Data description

The input data contains the exchanged volume (in dollars) of a certain set of stocks and aggregated dates over some periods of 5 minutes. Each line is defined by a unique ID and corresponds to a certain day (defined by “date”) and to a certain stock (defined by "product_id"). The exchanged volumes are the sum of the values exchanged on the market and are labelled with the time of the beginning of the aggregation (from 9:30am to 1:55pm).

The first line of the input file contains the header, and the columns are separated by commas. The three first corresponding columns are:

- the ID: the identification number, it is linked to the ID of the output file,
- the date: related to a certain day, shared by some stocks (for practical reasons, the days are mixed randomly),
- the product_id: the identification number of the stock, it is related to a particular company (the stock #236 corresponds to the same company in the training and test file).

The remaining columns are the volumes exchanged in dollars over the period of 5 minutes. The time labels are provided in military format. For example, the column labelled with 10:00:00 corresponds to the volume exchanged on the market between 10am and 10:05am.

Here is an example of an input file:
ID, date, product_id, 09:30:00, 09:35:00, ..., 13:55:00
0, 1, 236, 121984.4, 644041.2, ..., 38941.0
1, 1, 239, 2594.3, 2587.1, ..., 964.2
2, 1, 254, 96348.1, 73186.9, ..., 65156.9

The training output file contains the output for each ID, which shows the volume exchanged for the same stock and the same day between 2pm and 4pm. The first line of the file contains the header and columns are separated by commas:
ID, TARGET
0, 11887952.0
1, 770918.0
2, 3004214.0

Thanks to the file test_input, participants need to provide a test output file in the same format as the training output file (associating each ID with the predicted volume during 2pm-4pm)

Metric

The metric used in this challenge to rank participants is the average absolute relative error:

$$\frac{1}{N}\sum_{i=1}^{N}\left|1-\frac{\hat{f}(x_i)}{f(x_i)}\right|$$

Supported by