

**Projet Master 2 Informatique**

**Apprentissage, Information et Contenu (AIC)**

**Preparing a starting kit and a challenge bundle**

**Isabelle Guyon**

[**iguyon@lri.fr**](mailto:iguyon@lri.fr)

\*\* Please preferably communicate through[**Piazza**](https://piazza.com/class/jmih4dkp6896ty) **\*\***

**Version 2, Nov. 14, 2018**

**Getting started:**

1. Install on your computer **Anaconda Python 3.6**. If you already have another version of Python or Anaconda on your computer, use virtual environments:

conda update conda

conda create -n python3 python=3.6 anaconda

Then you will be able to use the 3.6 version by activating the virtual environment:

conda activate python3

To get out, use:

conda deactivate

This installation should include python 3.6, the spyder IDE and jupyter-notebook (verify).

1. Another possibility will be to use the docker codalab/codalab-legacy:py3. More on dockers at the end of this document. If you are going to use OTHER libraries than those provided in Anaconda, definitely use dockers!
2. [Clone](https://help.github.com/articles/cloning-a-repository/) the starting kit template: <https://github.com/madclam/m2aic2019/tree/master>. If you are new to Github, take a tutorial. Using Git and Github is essential to your project.
3. **When done creating your starting kit, push it to your own Github repo (your homework will be your Github repo URL).**
4. Each group has an email [groupname@chalearn.org](mailto:groupname@chalearn.org), which has been used to create accounts on:

<https://chalab.lri.fr/> and <https://codalab.lri.fr/>.

The **password** in the same for all accounts: **m2\_aic++**

1. **Starting kit structure:**

A starting kit should imperatively have the following structure, do NOT change it (with exception of the L2RPN group working on reinforcement learning for whom the data is replaced by the simulator):

* **README.md:** Short description of your task and credits.
* **README.ipynb:** Sample code reading data, calling sample code, generating sample submissions.
* **sample\_data/:** A SMALL data sub-sample (could be a subset of the training data, but should NOT include any validation or test examples on which the competitors will be tested).
* **sample\_result\_submission/:** Example prediction files, for the validation and test set.
* **sample\_code\_submission/:** Self-contained, ready-to-submit example code submission, including the predictive mode class model.py and a metadata file (with just a comment).
* **ingestion\_program/:** Program and libraries needed to run code submissions on the challenge platform and generate prediction results using the input\_data (labeled training data and unlabeled validation and test data) and model.py.
* **scoring\_program/:** Program and libraries needed to score the prediction results on the challenge platform using the target values (solution), also called reference\_data. A library of scoring functions to choose from is provided in libscore.py. All you need to do is to put the name of the chosen function in metric.txt. If needed, you can supply the code of your own metric in my\_metric.py. You can also use a scikit-learn metric.

1. **Data preprocessing and formatting:**

This year we have a variety of tasks, depending on the project you chose. Those include **classification**, **regression (survival), matrix factorization**, and **reinforcement learning**.

* **Data representation:** For the first (basic) version of the starting kit,all datasets will have to be transformed/preprocessed into a "**feature representation**". This especially important for text, speech, vision, or any datasets whose raw data consist of images or sequences. For example, for image data, you can use features calculated by a [pre-trained CNN](https://www.quora.com/Where-can-I-download-the-various-different-kinds-of-pre-trained-deep-learning-neural-networks). The Areal and PersoData groups working on computer vision applications will keep a non-preprocessed version of their dataset, for later (we’ll do a second version from raw data for more advanced students).
* **Missing data:** Avoid dataset will a large number of missing values (more than 10%). Represent the missing values with the symbol **NaN.**
* **Binary and categorical variables:** Represent all variables with a **NUMERIC code**. Avoid categorical variables with a large number of values, some of which appear very infrequently. To that end, you may group infrequent categories. Likewise, for binary variables, avoid datasets in which variables are very imbalanced (very sparse data). The target values (solutions) must also be in a numerical format.
* **Number of classes:** Avoid problems with a too large number of classes (>20) and have too few examples per class (less than 1000 in test data).
* **Size dataset:** Do not consider too small datasets of less than 50 000 examples. You must have at least **10 000 test examples** and preferably a large training set and a validation set of the same order of magnitude as the test set. We want enough samples, but the overall preprocessed compressed volume of data should not exceed **300 MB**. Quantizing values to enough precision (but not too much) is recommended, as part of preprocessing. Usually quantizing between 0 and 999 is enough. To further reduce data volume, you may also reduce the number of features by feature selection (this should be the last resort because it is good for the L2 students to struggle with many features). So no limit on the number of features.
* **Data homogeneity:** Avoid data that are non-homogeneous or non-identically-distributed. **Shuffle** the order of your examples randomly.

**Data format:** All pre-processed data must be produced in [**AutoML format**](https://github.com/codalab/chalab/wiki/Help:-Wizard-%E2%80%90-Challenge-%E2%80%90-Data)**:**.

This is imperative because you will all share the same data reader (called by the ingestion program). You will have to split the data into a **“training set”,** a development test set a.k.a.**”validation set”** for the challenge phase 1, and a (final) **“test set”** for the challenge phase 2. You should eventually produce ALL the following files, but, for the first version, produce at least those outlined in boldface:

**DataName\_train.data # This is the training matrix X**

**DataName\_train.solution # This is the training matrix Y (target values)**

**DataName\_valid.data # This is the first test set called “validation set” used during development**

**DataName\_valid.solution # and its target values (hidden to the participants)**

**DataName\_test.data # This is the FINAL test set**

**DataName\_test.solution # and its target values**

DataName\_feat.name

DataName\_label.name

DataName\_feat.type

DataName\_public.info

DataName\_private.info

We ask that you to:

+ Provide all the requested files (even those not in boldface) because this will greatly help the participants and this be useful documentation for future reference.

+ NOT provide “un-split” data in the simpler form: DataName.data and DataName.solution, because **providing your own data-split** is better for several reasons:

* You have more control and you can, for instance, ensure that each class has a balanced mix of examples (in the case of multiple data sources of multiple types of subjects), or, on the contrary, that you have different types of examples in training and test data (for instance if you want to address speaker independent speech recognition, samples from different speakers should be in training and test data).
* You are sure that the examples are well shuffled.
* You are sure that the example order matches your sample result submission.

**Data leakage:** This refers to features that inform inadvertently about the target values: running a feature selection algorithm often reveals such features. Typically, there may be a column with sample IDs highly correlated with the target. There are other forms of data leakage. For example if you forget to shuffle the examples and all the examples of the same class appear next to one another. Check <https://www.cs.umb.edu/~ding/history/470_670_fall_2011/papers/cs670_Tran_PreferredPaper_LeakingInDataMining.pdf> for other horror stories.

**Sample data and challenge data**: You need to prepare 2 datasets in AutoML format:

* **Challenge data:** your actual challenge data. In the challenge platform, it will be split into downloadable **public\_data** or **input\_data** (everything except the DataName\_valid.solution, DataName\_test.solution, and DataName\_private.info) and **reference\_data** not leaving the platform: DataName\_valid.solution for phase 1 and DataName\_test.solution for phase 2.
* **Sample data:** a small data subset carved out the challenge TRAINING data, for practice purposes only (do not compromise real validation or test data).

1. **Sample submissions:**

You must prepare two zip files:

**1) sample\_result\_submission.zip:** A zip file with NO directory structure containing 2 files:

* + **dataName**\_valid.predict
  + **dataName**\_test.predict

corresponding to predictions of the **target values** made by your sample code for the validation set and the test set, one sample per line. For the HADACA project, your participants must predict 2 matrices A and T of dim (N, C) and (C, M). To respect the format, please write matrix A and T’ one after the other to get a file with N+M lines and C columns.

Tentatively, you can put random values or constant values (with the right number of lines), but later you may want to create a “real” sample submission with the sample code provided in sample\_code\_submission. One easy way is to upload a code submission to Codalab, then “Download evaluation output from prediction step”, but you need to remove the metadata file!

**2)** **sample\_code\_submission.zip**: A zip file with NO directory structure containing a Python class model.py. This should be a basic but functional baseline method solving your problem. Using the “ingestion program” (see below), you can generate a sample\_result\_submission from your sample\_code\_submission. This zip file should also contain a metadata file with just a comment. It may or may not contain a pickle with a trained model (maybe you want to create two versions).

1. **Ingestion program:**

A sample ingestion program is provided (it should not need to be modified if you respect the AutoML format and the sunmission format). The ingestion program receives the code submission of the participants on the server and runs it. However, you can also run it locally to produce a result submission: python ingestion\_program/ingestion.py **public\_data** sample\_result\_submission ingestion\_program sample\_code\_submission

where public\_data/ should contain your real challenge data (but not the validation and test target values/solutions). **Remove the pickle in sample\_code\_submission, if any** (may cause an error).

1. **Scoring program:**

A scoring program is provided. It should not need to be modified, however, the scoring metric can be chosen by the organizers (you). Place your code in my\_metric.py, and indicate the name of your scoring function in metric.txt. Instead of writing your own code, you may also select one of the metrics for libscore.py and specify it in metric.txt:

|  |  |  |  |
| --- | --- | --- | --- |
| **Binary classification** | **Multi-class classification** | **Multi-label classification** | **Regression** |
| bac\_binary | bac\_multiclass | bac\_multilabel | abs\_regression |
| auc\_binary |  | auc\_multilabel | r2\_regression |
| pac\_binary | pac\_multiclass | pac\_multilabel |  |
| f1\_binary |  | f1\_multilabel |  |

Test your metric with:

python scoring\_program/score.py **reference\_data** sample\_result\_submission scoring\_output

where reference\_data/ should contain your validation and test target values/solutions: DataName\_valid.solution for phase 1 and DataName\_test.solution for phase 2, and scoring\_output is a directory that will be created.

1. **Jupyter-notebook:**

We provide a sample Jupyter notebook to help the participants getting started. Keep the basic structure of the template but replace the yellow place holders with relevant graphics and explanations:  
jupyter-notebook README.ipynb.

1. **From the starting kit to the challenge bundle:**

The challenge data and starting kit are the basis for the Codalab challenge. Here is your **check list:**

* **LOGO**: Prepare a **logo** (a small square jpg file) that we will refer to as **DataName\_logo.jpg.**
* **DATA:** The challenge **data** should have this structure:

**DataName/ DataName\_train.data**

**DataName\_train.solution**

**DataName\_valid.data**

**DataName\_valid.solution**

**DataName\_test.data**

**DataName\_test.solution**

DataName\_feat.name

DataName\_label.name

DataName\_feat.type

DataName\_public.info

DataName\_private.info

Zip it **with** the top directory structure and call it **DataName\_data.zip.**

**At this stage it is fine to use a small sample dataset as a place holder for debug purposes.**

* **STARTING KIT:** The **starting kit** should have this structure:

README.md.

README.ipynb

sample\_data/

sample\_result\_submission/

sample\_code\_submission/

ingestion\_program/

scoring\_program/

Zip it and call it **DataName\_starting\_kit.zip.**

* **PROBLEM:** The contents of **ingestion\_program/** should be un-changed:

metadata

ingestion.py

data\_converter.py

data\_io.py

data\_manager.py

Zip it without directory structure and call it **DataName\_ingestion.zip.**

* **METRIC:** Make a note of the name of the **metric** you used (e.g. **chosen\_metric**), declared in scoring\_function/metric.txt, if you used one of the metrics provided in libscores.py, OR save the code of your metric in a file **my\_metric.py.**
* **SAMPLE SUBMISSIONS:** You should also have prepared 2 **samples submissions**:

**sample\_code\_submission.zip** [containing an empty metadata file and model.py]

**sample\_result\_submission.zip** [this one should have been generated from the sample code submission using the challenge data (that you will put on the challenge platform) **NOT the sample data, which is just for debug purposes**; it should contain DataName\_valid.predict and DataName\_test.predict].

Upload your challenge to ChaLab

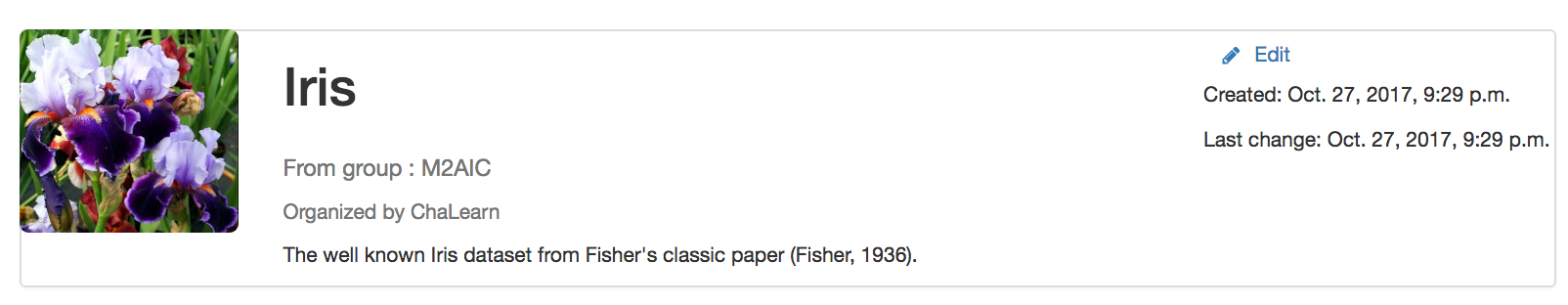
Go to <http://chalab.lri.fr/>. We created accounts with your group name ({vision, friend, ecolo, credit, [biomed}@chalearn.org](mailto:biomed%7d@chalearn.org), please use them. You are part of a class group that can use a pre-filled challenge template. To activate it, go to “Profile” and select “Private group M2AIC”:



The create a “New m2aic challenge”:



On the next page, click “Edit” at the top to replace the logo by **DataName\_logo.jpg** and the challenge information:



If you are using the Iris template, you only have to perform the green steps:

1. **Data:** Upload **DataName\_data.zip**. Note: **NO file larger than 300MB**. If your dataset is >50MB, we recommend to upload only the sample data to ChaLab. Then follow instructions in the next section to upload the large dataset to Codalab directly.
2. **Split:** Your data should be pre-split if you followed the instructions; nothing to do.
3. **Problem**: The program **DataName\_ingestion.zip** should already be there: nothing to do.
4. **Metric:** choose **chosen\_metric** in the list of default metrics or upload **my\_metric.py.**
5. **Protocol:** we impose a 2-phase challenge. Keep the date of the template. The beginning of phase 2 shoud be **APRIL 30, 2018:** nothing to do.
6. **Baseline:** Upload **DataName\_starting\_kit.zip**. No large sample data should be bundled with the starting kits. Files <50 MB please.
7. **Documentation:** Fill up the documentation pages appropriately. **Do not change the terms/rules.**

Then **Package & Publish:** Build your competition bundle and download the zip (challenge bundle): **DataName\_challenge\_bundle.zip**

**Upload the challenge bundle to Codalab**

1. Upload **DataName\_challenge\_bundle.zip** to <https://codalab.lri.fr/>.

To do this, use your account **groupname@chalearn.org**.

1. Test your challenge with your sample submissions: **sample\_code\_submission.zip** and **sample\_result\_submission.zip.**

To do this you will need to go to **Participate>View/Submit results.**

**WARNING: you will get errors if the correct docker is not specified. See page 11.**

**Large datasets**

If you have a very large dataset and only uploaded sample data instead of your real challenge dataset on ChaLab, follow these instructions to upload data directly to Codalab.

1. Separate the target values from the rest and create 3 archives:

* **DataName\_input\_data.zip**, containing:

DataName\_train.data

DataName\_train.solution

DataName\_valid.data

DataName\_test.data

DataName\_feat.name

DataName\_label.name

DataName\_feat.type

DataName\_public.info

* **DataName\_reference\_validation\_data.zip**, containing:

DataName\_valid.solution

DataName\_public.info

DataName\_private.info

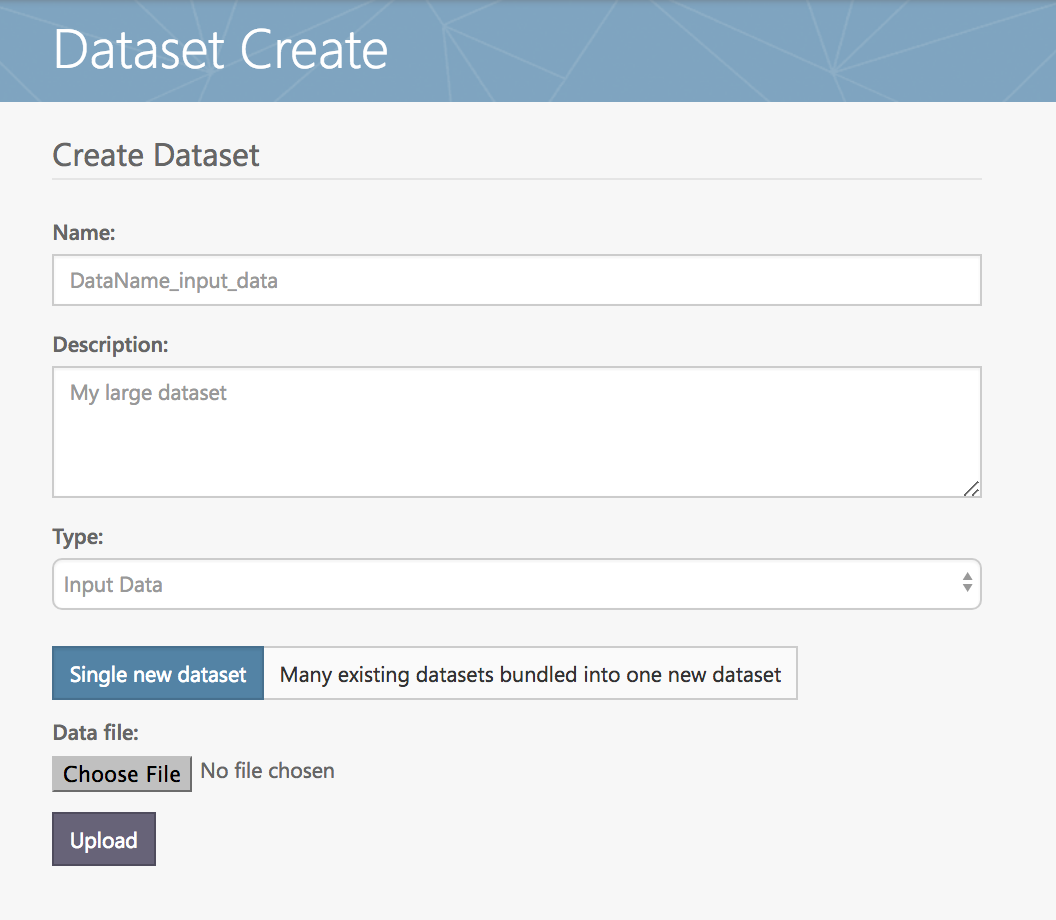
* **DataName\_reference\_test\_data.zip**, containing:

DataName\_test.solution

DataName\_public.info

DataName\_private.info

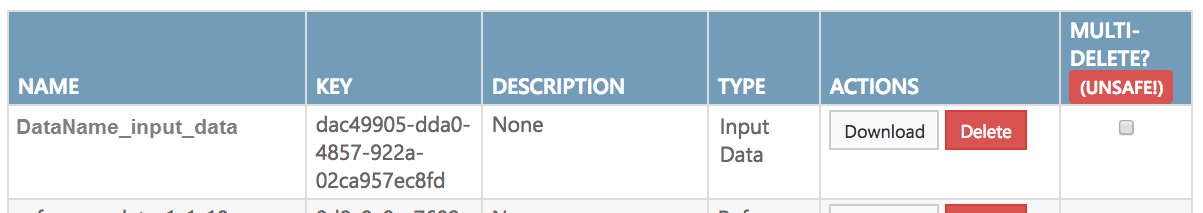
go to **My Competitions>My Datasets>Create Dataset.** Fill out the form:



Upload the file **DataName\_input\_data.zip**. After you upload, you should see your dataset in the data table. The KEY can be used to refer to it from the YAML file, both for public\_data and input\_data.

input\_data: dac49905-dda0-4857-922a-02ca957ec8fd

public\_data: dac49905-dda0-4857-922a-02ca957ec8fd

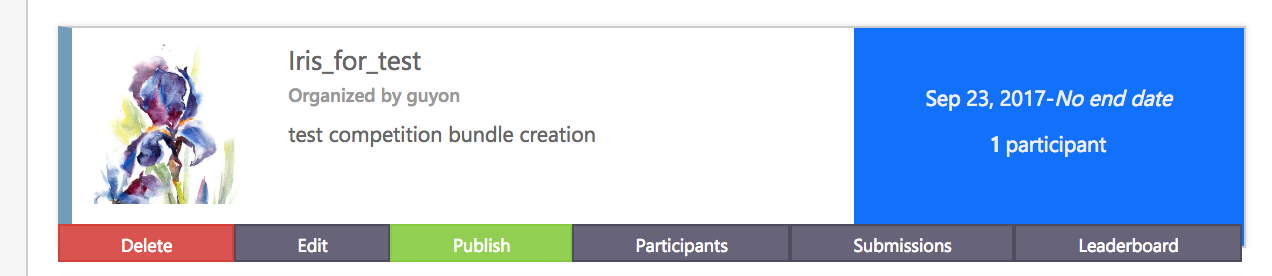


Do the same thing for **DataName\_reference\_data.zip**. IMPORTANT, use the correct data Type **(input\_data or reference\_data)** in the form when you create the dataset. You can then use the key obtained in the YAML file:

reference\_data: 0d8c9c0e-7609-4314-bf1c-3154cdee3462

You need only to specify public\_data in phase 1. Use the same input\_data and reference\_data keys in both phases.

You can also use the editor. In Competitions I’m running, find your competition and click “Edit”:

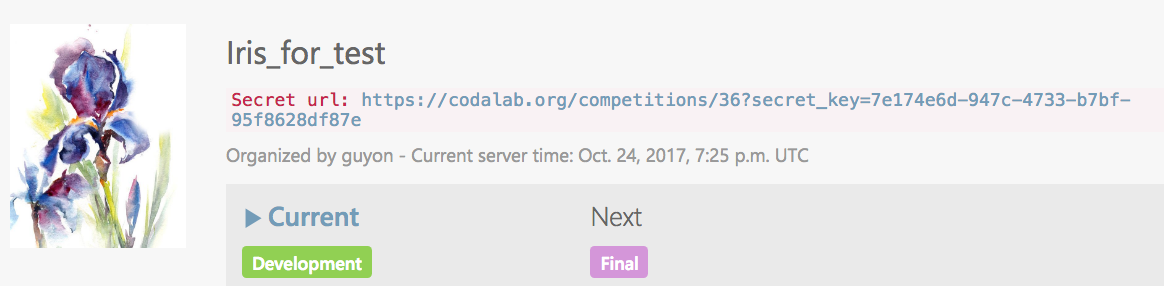


Find the menus for Input Data, Reference Data, and Public Data. Select the right dataset, as don’t forget to **SAVE YOUR CHANGES.**



**Editing your challenge on Codalab**

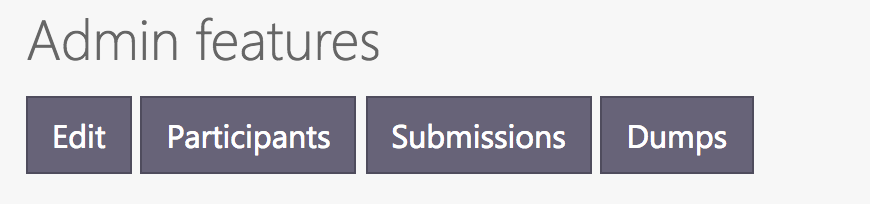
Once your challenge is uploaded to Codalab, your whole team can contribute to improving it. Before you make your challenge public (by clicking “Publish”) you can share the secret URL with others.



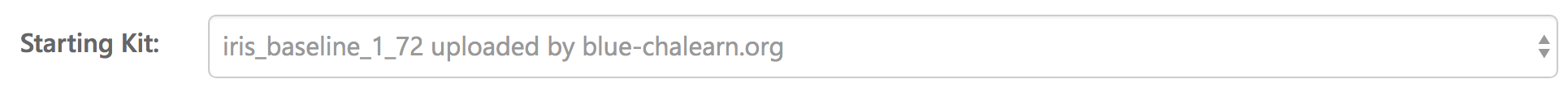
But this will not allow them to edit the challenge. To give them this privilege, go to the editor and add their Codalab ID to the list of admins:



The Codalab editor can be used to edit your challenge. However, try to minimize to avoid winding up with a broken site. Save your intermediate challenge bundles by using the **Dump** button.



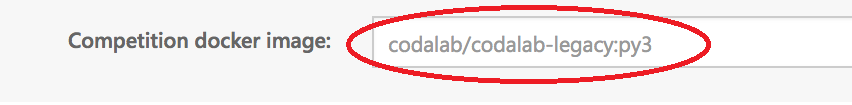
Editing your challenge allows you to update your starting kit. You can upload it in my datasets, then point to it from the editor:



If your score is an “accuracy” (better is higher), you want the leaderboard sorted in descending order, but if it is an “error” (better is lower), you want it in ascending order. You can change that:



We want to run the code into Python 3 dockers. Make sure the correct docker is indicated:



**Using dockers**

Codalab allows you to specify a docker in which you code will be running. Thus it is convenient to test your code in this docker ahead of time.

When you submit code to the Codalab platform, your code is executed inside a docker container. This environment can be exactly reproduced on your local machine by downloading the corresponding docker image. The codalab/codalab-legacy:py3 docker, which can be pulled from <https://hub.docker.com/r/codalab/codalab-legacy/tags/> contains a large number of pre-loaded programs, including Python 3 and many libraries including scikit-learn. See <https://github.com/codalab/codalab-dockers/tree/master/legacy-py3> for details.

To run the starting kit inside the Codalab docker from a terminal:

**Step 1) Create a temporary folder** to put everything you will need (data and starting kit):

mkdir ~/aux

cp starting\_kit.zip ~/aux

Copy any other data/code you want to that folder, as we will instruct you to map such local folder to be visible when you are within the docker. If you don’t do such mapping procedure, you will not be able to see what you need from within the docker.

**Step 2) Run the docker** (and map such folder, described above, to /home/aux, inside the docker):

docker run -it -v ~/aux:/home/aux codalab/codalab-legacy:py3

The first time around it takes a while because docker must download the entire docker image. if everything goes well, you will be inside the docker and the prompt will be like below:

root@c74a8b1ccaf7:/#

**Step 3) Run the starting kit.** Inside the docker, go to “/home/aux” and unzip and run the starting kit.

# cd /home/aux

# unzip starting\_kit.zip

# cd starting\_kit

# python3 ingestion\_program/ingestion.py sample\_data sample\_result\_submission ingestion\_program sample\_code\_submission

# python3 scoring\_program/score.py sample\_data sample\_result\_submission scoring\_output

# exit

WARNING: inside the docker, python 3.6 is called python3.

**Step 4) Run the jupyter notebook.**

Same thing as before, but add : **-p 8888:8888**

docker run -it -p 8888:8888 -v ~/aux:/home/aux codalab/codalab-legacy:py3

Go to a web browser and check that the notebook is running at http://localhost:8888/.

Then open README.pynb. WARNING: the default notebook kernel is Python 2, you’ll have to switch to Python 3. If you do not have a Python 3 kernel… you are in trouble. [Fixing it is complicated](https://ipython.readthedocs.io/en/latest/install/kernel_install.html).

**Modifying your docker**

The docker we recommend contains a lot of libraries, but you may need more for your specific competition. You can then make changes to it [following these instructions](https://github.com/codalab/codalab-competitions/wiki/User_Docker). But AVOID THIS if you can because you will then have another docker as everyone else, this may confused the L2 students. Also, you will need to test that it works fine on Codalab. If you make a new docker, it must be posted on [Docker Hub](https://hub.docker.com/). Preferable use a docker file to specify your changes and put them under revision control on Github.

How to use Git and Github

There are several good tutorials and cheat sheets on the Internet. You should start by cloning the repo of the starting kit, this is better than downloading the zip:

git clone https://github.com/madclam/m2aic2019.git

If you want to create another repo, [follow these instructions](https://help.github.com/articles/adding-an-existing-project-to-github-using-the-command-line/).