OpenML-Python

Citation for published version (APA):

Document status and date:
Published: 06/11/2019

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 30. Jan. 2020
OpenML-Python: an extensible Python API for OpenML

Matthias Feurer  
University of Freiburg, Freiburg, Germany  
FEURERM@CS.UNI-FREIBURG.DE

Jan N. van Rijn  
Leiden University, Leiden, Netherlands  
J.N.VAN.RIJN@LIACS.LEIDENUNIV.NL

Arlind Kadra  
University of Freiburg, Freiburg, Germany  
KADRAA@CS.UNI-FREIBURG.DE

Pieter Gijsbers  
Eindhoven University of Technology, Eindhoven, Netherlands  
P.GIJSBERS@TUE.NL

Neeratyoy Mallik  
University of Freiburg, Freiburg, Germany  
MALLIK@CS.UNI-FREIBURG.DE

Sahithya Ravi  
Eindhoven University of Technology, Eindhoven, Netherlands  
S.RAVI@TUE.NL

Andreas Müller  
Columbia University, New York, USA  
ACM2248@COLUMBIA.EDU

Joaquin Vanschoren  
Eindhoven University of Technology, Eindhoven, Netherlands  
J.VANSCHOREN@TUE.NL

Frank Hutter  
University of Freiburg, Freiburg & Bosch Center for Artificial Intelligence, Germany  
FH@CS.UNI-FREIBURG.DE

Abstract

OpenML is an online platform for open science collaboration in machine learning, used to share datasets and results of machine learning experiments. In this paper we introduce OpenML-Python, a client API for Python, opening up the OpenML platform for a wide range of Python-based tools. It provides easy access to all datasets, tasks and experiments on OpenML from within Python. It also provides functionality to conduct machine learning experiments, upload the results to OpenML, and reproduce results which are stored on OpenML. Furthermore, it comes with a scikit-learn plugin and a plugin mechanism to easily integrate other machine learning libraries written in Python into the OpenML ecosystem. Source code and documentation is available at https://github.com/openml/openml-python/.

Keywords: Python, Collaborative Science, Meta-Learning, Reproducible Research

1. Introduction

OpenML is a collaborative online machine learning (ML) platform, meant for sharing and building on prior empirical machine learning research (Vanschoren et al., 2014).

It goes beyond open data repositories, such as UCI (Dua and Graff, 2017), PMLB (Olson et al., 2018), the ‘datasets’ submodules in scikit-learn (Pedregosa et al., 2011) and tensor-
flow (Abadi et al., 2016), and the closed-source data sharing platform at Kaggle.com, since OpenML also collects millions of shared experiments on these datasets, linked to the exact ML pipelines and hyperparameter settings, and includes comprehensive logging and uploading functionalities which can be accessed programmatically via a REST API. However, sharing ML experiments adds significant complexity to most people’s workflows.

OpenML-Python is a seamless integration of OpenML into the popular Python ML ecosystem,1 that takes away this complexity by providing easy programmatic access to all OpenML data and automating the sharing of new experiments.2 In this paper, we introduce OpenML-Python’s core design, showcase its extensibility to new ML libraries, and give code examples for several common research tasks.

2. Use cases for the OpenML-Python API

OpenML-Python allows for easy dataset and experiment sharing by handling all communication with OpenML’s REST API. In this section, we briefly describe how the package can be used in several common machine learning tasks and highlight recent uses.

Working with datasets. OpenML-Python can retrieve the thousands of datasets on OpenML (all of them, or specific subsets) in a unified format, retrieve meta-data describing them, and search through them with filters. Datasets are converted from OpenML’s internal format into numpy, scipy or pandas data structures, which are standard for ML in Python. To facilitate contributions from the community, it allows people to upload new datasets in only two function calls, and to define new tasks on them (combinations of a dataset, train/test split and target attribute).

Publishing and retrieving results. Sharing empirical results allows anyone to search and download them in order to reproduce and reuse them in their own research. One goal of OpenML is to simplify the comparison of new algorithms and implementations to existing approaches by comparing to the results on OpenML. To this end we also provide an interface for integrating new machine learning libraries with OpenML and we have already integrated scikit-learn. OpenML-Python can then be used to set up and conduct machine learning experiments for a given task and flow (an ML pipeline including hyperparameters and random states), and publish reproducible results.

Use cases in published works. OpenML-Python has already been used to scale up studies with hundreds of consistently formatted datasets (Feurer et al., 2015; Fusi et al., 2018), supply large amounts of meta-data for meta-learning (Perrone et al., 2018), answer questions about algorithms such as hyperparameter importance (van Rijn and Hutter, 2018) and facilitate large-scale comparisons of algorithms (Strang et al., 2018).

3. High-level Design of OpenML-Python

The OpenML platform is organized around several entity types which describe different aspects of a machine learning study. It hosts datasets, tasks that define how models should be evaluated on them, flows that record the structure and other details of ML pipelines, and

---

2. Other clients already exist for R (Casalicchio et al., 2017) and Java (van Rijn 2016).
import openml; import numpy as np
import matplotlib.pyplot as plt

df = openml.evaluations.list_evaluations_setups(
    'predictive_accuracy', flow=[8353], task=[6],
    output_format='dataframe', parameters_in_separate_columns=True,
)  # Choose an SVM flow (e.g. 8353), and the dataset 'letter' (task 6).

hp_names = ['sklearn.svm.classes.SVC(16).C', 'sklearn.svm.classes.SVC(16).gamma']
df[hp_names] = df[hp_names].astype(float).apply(np.log)
C, gamma, score = df[hp_names[0]], df[hp_names[1]], df['value']
cntr = plt.tricontourf(C, gamma, score, levels=12, cmap='RdBu_r')
plt.colorbar(cntr, label='accuracy')
plt.xlim((min(C), max(C))); plt.ylim((min(gamma), max(gamma)))
plt.xlabel('C (log10)', size=16); plt.ylabel('gamma (log10)', size=16)
plt.title('SVM performance landscape', size=20)

Figure 1: Code for retrieving the predictive accuracy of an SVM classifier on the ‘letter’
dataset and creating a contour plot with the results.

runs that record the experiments evaluating specific flows on certain tasks. For instance,
an experiment (run) shared on OpenML can show how a random forest (flow) performs
on ‘iris’ (dataset) if evaluated with 10-fold cross-validation (task), and how to reproduce
that result. In OpenML-Python, all these entities are represented by classes, each defined
in their own submodule. This implements a natural mapping from OpenML concepts to
Python objects. While OpenML is an online platform, we facilitate offline usage as well.

Plugins. To allow users to automatically run and share machine learning experiments
with different libraries through the same OpenML-Python interface, we designed a plu-
gin interface that standardizes the interaction between machine learning library code and
OpenML-Python. We also created a plugin for scikit-learn (Pedregosa et al., 2011), as it
is one of the most popular Python machine learning libraries. This plugin can be used for
any library which follows the scikit-learn API (Buitinck et al., 2013).

A plugin’s responsibility is to convert between the libraries’ models and OpenML flows,
interact with its training interface and format predictions. For example, the scikit-learn
plugin can convert an OpenMLFlow to an Estimator (including hyperparameter settings),
train models and produce predictions for a task, and create an OpenMLRun object to up-
load the predictions to the OpenML server. The plugin also handles advanced procedures,
such as scikit-learn’s random search or grid search and uploading its traces (hyperparam-
eters and scores of each model evaluated during search).

We are working on more plugins, and anyone can contribute their own using the scikit-learn plugin
implementation as a reference.

4. Examples

We show two example uses of OpenML-Python to
demonstrate its API’s simplicity. First, we show how
to retrieve results and evaluations from the OpenML
server in Figure 1 (generating the plot on the right).
```
import openml
import sklearn.tree, sklearn.impute, sklearn.pipeline

# obtain a benchmark suite
benchmark_suite = openml.study.get_suite('OpenML-CC18')
clf = sklearn.pipeline.Pipeline(steps=[
    ('imputer', sklearn.impute.SimpleImputer()),
    ('estimator', sklearn.tree.DecisionTreeClassifier()),
])  # build a sklearn classifier
for task_id in benchmark_suite.tasks:  # iterate over all tasks
    task = openml.tasks.get_task(task_id)  # download the OpenML task
    run = openml.runs.run_model_on_task(clf, task)  # run classifier on splits
    # run.publish()  # upload the run to the server, optional
```

Figure 2: Training and evaluating a decision tree classifier from scikit-learn on each task of
the OpenML-CC18 benchmark suite (Bischl et al., 2019).

Second, in Figure 2 we show how to conduct experiments on a benchmark suite (Bischl et al.,
2019). Further examples, including how to create datasets and tasks and how OpenML-
Python was used in previous publications, can be found in the online documentation.3

5. Project development

The project has been set up for development through community effort from different re-
search groups, and has received contributions from numerous individuals. The package is
developed publicly through Github which also provides an issue tracker for bug reports,
feature requests and usage questions. To ensure a coherent and robust code base we use
continuous integration for Windows and Linux as well as automated type and style check-
ing. Documentation is also rendered on continuous integration servers and consists of a mix
of tutorials, examples and API documentation.

For ease of use and stability, we use well-known and established third-party packages
where needed. For instance, we build documentation using the popular sphinx Python doc-
umentation generator4, use an extension to automatically compile examples into documen-
tation and Jupyter notebooks5, and employ standard open-source packages for scientific
computing such as numpy, scipy (Virtanen et al., 2019), and pandas (McKinney, 2010).
The package is written in Python3 and open-sourced with a 3-Clause BSD License.3

6. Conclusion

OpenML-Python allows easy interaction with OpenML from within Python. It makes it easy
for people to share and reuse the data, meta-data, and empirical results which are generated
as part of an ML study. This allows for better reproducibility, simpler benchmarking and
easier collaboration on ML projects. Our software is shipped with a scikit-learn plugin and
has a plugin mechanism to easily integrate other ML libraries written in Python.

3. We provide documentation and code examples on http://openml.github.io/openml-python and host the
   project on http://github.com/openml/openml-python.
Acknowledgments

MF, NM and FH acknowledge funding by the Robert Bosch GmbH. AK, JvR and FH acknowledge funding by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant no. 716721. JV and PG acknowledge funding by the Data Driven Discovery of Models (D3M) program run by DARPA and the Air Force Research Laboratory. The authors also thank Bilge Celik, Victor Gal and everyone listed at https://github.com/openml/openml-python/graphs/contributors for their contributions.

References


