PMLB v1.0: an open source dataset collection for benchmarking machine learning methods

Authors

- **Trang T. Le**
  ⓘD [0000-0003-3737-6565](#) · ○ [trang1618](#) · 🐦 [trang1618](#)
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **William La Cava**
  ⓘD [0000-0002-1332-2960](#) · ○ [lacava](#) · 🐦 [w_la_cava](#)
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Joseph D. Romano**
  ⓘD [0000-0002-7999-4399](#) · ○ [jdromano2](#) · 🐦 [jdromano2](#)
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104; Center of Excellence in Environmental Toxicology, University of Pennsylvania, Philadelphia, PA 19104

- **John T. Gregg**
  ⓘD [0000-0002-2619-3440](#) · ○ [greggj2016](#)
  Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19087, USA.

- **Daniel J. Goldberg**
  ⓘD [0000-0003-4173-9867](#) · ○ [daniel0710goldberg](#)
  Department of Computer Science & Engineering, Washington University in St. Louis, St. Louis, MO 63130

- **Praneel Chakraborty**
  ⓘD [0000-0001-9586-0721](#) · ○ [praneelc](#)
  School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104; Wharton School, University of Pennsylvania, Philadelphia, PA 19104

- **Natasha L. Ray**
  ⓘD [0000-0001-6883-4624](#) · ○ [natray21](#)
  Princeton Day School, Princeton, NJ 08540

- **Daniel Himmelstein**
  ⓘD [0000-0002-3012-7446](#) · ○ [dhimmel](#) · 🐦 [dhimmel](#)
  Related Sciences; Department of Systems Pharmacology & Translational Therapeutics, University of Pennsylvania, Philadelphia, PA 19104

- **Weixuan Fu**
  ⓘD [0000-0002-6434-5468](#) · ○ [weixuanfu](#) · 🐦 [weixuanfu](#)
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Jason H. Moore**

  ⓘ [0000-0002-5015-1099](#) · ⓖ [EpistasisLab](#) · 🐦 [moorejh](#)

  Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19087 · Funded by National Institutes of Health Grant Nos. LM010098 and AI116794.

Summary

PMLB (Penn Machine Learning Benchmark) is an open source data repository containing a curated collection of datasets for evaluating and comparing machine learning (ML) algorithms. Compiled from a broad range of existing ML benchmark collections, PMLB synthesizes and standardizes hundreds of publicly available datasets from diverse sources such as the [UCI ML repository](#) and [OpenML](#) [1], enabling systematic assessment of different ML methods. These datasets cover a range of applications, from binary/multi-class classification to regression problems with combinations of categorical and continuous features. PMLB has both a Python interface (`pmlb`) and an R interface (`pmlbr`), both with detailed documentation that allows the user to access cleaned and formatted datasets using a single function call (`fetch_data`). PMLB also provides a comprehensive description of each dataset and advanced functions to explore the dataset space such as `nearest_datasets` and `filter_datasets` , which allow for smoother user experience and handling of data. The resource is designed to facilitate open source contributions in the form of datasets as well as improvements to curation.

Statement of need

Benchmarking is a standard practice to illustrate the strengths and weaknesses of algorithms with regards to different problem characteristics. In ML, benchmarking often involves assessing the performance of specific ML models — namely, how well they predict labels for new samples (supervised learning) or how well they organize and/or represent data with no pre-existing labels (unsupervised learning). The extent to which ML methods achieve these aims is typically evaluated over a group of benchmark datasets [2,3]. PMLB was designed to provide a suite of such datasets with uniform formatting, as well as the framework for conducting automatic evaluation of the different algorithms.

The original release of PMLB (v0.2) [4] received positive feedback from the ML community, reflecting the pressing need for a collection of standardized datasets to evaluate models without intensive preprocessing and dataset curation. As the repository becomes more widely used, community members have requested new features such as additional information about the datasets, as well as new functions to select datasets given specific criteria. In this paper, we review the original functionality and present new enhancements that facilitate frictionless interaction with the repository, both from the perspective of database contributors and end-users.

Differentiating attributes

New datasets with rich metadata

Since its previous major release (0.2) [4], we have made substantial improvements in the collection of new datasets as well as other helpful supporting features. Furthermore we have redesigned the repository structure, and PMLB now includes benchmark datasets for regression problems (Fig. [1](#)). To fulfill [requests made by several users](#), each dataset also includes a `metadata.yaml` file that contains general descriptive information about the dataset itself (an example can be viewed [here](#)). Specifically, for each dataset, the metadata file includes a web address to the original source of the

dataset, a text description of the dataset's purpose, the publication associated with the dataset generation, the type of learning problem it was designed for (i.e., classification or regression), keywords (e.g., "simulation", "ecological", "bioinformatics"), and a description of individual features and their coding schema (e.g., 'non-promoter'= 0, 'promoter'= 1). Metadata files are supported by a standardized format that is formalized using JSON-Schema (version `draft-07`) [5] — upcoming releases of PMLB will include automated validation of datasets and metadata files to further improve ease of contribution and data accuracy.
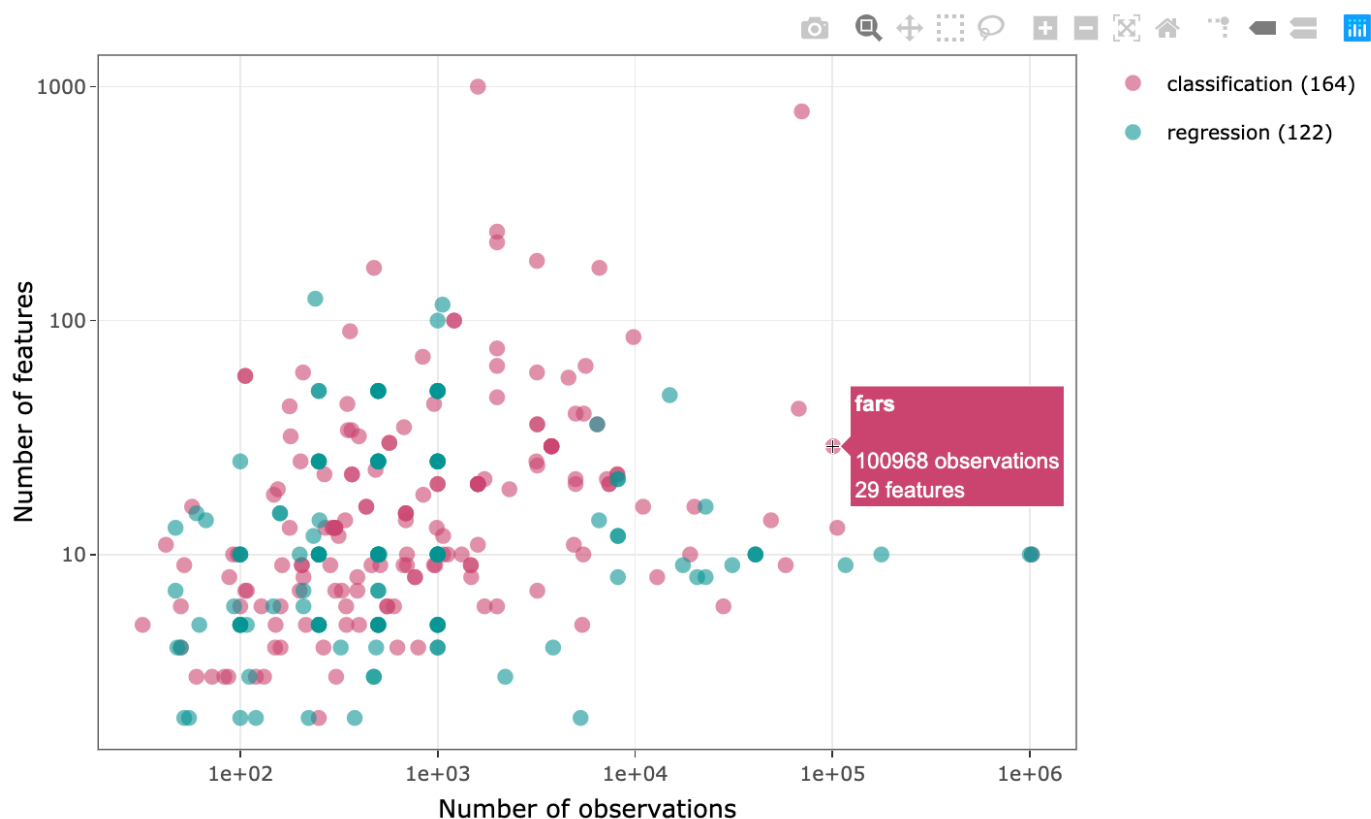

Figure 1: Characteristics of datasets in the PMLB collection

Aside from the authors of this paper, a number of open source contributors have been invaluable in providing manually-curated metadata. In addition, careful examination by open source contributors have led to important bug fixes, such as a correction to the target column in the bupa dataset.

User-friendly interfaces

On PMLB's home page, users can now browse, sort, filter, and search datasets from a lookup table of datasets with summary statistics (Fig. 2). To select datasets with numerical values for specific metaparameters (e.g., number of observations, number of features, class balance, etc.), users can type ranges in the box at the bottom of each numeric column in the format `low ... high`. For example, if the user wants to view all classification datasets with 80 to 100 observations, they would select `classification` at the bottom of the `Task` column, and type `80 ... 100` at the bottom of the `n_observations` column. The `CSV` button allows the user to download the table's contents with any active filters applied.

Dataset	n_observations	n_features	n_classes	Endpoint	Imbalance	Task	Metadata
analcatdata_asbestos	83	3	2	categorical	0.01	classification	⬤
analcatdata_creditscore	100	6	2	categorical	0.21	classification	⬤
analcatdata_cyyoung8092	97	10	2	categorical	0.26	classification	⬤
analcatdata_cyyoung9302	92	10	2	categorical	0.34	classification	⬤
lupus	87	3	2	categorical	0.04	classification	⬤
postoperative_patient_data	88	8	2	categorical	0.21	classification	⬤
All	80.00 ... 100.00	All	All	All	All	["classific	All

CSV · Show 10 entries · Search:

Figure 2: Dataset summary statistics table, with advanced searching, filtering, and sorting features

On the website, we have also published a detailed but concise contribution guide with step-by-step instructions on how to add new datasets, submit edits for existing datasets, or improve the provided Python or R code. When a new dataset is added, summary statistics (e.g., number of observations, number of classes, etc.) are automatically computed, a profiling report is generated (see below), a corresponding metadata template is added to the dataset folder, and PMLB's list of available dataset names is updated. Other checks included in the continuous integration workflow help to reduce the amount of work required from both contributers and code reviewers.

In addition to the Python interface for PMLB, we have included an R library, both of which can be installed with a single command. The R library has been adapted from a separate repository that is currently unmaintained, but released under the GPL-2 license. However, because the original source code was released under the GNU General Public License, version 2. The R library also includes a number of detailed "vignette" documents to help new users learn how to use the software.

PMLB now includes original data rows with missing data (i.e., NA). The new version of PMLB also allows the user to select datasets most similar to one of their own using the `nearest_datasets` function. Here, the similarity between datasets is configurable to any number of metadata characteristics (e.g., number of samples, number of features, number of target classes, etc.). This functionality is helpful for users who wish to find PMLB datasets with similar characteristics to their own in order to test or optimize methods (e.g. hyperparameter tuning) for their desired problem without the risk of over-fitting to their dataset.

An API reference guide that details all user-facing functions and variables in PMLB's Python and R libraries is included on the PMLB website.

Pandas profiling reports

For each dataset, we use `pandas-profiling` to generate summary statistic reports. In addition to the descriptive statistics provided by the commonly-used `pandas.describe` (Python) [6] or `skimr::skim` (R) functions, `pandas-profiling` gives a more extensive exploration of the dataset, including correlation structure within the dataset and flagging of duplicate samples. Browsing a report allows users and contributors to easily assess dataset quality and make any necessary changes. For example, if a feature is flagged by `pandas-profiling` as having a single value replicated in all samples, it is likely that this feature is uninformative for ML analysis and should be removed from the dataset.

The profiling reports can be accessed by clicking on the dataset name in the interactive data table or the data point in the interactive chart on the PMLB website. Alternatively, all reports can be viewed on the repository's gh-pages branch, or generated manually by users on their local computing resources.

Space efficiency

We have significantly reduced the size of the PMLB source repository by using Git Large File Storage (LFS) to efficiently track changes in large database source files [7]. Users who would like to interact with the entire repository (including the complete database sources) locally can do so by either installing Git LFS and cloning the PMLB repository, or by downloading a ZIP archive of the repository from GitHub in a web browser.

References

1. **OpenML**
   Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, Luis Torgo
   ACM SIGKDD Explorations Newsletter (2014-06-16) https://doi.org/gf238r
   DOI: 10.1145/2641190.2641198

2. **Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition**
   J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel
   Neural Networks (2012-08) https://doi.org/f3z6dz
   DOI: 10.1016/j.neunet.2012.02.016

3. **An empirical comparison of supervised learning algorithms**
   Rich Caruana, Alexandru Niculescu-Mizil
   Association for Computing Machinery (ACM) (2006) https://doi.org/bmstc2
   DOI: 10.1145/1143844.1143865

4. **PMLB: a large benchmark suite for machine learning evaluation and comparison**
   Randal S. Olson, William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, Jason H. Moore
   BioData Mining (2017-12-11) https://doi.org/gfrbw5
   DOI: 10.1186/s13040-017-0154-4 · PMID: 29238404 · PMCID: PMC5725843

5. **Foundations of JSON Schema**
   Felipe Pezoa, Juan L. Reutter, Fernando Suarez, Martín Ugarte, Domagoj Vrgoč
   Association for Computing Machinery (ACM) (2016) https://doi.org/ghcsq4
   DOI: 10.1145/2872427.2883029

6. **Data Structures for Statistical Computing in Python**
   Wes McKinney
   SciPy (2010) https://doi.org/ggr6q3
   DOI: 10.25080/majora-92bf1922-00a

7. **Ten Simple Rules for Taking Advantage of Git and GitHub**
   Yasset Perez-Riverol, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe da Veiga Leprevost, Christian Fufezan, Tobias Ternent, Stephen J. Eglen, Daniel S. Katz, … Juan Antonio Vizcaíno
   PLOS Computational Biology (2016-07-14) https://doi.org/gbrb39
   DOI: 10.1371/journal.pcbi.1004947 · PMID: 27415786 · PMCID: PMC4945047