Package 'pmlbr'

July 23, 2025

Title Interface to the Penn Machine Learning Benchmarks Data Repository

Description Check available classification and regression data sets from the PMLB repository and download them.
The PMLB repository (<https://github.com/EpistasisLab/pmlbr>) contains a curated collection of data sets for evaluating and comparing machine learning algorithms.
These data sets cover a range of applications, and include binary/multi-class classification problems and
regression problems, as well as combinations of categorical, ordinal, and continuous features.
There are currently over 150 datasets included in the PMLB repository.

Version 0.3.0

BugReports https://github.com/EpistasisLab/pmlbr/issues

Depends R (>= 3.2.0)

Imports utils, FNN, stats

License GPL-2 | file LICENSE

URL https://github.com/EpistasisLab/pmlbr

Encoding UTF-8

RoxygenNote 7.3.2

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Trang Le [aut, cre] (https://trang.page/),
makeyourownmaker [aut] (https://github.com/makeyourownmaker),
Jason Moore [aut] (http://www.epistasisblog.org/),
University of Pennsylvania [cph]

Maintainer Trang Le <grixor@gmail.com>

Repository CRAN

Date/Publication 2025-02-28 03:10:02 UTC

Contents

`classification_datasets` *Classification datasets*

Description

Classification datasets

Usage

```
classification_datasets()
```

Value

A character vector of classification dataset names.

Examples

```
if (interactive()) {
  sample(classification_datasets(), 10)
}
```

compute_imbalance *Computes imbalance value for a given dataset.*

Description

Computes imbalance value for a given dataset.

Usage

```
compute_imbalance(target_col)
```

Arguments

target_col	Factor or character vector of target column.

Value

A value of imbalance metric, where zero means that the dataset is perfectly balanced and the higher the value, the more imbalanced the dataset.

dataset_names *All available datasets*

Description

All available datasets

Usage

```
dataset_names()
```

Value

A character vector of all dataset names.

Examples

```
if (interactive()) {
  sample(dataset_names(), 10)
}
```

fetch_data *fetch_data function*

Description

Download a data set from the PMLB repository, (optionally) store it locally, and return the data set. You must be connected to the internet if you are fetching a data set that is not cached locally.

Usage

```
fetch_data(
  dataset_name,
  return_X_y = FALSE,
  local_cache_dir = NA,
  dropna = TRUE
)
```

Arguments

dataset_name	The name of the data set to load from PMLB
return_X_y	Boolean. Whether to return the data with the features and labels stored in separate data structures or a single structure (can be TRUE or FALSE, defaults to FALSE)
local_cache_dir	The directory on your local machine to store the data files in (defaults to NA, indicating cache will not be used)
dropna	Boolean. Whether rows with NAs should be automatically dropped. Default to TRUE.

See Also

pmlb_metadata.

Examples

```
# Features and labels in single data frame
if (interactive()){
  penguins <- fetch_data("penguins")
  head(penguins)

  # Features and labels stored in separate data structures
  penguins <- fetch_data("penguins", return_X_y = TRUE)
  penguins$x # data frame
  penguins$y # vector
}
```

get_type *Get type/class of given vector.*

Description

Get type/class of given vector.

Usage

```
get_type(x, include_binary = FALSE)
```

Arguments

x	Input vector.
include_binary	Boolean. Whether binary should be counted separately from categorical.

Value

Type/class of ‘x‘.

graceful_download *Download a File Gracefully with Retry Mechanism*

Description

Attempts to download a file from a specified URL, retrying a set number of times if the download fails. This function meets CRAN's requirement for gracefully handling the use of internet resources by catching errors and returning a warning message if the download ultimately fails.

Usage

```
graceful_download(url, destfile, retries = 3)
```

Arguments

url	Character. The URL of the file to download.
destfile	Character. The path to the destination file where the downloaded content will be saved.
retries	Integer. The maximum number of download attempts (default is 3).

Value

Logical. Returns ‘TRUE‘ if the download succeeds, ‘FALSE‘ otherwise.

Examples

```
## Not run:
dataset_url <- "https://example.com/dataset.csv"
tmp <- tempfile(fileext = ".csv")
success <- download_file_gracefully(dataset_url, tmp)
if (!success) {
  message("Continuing gracefully without the dataset.")
}

## End(Not run)
```

`nearest_datasets` *Select nearest datasets given input 'x'.*

Description

If 'x' is a data.frame object, computes dataset characteristics. If 'x' is a character object specifying dataset name from PMLB, use the already computed dataset statistics/characteristics in 'summary_stats'.

Usage

```
nearest_datasets(x, ...)

## Default S3 method:
nearest_datasets(x, ...)

## S3 method for class 'character'
nearest_datasets(
  x,
  n_neighbors = 5,
  dimensions = c("n_instances", "n_features"),
  target_name = "target",
  ...
)

## S3 method for class 'data.frame'
nearest_datasets(
  x,
  y = NULL,
  n_neighbors = 5,
  dimensions = c("n_instances", "n_features"),
  task = c("classification", "regression"),
  target_name = "target",
  ...
)
```

Arguments

x	Character string of dataset name from PMLB, or data.frame of n_samples x n_features(or n_features+1 with a target column)
...	Further arguments passed to each method.
n_neighbors	Integer. The number of dataset names to return as neighbors.
dimensions	Character vector specifying dataset characteristics to include in similarity calculation. Dimensions must correspond to numeric columns of [all_summary_stats.tsv](https://github.com/E If 'all' (default), uses all numeric columns.
target_name	Character string specifying column of target/dependent variable.
y	Vector of target column. Required when 'x" does not contain the target column.
task	Character string specifying classification or regression for summary stat generation.

Value

Character string of names of most similar datasets to df, most similar dataset first.

Examples

```
if (interactive()){
  nearest_datasets('penguins')
  nearest_datasets(fetch_data('penguins'))
}
```

---

pmlb *pmlb: R interface to the Penn Machine Learning Benchmarks data repository*

Description

The PMLB repository contains a curated collection of data sets for evaluating and comparing machine learning algorithms. These data sets cover a range of applications, and include binary/multiclass classification problems and regression problems, as well as combinations of categorical, ordinal, and continuous features. There are approximately 290 data sets included in the PMLB repository and there are no missing values in these data sets.

Details

This R library includes summaries of the classification and regression data sets but does NOT include any of the PMLB data sets. The data sets can be downloaded using the `fetch_data` function which is similar to the corresponding PMLB python function.

See `fetch_data`, `pmlb_metadata` for usage examples and further information.

If you use PMLB in a scientific publication, please consider citing the following paper:

Randal S. Olson, William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, and Jason H. Moore (2017).

PMLB: a large benchmark suite for machine learning evaluation and comparison
https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0154-4
BioData Mining 10, page 36.

Author(s)

Maintainer: Trang Le <grixor@gmail.com> (https://trang.page/)

Authors:

- makeyourownmaker <makeyourownmaker@gmx.com> (https://github.com/makeyourownmaker)
- Jason Moore <jhmoore@upenn.edu> (http://www.epistasisblog.org/)

Other contributors:

- University of Pennsylvania [copyright holder]

See Also

Useful links:

- https://github.com/EpistasisLab/pmlbr
- Report bugs at https://github.com/EpistasisLab/pmlbr/issues

---

`pmlb_metadata` *Get metadata for all datasets in PMLB.*

Description

Metadata like summary statistics and names of available datasets on the PMLB repository.

Usage

```
pmlb_metadata()
```

Value

A list containing summary_stats, dataset_names, classification_datasets, and regression_datasets

Examples

```
if (interactive()) {
  sample(pmlb_metadata()$dataset_names, 10)
}
```

regression_datasets *Regression datasets*

Description

Regression datasets

Usage

```
regression_datasets()
```

Value

A character vector of regression dataset names.

Examples

```
if (interactive()) {
  sample(regression_datasets(), 10)
}
```

summary_stats *Summary statistics*

Description

Summary statistics

Usage

```
summary_stats()
```

Value

A dataframe of summary statistics of all available datasets, including number of instances/rows, number of columns/features, task, etc.

Examples

```
if (interactive()) {
  head(summary_stats())
}
```

Index