Package 'trainsplit'

January 13, 2025

Title Split a Dataframe, Tibble, or Data.table into Training and Test Sets

Version 1.2

Description Split a dataframe, tibble, or data.table into training and test sets. Return either a list, an index, or directly assign training and test sets into memory.

URL https://github.com/eastnile/trainsplit

Encoding UTF-8

RoxygenNote 7.3.1

Imports data.table

Suggests tibble, dplyr

License MIT + file LICENSE

NeedsCompilation no

Author Zhaochen He [aut, cre] (<https://orcid.org/0000-0002-6579-5073>)

Maintainer Zhaochen He <eastnileuc@gmail.com>

Repository CRAN

Date/Publication 2025-01-13 17:10:02 UTC

Contents

`trainsplit` *trainsplit*

Description

Splits a dataframe, tibble, or data.table into a test set and training set. Specify either the number or percentage of observations to be put into training set.

Usage

```
trainsplit(
  data,
  ntrain = NULL,
  trainpct = NULL,
  round_ntrain = "round",
  seed = NULL,
  return = "parentenv"
)
```

Arguments

data	The dataset you want to split
ntrain	The number of observations to go into the training set. Must be >= 0 and <= nrow(data).
trainpct	Fraction of observations to go into training set. Must be >= 0 and =< 1. If set to 0 or 1, the empty test or training set will still inherit the same column names and types as the original dataset.
round_ntrain	What to do when nrow(data) * trainpct is not a whole number. Default behavior is to round the size of the training set. Use 'ceiling' or 'floor' to instead set the size of training set to next highest or lowest whole number.
seed	Sets the random seed; use this argument if you want to always get the same result. Note: sets seed only locally within the function.
return	Three return modes available: "parentenv" assigns the training and test sets into the environment that called the function with names based on the name of the original dataset; this is intended largely for an educational context. "list" will return a list with the training and test sets. "index" will return only the numerical index of the rows to be placed into the training set, which can then be manually subset by the user.

Value

Depends on "return" argument; either a list, an index, or NULL if return = "parentenv" was selected.

Examples

```
# Splits the training and test sets and assigns them into memory.
trainsplit(mtcars, trainpct = 0.75)
# Specify size of training set by number of rows, not percent:
trainsplit(mtcars, ntrain = 10)
# Size of training set rounds to one:
trainsplit(mtcars, trainpct = 0.01, round_ntrain = 'ceiling')
# Also works with data.table:
trainsplit(data.table::as.data.table(mtcars), trainpct = 0.75)
# Return a list containing the training/test sets instead:
trainsplit(mtcars, trainpct = 0.75, return = 'list')
```

Index