

# Package ‘splitTools’

March 23, 2023

**Title** Tools for Data Splitting

**Version** 0.3.3

**Description** Fast, lightweight toolkit for data splitting. Data sets can be partitioned into disjoint groups (e.g. into training, validation, and test) or into (repeated) k-folds for subsequent cross-validation. Besides basic splits, the package supports stratified, grouped as well as blocked splitting. Furthermore, cross-validation folds for time series data can be created. See e.g. Hastie et al. (2001)  [<doi:10.1007/978-0-387-84858-7 >](https://doi.org/10.1007/978-0-387-84858-7) for the basic background on data partitioning and cross-validation.

**License** GPL (>= 2)

**Encoding** UTF-8

**RoxygenNote** 7.2.3

**Imports** stats

**Suggests** knitr, ranger, rmarkdown, testthat (>= 3.0.0)

**URL** <https://github.com/mayer79/splitTools>

**BugReports** <https://github.com/mayer79/splitTools/issues>

**VignetteBuilder** knitr

**Config/testthat/edition** 3

**NeedsCompilation** no

**Author** Michael Mayer [aut, cre]

**Maintainer** Michael Mayer <mayermichael79@gmail.com>

**Repository** CRAN

**Date/Publication** 2023-03-23 18:50:02 UTC

## R topics documented:

create_folds . . . . .	2
create_timefolds . . . . .	3
multi_strata . . . . .	4
partition . . . . .	5

---

create_folds	<i>Create Folds</i>
--------------	---------------------

---

### Description

This function provides a list of row indices used for k-fold cross-validation (basic, stratified, grouped, or blocked). Repeated fold creation is supported as well.

### Usage

```
create_folds(
  y,
  k = 5,
  type = c("stratified", "basic", "grouped", "blocked"),
  n_bins = 10,
  m_rep = 1,
  use_names = TRUE,
  invert = FALSE,
  shuffle = FALSE,
  seed = NULL
)
```

### Arguments

y	Either the variable used for "stratification" or "grouped" splits. For other types of splits, any vector of the same length as the data intended to split.
k	Number of folds.
type	Split type. One of "stratified", "basic", "grouped", "blocked". The default is "stratified".
n_bins	Approximate numbers of bins for numeric y and type = "stratified".
m_rep	How many times should the data be split into k folds? Default is 1, i.e. no repetitions.
use_names	Should folds be named? Default is TRUE.
invert	Set to TRUE in order to receive out-of-sample indices. Default is FALSE, i.e. in-sample indices are returned.
shuffle	Should row indices be randomly shuffled within folds? Default is FALSE.
seed	Integer random seed.

### Details

By default, the function uses stratified splitting. This will balance the folds regarding the distribution of the input vector y. Numeric input is first binned into n\_bins quantile groups. If type = "grouped", groups specified by y are kept together when splitting. This is relevant for clustered or panel data. In contrast to basic splitting, type = "blocked" does not sample indices at random, but rather keeps them in sequential groups.

**Value**

If `invert = FALSE` (the default), a list with in-sample row indices. If `invert = TRUE`, a list with out-of-sample indices.

**Examples**

```
y <- rep(c(letters[1:4]), each = 5)
create_folds(y)
create_folds(y, k = 2)
create_folds(y, k = 2, m_rep = 2)
create_folds(y, k = 3, type = "blocked")
```

---

<code>create_timefolds</code>	<i>Creates Folds for Time Series Data</i>
-------------------------------	---

---

**Description**

This function provides a list with in- and out-of-sample indices per fold used for time series k-fold cross-validation, see Details.

**Usage**

```
create_timefolds(y, k = 5, use_names = TRUE, type = c("extending", "moving"))
```

**Arguments**

<code>y</code>	Any vector of the same length as the data intended to split.
<code>k</code>	Number of folds.
<code>use_names</code>	Should folds be named? Default is TRUE.
<code>type</code>	Should in-sample data be "extending" over the folds (default) or consist of one single fold ("moving")?

**Details**

The data is first partitioned into  $k+1$  sequential blocks  $B_1$  to  $B_{k+1}$ . Each fold consists of two index vectors: one with in-sample row numbers, the other with out-of-sample row numbers. The first fold uses  $B_1$  as in-sample and  $B_2$  as out-of-sample data. The second one uses either  $B_2$  (if `type = "extending"`) or  $B_1, B_2$  (if `type = "moving"`) as in-sample and  $B_3$  as out-of-sample data etc. until the  $k$ th fold with  $B_1, \dots, B_k$  ("extending") resp.  $B_k$  ("moving") as in-sample and  $B_{k+1}$  as out-of-sample data. This makes sure that out-of-sample data always follows in-sample data.

**Value**

A nested list with in-sample ("insample") and out-of-sample ("outsample") indices per fold.

**Examples**

```

y <- runif(100)
create_timefolds(y)
create_timefolds(y, use_names = FALSE)
create_timefolds(y, use_names = FALSE, type = "moving")

```

multi\_strata

*Create Strata from Multiple Features***Description**

This function is a helper function to create strata based on multiple criteria, i.e. the columns of a data.frame, which can then be used as input variable to the splitting functions. Currently, the function offers two strategies to create the strata, see Details.

**Usage**

```
multi_strata(df, strategy = c("kmeans", "interaction"), k = 3L)
```

**Arguments**

df	A data.frame. All columns of the provided data.frame are used to compute strata using the defined strategy (see Details).
strategy	A character. The strategy (either "kmeans" or "interaction") to compute the strata (see Details).
k	An integer. For strategy = "kmeans", it is the desired number of strata, while for strategy = "interaction", it is the approximate number of bins per numeric feature before forming all combinations.

**Details**

The default strategy to turn the columns of the data.frame into a one-dimensional stratification factor is "kmeans": It selects all numeric, logical, character, and factor columns, transforms them into a numeric matrix (ordered factors to integer, unordered factors or categoricals one-hot encoded), scales them and then runs a k-means cluster analysis aiming at k clusters. The strategy "interaction" selects the same columns. Then, it categorizes all numeric columns into approximately k bins. Then, all combinations of all columns are formed by calling `base::interaction()`.

**Value**

A vector containing the strata as a factor that can be passed further on to the splitting functions.

**See Also**

[partition](#), [create\\_folds](#).

**Examples**

```

y_multi <- data.frame(
  A = rep(c(letters[1:4]), each = 20),
  B = factor(sample(c(0, 1), 80, replace = TRUE)),
  c = rnorm(80)
)
y <- multi_strata(y_multi, k = 3)
folds <- create_folds(y, k = 5)

```

partition

*Split Data into Partitions***Description**

This function provides row indices for data splitting, e.g. into training, validation, and test. Different types of split strategies are supported ("basic", "stratified" (the default), "grouped", or "blocked"), see Details. The partition indices are either returned as a list with one element per partition (the default) or as vector of partition ids.

**Usage**

```

partition(
  y,
  p,
  type = c("stratified", "basic", "grouped", "blocked"),
  n_bins = 10,
  split_into_list = TRUE,
  use_names = TRUE,
  shuffle = FALSE,
  seed = NULL
)

```

**Arguments**

<code>y</code>	Either the variable used for "stratification" or "grouped" splits. For other types of splits, any vector of the same length as the data intended to split.
<code>p</code>	A vector with split probabilities per partition, e.g. <code>c(train = 0.7, valid = 0.3)</code> . Names are passed to the output.
<code>type</code>	Split type. One of "stratified", "basic", "grouped", "blocked". The default is "stratified".
<code>n_bins</code>	Approximate numbers of bins for numeric <code>y</code> and <code>type = "stratified"</code> .
<code>split_into_list</code>	Should the resulting partition vector be split into a list or not? Default is <code>TRUE</code> .
<code>use_names</code>	Should names of <code>p</code> be used as partition names? Default is <code>TRUE</code> .
<code>shuffle</code>	Should row indices be randomly shuffled within partition? Default is <code>FALSE</code> . Shuffling is only possible when <code>split_into_list = TRUE</code> .
<code>seed</code>	Integer random seed.

### Details

By default, the function uses stratified splitting. This will balance the partitions as good as possible regarding the distribution of the input vector *y*. Numeric input is first binned into *n\_bins* quantile groups. If *type = "grouped"*, groups specified by *y* are kept together when splitting. This is relevant for clustered or panel data. In contrast to basic splitting, *type = "blocked"* does not sample indices at random, but rather keeps them in groups: e.g. the first 80% of observations form a training set and the remaining 20% are used for testing.

### Value

A list with row indices per partition (if *split\_into\_list = TRUE*) or a vector of partition ids.

### Examples

```
y <- rep(c(letters[1:4]), each = 5)
partition(y, p = c(0.7, 0.3), seed = 1)
partition(y, p = c(0.7, 0.3), split_into_list = FALSE, seed = 1)
p <- c(train = 0.8, valid = 0.1, test = 0.1)
partition(y, p, seed = 1)
partition(y, p, split_into_list = FALSE, seed = 1)
partition(y, p, split_into_list = FALSE, use_names = FALSE, seed = 1)
partition(y, p = c(0.7, 0.3), type = "grouped")
partition(y, p = c(0.7, 0.3), type = "blocked")
```

# Index

`create_folds`, [2](#), [4](#)  
`create_timefolds`, [3](#)  
`multi_strata`, [4](#)  
`partition`, [4](#), [5](#)