Package: hclusteasy (via r-universe)

August 24, 2024

Title Determining Hierarchical Clustering Easily

Version 0.1.0

Description Facilitates hierarchical clustering analysis with functions to read data in 'txt', 'xlsx', and 'xls' formats, apply normalization techniques to the dataset, perform hierarchical clustering and construct scatter plot from principal component analysis to evaluate the groups obtained.

License GPL-2

URL https://github.com/tsukubai/hclusteasy

BugReports https://github.com/tsukubai/hclusteasy/issues

Depends R (>= 3.6)

Imports clusterSim, factoextra, readxl, stats, utils

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.1

Repository https://tsukubai.r-universe.dev

RemoteUrl https://github.com/tsukubai/hclusteasy

RemoteRef HEAD

RemoteSha b156f8e050a446baaa31465ec81bfbf6b2d5a899

Contents

`hca` *Generate and Select Groups with Hierarchical Clustering*

Description

Perform hierarchical clustering and generate groups based on sample dissimilarity using the Euclidean method.

Usage

```
hca(data, method = ”complete”, num.groups = 3)
```

Arguments

`data` Dataset in `data.frame` format.

`method` Method of hierarchical clustering, considering: "ward.D", "ward.D2", "single", "complete", "average" (UPGMA), "mcquitty" (WPGMA), "median" (WPGMC) or "centroid" (UPGMC). Default is "complete".

`num.groups` Number of groups to cut. Default is three.

Value

A vector of integers, where each element represents the group assigned to each observation in the original dataset.

Examples

```
# Load the required package
library(hclusteasy)


# Read the 'iris' dataset from the package
data(”iris_uci”)

# Remove column 'Species' from the iris dataset
iris <- iris_uci[, -5]


# Apply hierarchical cluster and selecting groups
g <- hca(iris)
```

`iris_uci` *Iris Dataset*

Description

This dataset contains 150 flower samples distributed among 3 iris species classes: Setosa, Versicolor, and Virginica. It consists of 5 columns, including 4 attributes measured in centimeters: sepal length and width, and petal length and width, along with a column indicating the iris species. This dataset was introduced by Ronald A. Fisher in 1936 in his classic paper on linear discriminant analysis.

Usage

```
data("iris_uci")
```

Source

https://archive.ics.uci.edu

References

Fisher, R. A. (1988). Iris. UCI Machine Learning Repository. doi:10.24432/C56C76.

`normalization` *Apply Normalization Techniques to the Dataset*

Description

Perform data normalization.

Usage

```
normalization(data, type = "n0", norm = "column", na.remove = FALSE)
```

Arguments

`data` Dataset in `data.frame` format.

`type` Type of normalization. Default is "n1".

- n0: without normalization
- n1: standardization ((x-mean)/sd)
- n2: positional standardization ((x-median)/mad)
- n3: unitization ((x-mean)/range)
- n3a: positional unitization ((x-median)/range)
- n4: unitization with zero minimum ((x-min)/range)

- n5: normalization in range <-1,1> ((x-mean)/max(abs(x-mean)))
- n5a: positional normalization in range <-1,1> ((x-median)/max(abs(x-median)))
- n6: quotient transformation (x/sd)
- n6a: positional quotient transformation (x/mad)
- n7: quotient transformation (x/range)
- n8: quotient transformation (x/max)
- n9: quotient transformation (x/mean)
- n9a: positional quotient transformation (x/median)
- n10: quotient transformation (x/sum)
- n11: quotient transformation (x/sqrt(SSQ))
- n12: normalization ((x-mean)/sqrt(sum((x-mean)^2)))
- n12a: positional normalization ((x-median)/sqrt(sum((x-median)^2)))
- n13: normalization with zero being the central point ((x-midrange)/(range/2))

`norm` Defines whether the normalization will be done by "column" or by "row". Default is "column".

`na.remove` A `logical` value indicating whether NA values should be excluded before performing normalization calculations. Default is FALSE.

Value

Normalized dataset in `data.frame` foramt.

Examples

```
# Load the required package
library(hclusteasy)


# Read the dataset 'iris' from the package
data("iris_uci")

# Remove the column 'Species' from the iris dataset
iris <- iris_uci[, -5]


# Apply normalization to the iris dataset
irisN <- normalization(iris, type = "n1")
```

`pca` *Plot Principal Component Analysis Results*

Description

Apply PCA (Principal Component Analysis) to the data and construct a scatter plot of the first two principal components.

Usage

```
pca(data, groups = "none")
```

Arguments

`data`	Dataset in `data.frame` format.
`groups`	Groups to color observations and draw ellipses around each group of samples with a confidence level of 0.98. Default is "none".

Value

A `ggplot`.

Examples

```
# Load the required package
library(hclusteasy)


# Read the 'iris' dataset from the package
data("iris_uci")

# Select column "Species" (groups) in the iris dataset
species <- iris_uci[, 5]

# Remove column "Species" in the iris dataset
iris <- iris_uci[, -5]


# Apply pca and ploting the two firsts components without groups
pca(iris)

# Apply pca and ploting the first two components with groups
pca(iris, groups = species)
```

`read.data` *Read Files in txt, xls, or xlsx Formats*

Description

Read datasets files in `txt`(space-separated), `xls` or `xlsx` and return the data as a `data.frame`.

Usage

```
read.data(path, col.names = FALSE, col.types = NULL)
```

Arguments

`path` Path to the `txt`(space-separated), `xls` or `xlsx` file.

`col.names` Logical value indicating whether the first row of the dataset should be used as column names. Use `TRUE` to use the first row as column names or `FALSE` otherwise. Default is `FALSE`.

`col.types` Character or a character vector specifying the data types for each column. Possible values are: "skip" , "guess" , "logical" , "numeric", "date" , "text" , or "list" . Default, it is `NULL`, which means the data types will be determined automatically ("guess"). Note that `txt` files do not support the `col.types` parameter.

Value

Dataset in `data.frame` format.

Examples

```
# Load the package
library(hclusteasy)

# Set the file path
file_path <- system.file("extdata",
                          "iris_uci.xlsx",
                           package = "hclusteasy")


# Read a .xlsx dataset
iris <- read.data(file_path,col.names = TRUE)
```

`wine_uci` *Wine Dataset*

Description

It consists of a dataset containing 178 wine samples distributed into 3 distinct classes. It has 14 columns, comprising 13 chemical attributes such as alcohol content, malic acid amount, ash, alkalinity of ash, magnesium, phenols, flavonoids, proanthocyanins, color intensity, hue, OD280/OD315 ratio, and proline, along with one column indicating the wine class. This dataset was introduced by Forina et al. in 1991 in a study on the chemical analysis of wines grown in the regions of Italy.

Usage

```
data("wine_uci")
```

Source

https://archive.ics.uci.edu

References

Aeberhard, Stefan and Forina, M. (1991). Wine. UCI Machine Learning Repository. doi:10.24432/C5PC7J.

Index