Package ‘MachineLearning’

March 15, 2019

Title Machine Learning Algorithms for Innovation in Tourism

Version 0.1.3

Description
A collection of routines created in the collaboration framework in tourism innovation between
The package provides a set of machine learning tools for pattern
detection, association and classification rules and feature selection

Depends R (>= 3.3.0)
Imports NbClust, dplyr, formula.tools, magrittr, arules, rpart,
     FSelectorRcpp, crayon, ggplot2, rpart.plot

License AGPL-3

URL https://datascienceumh.github.io/MachineLearning/

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Suggests spelling, testthat

Language en-GB

NeedsCompilation no

Author Agustin Perez-Martin [aut] (<https://orcid.org/0000-0003-4994-3176>),
     Agustin Perez-Torregrosa [cre, aut]
     (<https://orcid.org/0000-0001-5658-4795>),
     Alejandro Rabasa-Dolado [aut] (<https://orcid.org/0000-0002-6243-9831>),
     Nuria Molla-Campello [aut] (<https://orcid.org/0000-0002-6448-7116>),
     Jesus Javier Rodriguez-Sala [aut]
     (<https://orcid.org/0000-0002-3796-0692>),
     Agencia Valenciana de Turisme [ctb, cph, fnd],
     Miguel Hernandez University [ctb, cph, fnd]

Maintainer Agustin Perez-Torregrosa <agustin.perez01@umh.es>
**Description**

This is a rule-based machine learning method to discover interesting relationships between a consequent and an antecedent (or group of antecedents) in large databases.

**Usage**

```r
AssociationRules(data, support = 0.2, confidence = 0.1, minlength = 2)
```

**Arguments**

- **data**: a data frame with discrete variables.
- **support**: a numeric value for the minimum support of the antecedents (default: 0.2).
- **confidence**: a numeric value for the minimum confidence of confidence in rule/association method (default: 0.8)
- **minlength**: an integer value for the minimal number of items per item set (default: 2 item)

**Value**

A MLA object of subclass Association
Examples

```r
# Load a Dataset
data(EGATUR)
# Generate an association rules with apriori, remember only support discretized variables,
# in this remove numerical variables.
Rules <- AssociationRules(EGATUR[,c(2,4,5,8)])
```

---

CART

Fit and graph a cart model

Description

Classification And Regression Tree is a simple technique to fit a relationship between numerical variables partitioning the target variable by a range of values of the explanatory variables. This function fits and graphs a cart model with a previous separation of training a testing datasets.

Usage

```r
CART(formula, data, p = 0.7, nodes_min = 2, nodes_max = 18,
     includedata = FALSE, seed = NULL, ...)
```

Arguments

- `formula`: a formula of the form `y ~ x1 + x2 + ...`
- `data`: the data frame that contains the variables specified in `formula`.
- `p`: the percentage of the training dataset to be obtained randomly.
- `nodes_min`: Number of minimum nodes.
- `nodes_max`: Number of maximum nodes.
- `includedata`: logicals. If TRUE the training and testing datasets are returned.
- `seed`: a single value, interpreted as an integer, or NULL. The default value is NULL, but for future checks of the model or models generated it is advisable to set a random seed to be able to reproduce it.
- `...`: further arguments passed to or from other methods.

Value

A MLA object of subclass CART

Examples

```r
# Load a Dataset
# Not run:
data(EGATUR)
CART(GastoTotalD~pais+aloja+motivo, data=EGATUR)

# End(Not run)
```
Clustering

A simple and powerful function to create clusters with KMeans

Description

This is a modified kmeans clustering technique to automatize the number of groups or clusters that can be partitioned the sample. Several techniques are used to obtain the best number of clusters.

Usage

```r
Clustering(data, n = "auto", n_max = 10, iter.max = 10,
          auto_criterion = c("explainwss", "db", "ratkowsky", "ball",
                             "friedman"), confidenceWSS = 0.9, agregate_method = median)
```

Arguments

- **data**: Data frame which numeric variables.
- **n**: maximal number of clusters, between 2 and (number of objects - 1), greater or equal to n_min. By default, n_max=10.
- **n_max**: the maximum number of iterations allowed.
- **auto_criterion**: the available criterions are: "explainwss", "db", "ratkowsky", "ball" and "friedman".
- **confidenceWSS**: a confidence interval for criterion WSS.
- **agregate_method**: a function to agregate results of different methods. Default value=median

Details

Several methods are available in order to obtain the best number of clusters: explainwss = Within-cluster Sum of Square db = Davies–Bouldin index (DBI). Davies and Bouldin (1979) ratkowsky = Ratkowsky and Lance (1978) ball = Ball and Hall (1965) friedman = Friedman and Rubin (1967)

@return A MLA object of subclass Clustering

Examples

```r
## Load a Dataset
## Not run:
data(EGATUR)
modelFit <- Clustering(data=EGATUR[,c("A13","gastototal")])

## End(Not run)
```
Description

CREA-RBS is a rule reduction method for allocating a significance value to each rule in the system so that experts may select the rules that should be considered as preferable and understand the exact degree of correlation between the different rule attributes.

Arguments

- **formula**: a formula of the form \( y \sim x_1 + x_2 + \ldots \)
- **data**: the data frame that contains the variables specified in `formula`.

Details

Significance is calculated from the antecedent frequency and rule frequency parameters for each rule; if the first one is above the minimal level and rule frequency is in a critical interval, its significance ratio is computed by the algorithm. These critical boundaries are calculated by an incremental method and the rule space is divided according to them. The significance function is defined for these intervals.

Value

A MLA object of subclass CREA-RBS

References


Examples

```r
## Load a Dataset
data(EGATUR)
## Generate a CREA-RBS model, remember only support discretized variables
CREA.RBS(GastoTotalD=pais+aloja+motivo,data=EGATUR)
```
EGATUR dataset

Description

Tourist Expenditure Survey (EGATUR) is the response by the Spanish Tourist Authorities to the growing need for information by this sector.

Usage

data("EGATUR")

Format

A data frame with 30541 observations on the following 13 variables.

- mm_aaaa a numeric vector
- pais a factor with country names
- A13 a numeric vector
- aloja a factor with levels Hotels Rest of market Over-The-Counter Accommodation
- motivo a factor with levels Leisure Business Others
- gastototal a numeric vector
- factoregatur a numeric vector
- GastoTotalD a factor with levels [17,1.67e+03] [1.67e+03,5.51e+03] [5.51e+03,1.99e+04]
- A13_D a factor with levels [1,3] [3,4] [4,6] [6,7] [7,14] [14,180]

Details

Tourist Expenditure Survey (EGATUR) is the response by the Spanish Tourist Authorities to the growing need for information by this sector, which is one of the major driving forces of the Spanish economy. The information provided by this survey makes it possible to ascertain with a greater degree of precision the volume of tourist expenditure by foreign visitors coming to Spain each month by different concepts, and to also analyse key aspects of their tourist behaviour. EGATUR makes it possible to improve strategic knowledge of variables regarding fundamental expenditure and tourist behaviour by visitors from other countries, and to a large extent compensate the loss of information which, in order to estimate the income and payment entries for tourism in the Balance of Payments, was being used by the Bank of Spain prior to the introduction of the Euro. Likewise it provides highly relevant information in terms of National Accounts estimates and, in particular, in estimating the main groups of the recent Tourism Satellite Account for Spain.

References


Examples

data(EGATUR)
### plot.MLA

**Plot MLA object**

**Description**

This function plot an MLA object. It is a method for the generic function `plot`.

**Usage**

```r
### S3 method for class 'MLA'
plot(x, simply = FALSE, ...)
```

**Arguments**

- **x** object of class "MLA"
- **simply** Allow to simplify the view of nodes, in case of MLA object is a CART. Default value is FALSE.
- **...** further arguments passed to or from other methods.

### plotCART

**Plot rpart or MLA object**

**Description**

This function plot an MLA object or a rpart.

**Usage**

```r
plotCART(x, ownlabs = TRUE)
```

**Arguments**

- **x** object of class "MLA" or "rpart".
- **ownlabs** Allow to simplify the view of nodes. Default value is TRUE.
- **...** further arguments passed to or from other methods.
print.MLA  

*Print an MLA Object*

**Description**

This function prints an MLA object. It is a method for the generic function print.

**Usage**

```r
## S3 method for class 'MLA'
print(x, first = 100, digits = getOption("digits"), ...)
```

**Arguments**

- **x** object of class "MLA"
- **first** When the print command shows a set of rules limits the number of rules.
- **digits** minimal number of significant digits.
- **...** further arguments passed to or from other methods.

---

`sampler`

*Splitting your dataset in training and testing*

**Description**

A training/test partition are created by `sampler` function.

**Usage**

```r
sampler(data, p, seed = NULL)
```

**Arguments**

- **data** the data frame that contains the variables to be separated.
- **p** the percentage of the training dataset to be obtained randomly. It can be expressed in either decimal fraction (such as 0.7) or percent (such as 72.12).
- **seed** a single value, interpreted as an integer, or NULL. The default value is NULL, but for future checks of the model or models generated it is advisable to set a random seed to be able to reproduce it.

**Examples**

```r
# The best way to demostrate the functionality is test the function
Sampling <- sampler(EGATUR, p=0.7)
```
VariableRanker  

Ranks of importance variables

Description
A Ranker of variables

Usage
VariableRanker(formula, data, based = "gainratio", ...)

Arguments

formula    a formula of the form y ~ x1 + x2 + ...
data       the data frame that contains the variables specified in formula.
based      methodology used to rank variables. The options available are informationgain, gainratio and symmetrical.uncertainty.
...        further arguments passed to or from other methods.

Value
A MLA object of subclass Var-Rank

Examples

## Load a Dataset
data(EGATUR)
VariableRanker(formula=GastoTotalD~pais+aloja+motivo,EGATUR)
Index

*Topic datasets
  EGATUR, 6

AssociationRules, 2

CART, 3
Clustering, 4
CREA.RBS, 5

EGATUR, 6
plot.MLA, 7
plotCART, 7
print.MLA, 8

sampler, 8

VariableRanker, 9