Package 'structree'

June 5, 2020

Type Package

Title Tree-Structured Clustering

Version 1.1.7

Date 2020-06-04

Author Moritz Berger

Depends mgcv, lme4, penalized

Maintainer Moritz Berger <Moritz.Berger@imbie.uni-bonn.de>

Description Tree-
structured modelling of categorical predictors (Tutz and Berger (2018), <doi:10.1007/s11634-017-0298-6>) or measurement
units (Berger and Tutz (2018), <doi:10.1080/10618600.2017.1371030>).

License GPL-3

LazyLoad yes

RoxygenNote 7.1.0

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2020-06-05 09:40:02 UTC

R topics documented:

CTB *Achievement Test from CTB/McGraw-Hill*

Description

The data set contains results of an achievement test that measures different objectives and subskills of subjects in mathematics and science. Inter alia, the students had to respond to 56 multiple-choice items (31 mathematics, 25 science). For the original description, see Section 5.6 of Chapter 5 in De Boeck and Wilson (2004).

Usage

```
data(CTB)
```

Format

A data frame containing 1211 observations on 9 variables:

score number of correctly solved items (metric)

school school ID (nominal)

size number of students in the school, in hundreds (metric)

bachelor transformed and standardized percentage of adults with BA degree or higher in area with school zip code (metric)

born transformed and standardized percentage of adults in the school area who were born in the state where they now reside (metric)

mortgage transformed and standardized median of the monthly mortgage in the school area (metric)

language transformed and standardized percentage of foreign language households in the school area (metric)

type type of school (1: catholic, 2: private, 3: public)

gender gender (0: male, 1: female)

References

De Boeck, P. and M. Wilson (2004). Explanatory item response models: A generalized linear and nonlinear approach. Springer Verlag.

Examples

```
data(CTB)

y <- CTB$score
x <- CTB$gender

hist(y)
table(x)
```

guPrenat *Prenatal Care in Guatemala*

Description

A data set derived from the National Survey of Maternal and Child Health in Guatemala in 1987. The data contains observations of children that were born in the 5-year period before the survey.

Usage

```
data(guPrenat)
```

Format

A data frame containing 1211 observations on 9 variables:

cluster community (nominal)

prenat prenatal care (0: traditional, 1: modern)

motherAge mother 25 years or older (0: no, 1: yes)

indig mother's ethnicity (nominal)

momEd mother's level of education (nominal)

husEd husband's level of education (nominal)

husEmpl husband's employment status (nominal)

toilet modern toilet in house (0: no, 1: yes)

TV frequency of TV usage (nominal)

References

Rodriguez, Germa'n and Goldman, Noreen (1995), "Improved estimation procedures for multilevel models with binary response: a case-study", Journal of the Royal Statistical Society, Series A, 164, 339-355.

Douglas Bates and Martin Maechler and Ben Bolker (2014). mlmRev: Examples from Multilevel Modelling Software Review. R package version 1.0-6. https://CRAN.R-project.org/package=mlmRev

Examples

```
data(guPrenat)

y <- guPrenat$prenat
community <- guPrenat$cluster

table(y)
hist(table(community))
```

plot.structree *Plotting Results of Tree-Structured Clustering*

Description

Takes a fitted structree object and plots the results of the tree component of the model.

Usage

```
## S3 method for class 'structree'
plot(x, select = NULL, paths = FALSE,
  result = FALSE, ask = FALSE, xlab = NULL, ylab = NULL,
  main = NULL, lwd = 1, cex.txt = 1, cex.axis = 1, cex.lab = 1,
  cex.main = 1, ...)
```

Arguments

x	Object of class structree.
select	Elements of the tree component that are plotted; if select is not specified, by default all components are pictured in one plot.
paths	If true, the coefficient paths are plotted.
result	If true, the resulting partition is displayed.
ask	If true, each element chosen by select is plotted seperately.
xlab	Label of x-axis.
ylab	Label of y-axis.
main	Title of the plot.
lwd	Linewidth.
cex.txt	Size of the text.
cex.axis	Size of the axis.
cex.lab	Size of the labels.
cex.main	Size of title.
...	Further arguments passed to or from other methods.

Details

By default the function pictures the estimated trees against all splits. If select=NULL the trees for all the predictors will be plotted.

Author(s)

Moritz Berger <Moritz.Berger@imbie.uni-bonn.de>
http://www.imbie.uni-bonn.de/personen/dr-moritz-berger/

References

Tutz, Gerhard and Berger, Moritz (2018): Tree-structured modelling of categorical predictors in regression, Advances in Data Analysis and Classification 12(3), 737-758.

Berger, Moritz and Tutz, Gerhard (2018): Tree-structured clustering in fixed effects models, Journal of Computational and Graphical Statistics 27(2), 380-392.

See Also

structree

Examples

```
data(rent)

## Not run:
mod <- structree(nmqm~tr(bez)+tr(bj)+tr(rooms)+badkach0,data=rent,
                 family=gaussian,stop_criterion="CV")

plot(mod, paths=TRUE)

## End(Not run)
```

---

rent *Munich Rent Data*

Description

The data set is part of the Munich rent index in 2003. It is available from the data archive of the Department of Statistics at the University of Munich (http://www.statistik.lmu.de/service/datenarchiv).

Usage

```
data(rent)
```

Format

A data frame containing 2053 observations on 11 variables:

nmqm net rent per square meter (metric)

wfl floor space (metric)

rooms number of rooms (ordinal)

bj year of construction (ordinal)

bez residential area (norminal)

ww0 hot water supply (1: no, 0: yes)

zh0 central heating (1: no, 0: yes)

badkach0 tiled bathroom (1: no, 0: yes)

badextra supplementary equipment in bathroom (1: yes, 0: no)

kueche well equipped kitchen (1: yes, 0: no)

quality quality of residential area (ordinal)

References

Fahrmeir, L. and Kuenstler, R. and Pigeot, I. and Tutz, G. (2004): Statistik: der Weg zur Datenanalyse. 5. Auflage, Springer, Berlin.

Examples

```
data(rent)

y <- rent$nmqm
X <- rent[,-1]

boxplot(y)
summary(X)
```

`structree` *Tree-Structured Clustering*

Description

Fusion of categories of ordinal or nominal predictors or fusion of measurement units by tree-structured clustering.

Usage

```
structree(formula, data, family = gaussian, stop_criterion = c("AIC",
  "BIC", "CV", "pvalue"), splits_max = NULL, fold = 5, alpha = 0.05,
  grid_value = NULL, min_border = NULL, ridge = FALSE,
  constant_covs = FALSE, trace = TRUE, plot = TRUE, k = 10,
  weights = NULL, offset = NULL, ...)

## S3 method for class 'structree'
print(x, ...)

## S3 method for class 'structree'
coef(object, ...)
```

Arguments

formula	Object of class formula: a symbolic description of the model to be fitted. See detail.
data	Data.frame of class data.frame containing the variables of the model.
family	a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. See family for details of family functions.
stop_criterion	Criterion to determine the optimal number of splits in the tree component of the model; one out of "AIC", "BIC", "CV" and "pvalue".
splits_max	Maximal number of splits in the tree component.
fold	Number of folds; only for stop criterion "CV".
alpha	Significance level; only for stop criterion "pvalue".
grid_value	An optional parameter; grid_value is a scalar giving the minimal distance between two adjacent observation units that are used as candidates for splitting; only for repeated measurements.
min_border	An optional parameter; min_border is a integer giving the minimal size of the outer nodes of the tree; only for repeated measurements.
ridge	If true, a small ridge penalty is added to obtain the order of measurement units; only for repeated measurements.
constant_covs	Must be set to true, if constant covariates are available; only for repeated measurments (currently only available for Gaussian response).
trace	If true, information about the estimation progress is printed.
plot	If true, the smooth components of the model are plottet; only for categorical predictors.
k	Dimension of the B-spline basis that is used to fit smooth components. For details see s; only for categorical predictors.
weights	An optional vector of prior weights to be used in the fitting process; see also glm.
offset	An a priori known component to be included in the linear predictor during fitting; see also glm.
...	Further arguments passed to or from other methods.
x, object	Object of class "structree".

Details

A typical formula has the form response ~ predictors, where response is the name of the response variable and predictors is a series of terms that specify the predictor of the model.

For an ordinal or nominal predictors z one has to enter tr(x) into the formula.

For smooth components x one has to enter s(x) into the formula; currently not implemented for repeated measurements.

For fixed effects z of observation units u one has to enter tr(z|u) into the formula. An unit-specific intercept is specified by tr(1|u).

The framework only allows for categorical predictors or observations units in the tree component, but not both. All other predictors with a linear term are entered as usual by x1+...+xp.

Value

Object of class "structree". An object of class "structree" is a list containing the following components:

coefs_end	all coefficients of the estimated model
partitions	list of matrices containing the partitions of the predictors in the tree component including all iterations
beta_hat	list of matrices with the fitted coefficients in the tree component including all iterations
which_opt	number of the optimal model (total number of splits-1)
opts	number of splits per predictor in the tree component
order	list of ordered split-points of the predictors in the tree component
tune_values	value of the stopping criterion that determine the optimal model
group_ID	list of the group IDs for each observations
coefs_group	list of coefficients of the estimated model
y	Response vector
DM_kov	Design matrix

Author(s)

Moritz Berger <Moritz.Berger@imbie.uni-bonn.de>
http://www.imbie.uni-bonn.de/personen/dr-moritz-berger/

References

Tutz, Gerhard and Berger, Moritz (2018): Tree-structured modelling of categorical predictors in regression, Advances in Data Analysis and Classification 12(3), 737-758.

Berger, Moritz and Tutz, Gerhard (2018): Tree-structured clustering in fixed effects models, Journal of Computational and Graphical Statistics 27(2), 380-392.

See Also

plot.structree

Examples

```
data(rent)

## Not run:
mod <- structree(nmqm~tr(bez)+tr(bj)+tr(rooms)+badkach0,data=rent,
                 family=gaussian,stop_criterion="CV")

print(mod)
coef(mod)

## End(Not run)
```

Index