Package ‘dummy’

July 22, 2025

Type Package

Title Automatic Creation of Dummies with Support for Predictive Modeling

Version 0.1.3

Date 2015-05-07

Author Michel Ballings and Dirk Van den Poel

Maintainer Michel Ballings <michel.ballings@GMail.com>

Description Efficiently create dummies of all factors and character vectors in a data frame. Support is included for learning the categories on one data set (e.g., a training set) and deploying them on another (e.g., a test set).

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2015-05-07 18:12:24

Contents

`categories` *Extraction of Categorical Values as a Preprocessing Step for Making Dummy Variables*

Description

`categories` stores all the categorical values that are present in the factors and character vectors of a data frame. Numeric and integer vectors are ignored. It is a preprocessing step for the `dummy` function. This function is appropriate for settings in which the user only wants to compute dummies for the categorical values that were present in another data set. This is especially useful in predictive modeling, when the new (test) data has more or other categories than the training data.

Usage

```
categories(x, p = "all")
```

Arguments

x	data frame containing factors or character vectors that need to be transformed to dummies. Numerics, dates and integers will be ignored.
p	select the top p values in terms of frequency. Either "all" (all categories in all variables), an integer scalar (top p categories in all variables), or a vector of integers (number of top categories per variable in order of appearance.

Value

A list containing the variable names and the categories

Author(s)

Authors: Michel Ballings, and Dirk Van den Poel, Maintainer: <Michel.Ballings@GMail.com>

See Also

dummy

Examples

```
#create toy data
(traindata <- data.frame(var1=as.factor(c("a","b","b","c")),
                         var2=as.factor(c(1,1,2,3)),
                         var3=c("val1","val2","val3","val3"),
                         stringsAsFactors=FALSE))
(newdata <- data.frame(var1=as.factor(c("a","b","b","c","d","d")),
                       var2=as.factor(c(1,1,2,3,4,5)),
                       var3=c("val1","val2","val3","val3","val4","val4"),
                       stringsAsFactors=FALSE))

categories(x=traindata,p="all")
categories(x=traindata,p=2)
categories(x=traindata,p=c(2,1,3))
```

dummy *Automatic Dummy Variable Creation with Support for Predictive Contexts*

Description

dummy creates dummy variables of all the factors and character vectors in a data frame. It also supports settings in which the user only wants to compute dummies for the categorical values that were present in another data set. This is especially useful in the context of predictive modeling, in which the new (test) data has more or other categories than the training data.

Usage

```
dummy(x, p = "all", object = NULL, int = FALSE, verbose = FALSE)
```

Arguments

`x` a data frame containing at least one factor or character vector

`p` Only relevant if object is NULL. Select the top p values in terms of frequency. Either "all" (all categories in all variables), an integer scalar (top p categories in all variables), or a vector of integers (number of top categories per variable in order of appearance).

`object` output of the `categories` function. This parameter is to be used when dummies should be created only of categories present in another data set (e.g., training set)

`int` should the dummies be integers (TRUE) or factors (FALSE)

`verbose` logical. Used to show progress

Value

A data frame containing dummy variables

Author(s)

Authors: Michel Ballings, and Dirk Van den Poel, Maintainer: <`Michel.Ballings@GMail.com`>

See Also

`categories`

Examples

```
#create toy data
(traindata <- data.frame(var1=as.factor(c("a","b","b","c")),
                         var2=as.factor(c(1,1,2,3)),
                         var3=c("val1","val2","val3","val3"),
                         stringsAsFactors=FALSE))
(newdata <- data.frame(var1=as.factor(c("a","b","b","c","d","d")),
                       var2=as.factor(c(1,1,2,3,4,5)),
                       var3=c("val1","val2","val3","val3","val4","val4"),
                       stringsAsFactors=FALSE))
#create dummies of training set
(dummies_train <- dummy(x=traindata))
#create dummies of new set
(dummies_new <- dummy(x=newdata))

#how many new dummy variables should not have been created?
sum(! colnames(dummies_new) %in% colnames(dummies_train))

#create dummies of new set using categories found in training set
(dummies_new <- dummy(x=newdata,object=categories(traindata,p="all")))
```

```
#how many new dummy variables should not have be created?
sum(! colnames(dummies_new) %in% colnames(dummies_train))


#create dummies of training set,
#using the top 2 categories of all variables found in the training data
dummy(x=traindata,p=2)

#create dummies of training set,
#using respectively the top 2,3 and 1 categories of the three
#variables found in training data
dummy(x=traindata,p=c(2,3,1))

#create all dummies of training data
dummy(x=traindata)
```

dummyNews *Display the NEWS file*

Description

dummyNews shows the NEWS file of the dummy package.

Usage

```
dummyNews()
```

Author(s)

Authors: Michel Ballings and Dirk Van den Poel, Maintainer: <Michel.Ballings@GMail.com>

Examples

```
dummyNews()
```

Index