Package 'catSplit'

October 20, 2021

Type Package

Title Encode Categorical Variables with Split Information from CART

Version 0.1.0

Description
Use primary and surrogate split information from CART (Classification and Regression Trees - Breiman L (1984)) as the
vector representation for a categorical variable. Outputs binary
columns for each categorical variable making use of target information.

License MIT + file LICENSE

Encoding UTF-8

Imports rpart, caret, dplyr, stats, utils, data.table, stringr,
OpenML, farff

RoxygenNote 7.1.2

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Mine Gazioglu [aut, cre],
Mustafa Baydogan [ctb]

Maintainer Mine Gazioglu <mine.gazioglu40@gmail.com>

Repository CRAN

Date/Publication 2021-10-20 11:50:02 UTC

R topics documented:

catSplitEncoding *Encode categorical variables using split information of CART*

Description

Encode categorical variables using split information of CART

Usage

```
catSplitEncoding(
  targetVariable,
  trainData,
  testData,
  problemType,
  datasetName,
  catVariables
)
```

Arguments

targetVariable	target variable that we want to predict.
trainData	training data.
testData	testing data.
problemType	classification or regression.
datasetName	Name of the dataset, could be any string name.
catVariables	List of categorical variables in the dataset.

Value

dataframe that is the encoding of categorical variables.

Examples

```
library("OpenML")
library("farff")
library("stringr")
library("stats")
library("data.table")
library("rpart")
library("catSplit")



# An example dataset from OpenML
datInfo <- getOMLDataSet(data.id = 41283, verbosity = 0)
```

```
targetVariable <- datInfo$target.features
dat <- datInfo$data
datasetName <- datInfo$desc$name
catVariables <- names(Filter(is.factor, dat))
# Remove target variable from catVariables
catVariables <- catVariables[!(catVariables %in% targetVariable)]
problemType <- "classification"
# Split dat to train and test sets
smp_size <- floor(0.75 * nrow(dat))
train_ind <- sample(seq_len(nrow(dat)), size = smp_size)
train <- as.data.frame.matrix(dat[train_ind, ])
test <- as.data.frame.matrix(dat[-train_ind, ])
# Outputs a list containing 2 files: encoding frame for train data, encoding frame for test data
train_and_test_cat = catSplitEncoding(targetVariable = targetVariable,
                                                    trainData = train,
                                                    testData = test,
                                                    problemType = problemType,
                                                    datasetName = datasetName,
                                                    catVariables = catVariables)
# Get transformed train and test sets from the output list
trainCat = train_and_test_cat[1]
testCat = train_and_test_cat[2]

# Drop categorical variables from the original train and test data
trainData <- train[!names(train) %in% catVariables]
testData <- test[!names(test) %in% catVariables]

# Merge encoding frame and original data
train <- cbind(trainCat, trainData)
test <- cbind(testCat, testData)
```

Index