

Package ‘missRanger’

June 14, 2019

Title Fast Imputation of Missing Values

Version 2.0.1

Description Alternative implementation of the beautiful 'MissForest' algorithm used to impute mixed-type data sets by chaining random forests, introduced by Stekhoven, D.J. and Bühlmann, P. (2012) <doi:10.1093/bioinformatics/btr597>. Under the hood, it uses the lightning fast random jungle package 'ranger'. Between the iterative model fitting, we offer the option of using predictive mean matching. This firstly avoids imputation with values not already present in the original data (like a value 0.3334 in 0-1 coded variable). Secondly, predictive mean matching tries to raise the variance in the resulting conditional distributions to a realistic level. This would allow e.g. to do multiple imputation when repeating the call to missRanger().

Depends R (>= 3.5.0)

License GPL(>= 2)

Encoding UTF-8

LazyData true

Type Package

Date 2019-06-14

Imports stats, FNN (>= 1.1), ranger (>= 0.10)

Author Michael Mayer [aut, cre, cph]

Maintainer Michael Mayer <mayermichael79@gmail.com>

RoxygenNote 6.1.1

NeedsCompilation no

R topics documented:

allVarsTwoSided	2
generateNA	2
imputeUnivariate	3
missRanger	3
pmm	5

Index	6
--------------	----------

allVarsTwoSided	<i>Extraction of Variable Names from Two-Sided Formula.</i>
-----------------	---

Description

Takes a formula and a data frame and returns all variable names in both the lhs and the rhs. lhs and rhs are evaluated separately. This is relevant if both sides contain a "." (= all variables).

Usage

```
allVarsTwoSided(formula, data)
```

Arguments

formula	A two-sided formula object.
data	A data.frame. Primarily used to deal with "." in the formula.

Value

A list with two character vectors of variable names.

Examples

```
allVarsTwoSided(Species + Sepal.Width ~ Petal.Width, iris)
allVarsTwoSided(. ~ ., iris)
allVarsTwoSided(.~Species ~ Sepal.Width, iris)
allVarsTwoSided(. ~ Sepal.Width, iris)
```

generateNA	<i>Adds Missing Values to a Data Set</i>
------------	--

Description

Takes a data frame and replaces randomly part of the values by missing values.

Usage

```
generateNA(data, p = 0.1, seed = NULL)
```

Arguments

data	A data.frame.
p	Proportion of missing values to approximately add to each column of data.
seed	An integer seed.

Value

data with missing values.

Examples

```
head(generateNA(iris))
```

imputeUnivariate	<i>Univariate Imputation</i>
------------------	------------------------------

Description

Fills missing values of a vector of any type by sampling with replacement from the non-missing values. Requires at least one non-missing value to run.

Usage

```
imputeUnivariate(x, seed = NULL)
```

Arguments

x	A vector of any type possibly containing missing values.
seed	An integer seed.

Value

A vector of the same length and type as x but without missing values.

Examples

```
imputeUnivariate(c(NA, 0, 1, 0, 1))
imputeUnivariate(c("A", "A", NA))
imputeUnivariate(as.factor(c("A", "A", NA)))

# Impute a whole data set univariately
ir <- generateNA(iris)
head(imputed <- do.call(data.frame, lapply(ir, imputeUnivariate)))
```

missRanger	<i>Fast Imputation of Missing Values by Chained Random Forests</i>
------------	--

Description

Uses the "ranger" package [1] to do fast missing value imputation by chained random forests, see [2] and [3]. Between the iterative model fitting, it offers the option of predictive mean matching. This firstly avoids imputation with values not present in the original data (like a value 0.3334 in a 0-1 coded variable). Secondly, predictive mean matching tries to raise the variance in the resulting conditional distributions to a realistic level and, as such, allows to do multiple imputation when repeating the call to missRanger(). The iterative chaining stops as soon as maxiter is reached or if the average out-of-bag estimate of performance stops improving. In the latter case, except for the first iteration, the second last (i.e. best) imputed data is returned.

Usage

```
missRanger(data, formula = . ~ ., pmm.k = 0L, maxiter = 10L,
  seed = NULL, verbose = 1, returnOOB = FALSE, case.weights = NULL,
  ...)
```

Arguments

<code>data</code>	A <code>data.frame</code> or <code>tibble</code> with missing values to impute.
<code>formula</code>	A two-sided formula specifying variables to be imputed (left hand side) and variables used to impute (right hand side). Defaults to <code>. ~ .</code> , i.e. use all variables to impute all variables. If e.g. all variables (with missings) should be imputed by all variables except variable "ID", use <code>. ~ . - ID</code> . Note that a "." is evaluated separately for both sides of the formula. Note that variables with missings must appear in the left hand side if they should be used on the right hand side.
<code>pmm.k</code>	Number of candidate non-missing values to sample from in the predictive mean matching step. 0 to avoid this step.
<code>maxiter</code>	Maximum number of chaining iterations.
<code>seed</code>	Integer seed to initialize the random generator.
<code>verbose</code>	Controls how much info is printed to screen. 0 to print nothing. 1 (default) to print a "." per iteration and variable, 2 to print the OOB prediction error per iteration and variable (1 minus R-squared for regression).
<code>returnOOB</code>	Logical flag. If TRUE, the final average out-of-bag prediction error is added to the output as attribute "oob".
<code>case.weights</code>	Vector with weight per observation in the data set used in fitting the random forests.
<code>...</code>	Arguments passed to <code>ranger</code> . If the data set is large, better use less trees (e.g. <code>num.trees = 100</code>) and/or a low value of <code>sample.fraction</code> . The following arguments are incompatible: <code>data</code> , <code>write.forest</code> , <code>probability</code> , <code>split.select.weights</code> , <code>dependent.variable.name</code> , and <code>classification</code> .

Value

An imputed `data.frame`.

References

- [1] Wright, M. N. & Ziegler, A. (2016). `ranger`: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, in press. <http://arxiv.org/abs/1508.04409>.
- [2] Stekhoven, D.J. and Bühlmann, P. (2012). 'MissForest - nonparametric missing value imputation for mixed-type data', *Bioinformatics*, 28(1) 2012, 112-118. <https://doi.org/10.1093/bioinformatics/btr597>.
- [3] Van Buuren, S., Groothuis-Oudshoorn, K. (2011). `mice`: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1-67. <http://www.jstatsoft.org/v45/i03/>

Examples

```
irisWithNA <- generateNA(iris)
irisImputed <- missRanger(irisWithNA, pmm.k = 3, num.trees = 100)
head(irisImputed)
head(irisWithNA)

# With extra trees algorithm
irisImputed_et <- missRanger(irisWithNA, pmm.k = 3, num.trees = 100, splitrule = "extratrees")
head(irisImputed_et)

# Do not impute Species. Note: Since this variable contains missings, it cannot be used
# to impute the other variables as well.
```

```
irisImputed <- missRanger(irisWithNA, . ~ Species ~ ., pmm.k = 3, num.trees = 100)

# Impute univariately only.
irisImputed <- missRanger(irisWithNA, . ~ 1)

# Use Species and Petal.Length to impute Species and Petal.Length.
irisImputed <- missRanger(irisWithNA, Species + Petal.Length ~ Species + Petal.Length,
  pmm.k = 3, num.trees = 100)
```

pmm

Predictive Mean Matching

Description

This function is used internally only but might help others to implement an efficient way of doing predictive mean matching on top of any prediction based missing value imputation. It works as follows: For each predicted value of a vector `xtest`, the closest `k` predicted values of another vector `xtrain` are identified by `k`-nearest neighbour. Then, one of those neighbours is randomly picked and its corresponding observed value in `ytrain` is returned.

Usage

```
pmm(xtrain, xtest, ytrain, k = 1L, seed = NULL)
```

Arguments

<code>xtrain</code>	Vector with predicted values in the training data set.
<code>xtest</code>	Vector with predicted values in the test data set.
<code>ytrain</code>	Vector with observed response in the training data set.
<code>k</code>	Number of nearest neighbours to choose from. Set <code>k = 0</code> if no predictive mean matching is to be done.
<code>seed</code>	Integer random seed.

Value

Vector with predicted values in the test data set based on predictive mean matching.

Examples

```
pmm(xtrain = c(0.2, 0.2, 0.8), xtest = 0.3, ytrain = c(0, 0, 1), k = 1) # 0
pmm(xtrain = c(0.2, 0.2, 0.8), xtest = 0.3, ytrain = c(0, 0, 1), k = 3) # 0 or 1
pmm(xtrain = c("A", "A", "B"), xtest = "B", ytrain = c("B", "A", "B"), k = 1) # B
pmm(xtrain = c("A", "A", "B"), xtest = "B", ytrain = c("B", "A", "B"), k = 2) # A or B
```

Index

`allVarsTwoSided`, 2

`generateNA`, 2

`imputeUnivariate`, 3

`missRanger`, 3

`pmm`, 5