Package 'MulvariateRandomForestVarImp'

October 12, 2022

Title Variable Importance Measures for Multivariate Random Forests

Version 0.0.2

Description Calculates two sets of post-hoc variable importance measures for multivariate random forests. The first set of variable importance measures are given by the sum of mean split improvements for splits defined by feature j measured on user-defined examples (i.e., training or testing samples). The second set of importance measures are calculated on a per-outcome variable basis as the sum of mean absolute difference of node values for each split defined by feature j measured on user-defined examples (i.e., training or testing samples). The user can optionally threshold both sets of importance measures to include only splits that are statistically significant as measured using an F-test.

License GPL (>= 3)

Encoding UTF-8

RoxygenNote 7.1.2

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Imports MultivariateRandomForest (>= 1.1.5), MASS (>= 7.3.0)

URL https://github.com/Megatvini/VIM/

BugReports https://github.com/Megatvini/VIM/issues

Depends R (>= 2.10)

NeedsCompilation no

Author Sikdar Sharmistha [aut],
Hooker Giles [aut],
Kadiyali Vrinda [ctb],
Dogonadze Nika [cre]

Maintainer Dogonadze Nika <nika.dogonadze@toptal.com>

Repository CRAN

Date/Publication 2021-12-15 10:40:05 UTC

R topics documented:

`MeanOutcomeDifference` *Mean Outcome Difference Importance Function*

Description

Mean Outcome Difference Importance Function

Usage

```
MeanOutcomeDifference(
  X,
  Y,
  sample_size = trunc(nrow(X) * 0.8),
  num_trees = 100,
  m_feature = ncol(X),
  min_leaf = 10,
  alpha_threshold = 0
)
```

Arguments

`X`	Feature matrix
`Y`	Target matrix
`sample_size`	Size of random subset for each tree generation
`num_trees`	Number of Trees to generate
`m_feature`	Number of randomly selected features considered for a split in each regression tree node, which must be positive integer and less than N (number of input features)
`min_leaf`	Minimum number of samples in the leaf node. If a node has less than or equal to min_leaf samples, then there will be no splitting in that node and this node will be considered as a leaf node. Valid input is positive integer, which is less than or equal to M (number of training samples)
`alpha_threshold`	threshold for split significant testing. If default value of 0 is specified, all the node splits will contribute to result, otherwise only those splits with improvement greater than 1-alpha critical value of an f-statistic do.

Details

For each split defined by feature j, the mean outcome difference importance function calculates the absolute difference in mean values per outcome between the left and right children nodes of the resultant split. With a multivariate outcome vector, this measure thus gives a vector of importance measures for feature j, i.e., it returns an outcome specific importance measure for feature j. If feature j is used in splitting M nodes of the tree, the resulting tree-specific importance measure is the sum of the node-specific absolute differences in mean nodal values per outcome calculated across all M nodes. For the multivariate random forest, the mean outcome difference importance measure for feature j is the average of the tree-specific measures across all trees in the forest.

If the alpha threshold is 0 all the splits defined by feature j will be used in computing the importance measure. The user also has the option of including only the significant node splits defined by feature j in the calculation of the importance measure. The significance of node splits is measured using an F-test. In this case, the user will matrix and the number of left and right node samples for the given node split.

Segal MR (1992) Tree-structured methods for longitudinal data. J. American Stat. Assoc. 87(418), 407-418.

Value

Vector of size N x 1

Examples

```
X = matrix(runif(50*5), 50, 5)
Y = matrix(runif(50*2), 50, 2)
MeanOutcomeDifference(X, Y)
```

`MeanSplitImprovement` *Mean Split Improvement Importance Function*

Description

Mean Split Improvement Importance Function

Usage

```
MeanSplitImprovement(
  X,
  Y,
  sample_size = trunc(nrow(X) * 0.8),
  num_trees = 100,
  m_feature = ncol(X),
  min_leaf = 10,
  alpha_threshold = 0
)
```

Arguments

X	Feature matrix
Y	Target matrix
sample_size	Size of random subset for each tree generation
num_trees	Number of Trees to generate
m_feature	Number of randomly selected features considered for a split in each regression tree node, which must be positive integer and less than N (number of input features)
min_leaf	Minimum number of samples in the leaf node. If a node has less than or equal to min_leaf samples, then there will be no splitting in that node and this node will be considered as a leaf node. Valid input is positive integer, which is less than or equal to M (number of training samples)
alpha_threshold	threshold for split significant testing. If default value of 0 is specified, all the node splits will contribute to result, otherwise only those splits with improvement greater than 1-alpha critical value of an f-statistic do.

Details

The mean split improvement importance function follows directly from Segal (1992) definition of the mean structure based split function. For each split defined by feature j, it calculates the difference between the within parent node sum of squares and the within children-nodes (left and right nodes) measured on either training or testing samples. If feature j is used in splitting M nodes of the tree, the resulting tree-specific importance measure is the sum of the node-specific differences calculated across all M nodes. The mean split improvement measure for feature j for the multivariate random forest is the average of the tree-specific measures across all trees in the forest.

If the alpha threshold is 0 all the splits defined by feature j will be used in computing the importance measure. The user also has the option of including only the significant node splits defined by feature j in the calculation of the importance measure. The significance of node splits is measured using an F-test. In this case, the user will need to threshold the alpha critical value of the F-statistic based on the number of outcome variables in the target matrix and the number of left and right node samples for the given node split.

Segal MR (1992) Tree-structured methods for longitudinal data. J. American Stat. Assoc. 87(418), 407-418.

Value

Vector of size N x 1

Examples

```
X = matrix(runif(50*5), 50, 5)
Y = matrix(runif(50*2), 50, 2)
MeanSplitImprovement(X, Y)
```

Index