Package 'CASMI'

October 31, 2022

Type Package

Title 'CASMI'-Based Functions

Version 1.0.0

Description Contains Coverage Adjusted Standardized Mutual Information ('CASMI')-based functions. 'CASMI' is a fundamental concept of a series of methods. For more information about 'CASMI' and 'CASMI'-related methods, please refer to the corresponding publications (for example, a feature selection method, Shi, J., Zhang, J., & Ge, Y. (2019) <doi:10.3390/e21121179>, and a dataset quality measurement method, Shi, J., Zhang, J., & Ge, Y. (2019) <doi:10.1109/ICHI.2019.8904553>) or contact the package author.

Imports EntropyEstimation, entropy, stats

License GPL-3

Encoding UTF-8

RoxygenNote 7.2.1

NeedsCompilation no

Author Jingyi (Catherine) Shi Developer [aut, cre, cph],
Jialin Zhang Dev [ctb]

Maintainer Jingyi (Catherine) Shi Developer <jshi@math.msstate.edu>

Repository CRAN

Date/Publication 2022-10-31 14:12:52 UTC

R topics documented:

AQI *AQI Index*

Description

A quantitative measure of dataset quality. The AQI Index score indicates the degree that how features are associated with the outcome in a dataset. (synonyms of "feature": "variable" "factor" "attribute")
For more information, please refer to the corresponding publication: Shi, J., Zhang, J. and Ge, Y. (2019), "An Association-Based Intrinsic Quality Index for Healthcare Dataset Ranking" <doi:10.1109/ICHI.2019.8904553>

Usage

```
AQI(data, alpha.filter = 0.2)
```

Arguments

`data`	data frame (features as columns and observations as rows). It requires at least one feature and only one outcome. The features must be discrete. The outcome variable (Y) must be in the last column.
`alpha.filter`	level of significance for the mutual information test of independence in step 2 (<doi:10.1109/ICHI.2019.8904553>). By default, ‘alpha.filter = 0.2‘.

Value

The AQI Index score.

Examples

```
## Generate a toy dataset: "data"
n=10000
x1=rbinom(n,3,0.5)+0.2
x2=rbinom(n,2,0.8)+0.5
x3=rbinom(n,5,0.3)
error=round(runif(n,min=-1,max=1))
y=x1+x3+error
data=data.frame(cbind(x1,x2,x3,y))
colnames(data) = c("feature1", "feature2", "feature3", "Y")

## Calculate the AQI score of "data"
AQI(data)
```

`CASMI.selectFeatures` **CASMI**-*Based Feature Selection*

Description

Selects the most relevant features toward an outcome. It automatically learns the number of features to be selected, and the selected features are ranked. The method automatically handles the feature redundancy issue. (synonyms of "feature": "variable" "factor" "attribute")
For more information, please refer to the corresponding publication: Shi, J., Zhang, J. and Ge, Y. (2019), "**CASMI**—An Entropic Feature Selection Method in Turing's Perspective" <doi:10.3390/e21121179>

Usage

```
CASMI.selectFeatures(data, alpha.filter = 0.1, alpha = 0.05)
```

Arguments

`data` data frame (features as columns and observations as rows). It requires at least one feature and only one outcome. The features must be discrete. The outcome variable (Y) must be in the last column.

`alpha.filter` level of significance for the mutual information test of independence in step 1 of the features selection. The smaller the alpha.filter, the fewer the features sent to step 2 (<doi:10.3390/e21121179>). By default, 'alpha.filter = 0.1'.

`alpha` level of significance for the confidence intervals in final results. By default, 'alpha = 0.05'.

Value

'CASMI.selectFeatures()' returns selected features and relevant information, including the estimated Kappa* for all selected features ('$KappaStar') and the corresponding confidence interval ('$KappaStarCI'). The selected features are ranked. The Standardized Mutual Information using the z estimator ('SMIz') and the corresponding confidence interval ('CI.SMIz.Lower' and 'CI.SMIz.Upper') are given for each selected feature. The p-value from the mutual information test of independence using the z estimator ('P-value.MIz') is given for each selected feature.

Examples

```
## Generate a toy dataset: "data"
## Features 1 and 3 are associated with Y, while feature 2 is irrelevant.
## The outcome variable Y must be in the last column. Features must be discrete.
n=10000
x1=rbinom(n,3,0.5)+0.2
x2=rbinom(n,2,0.8)+0.5
x3=rbinom(n,5,0.3)
error=round(runif(n,min=-1,max=1))
y=x1+x3+error
data=data.frame(cbind(x1,x2,x3,y))
colnames(data) = c("feature1", "feature2", "feature3", "Y")
```

```
## Select features and relevant results from the toy dataset "data"
CASMI.selectFeatures(data)
```

Index