Package ‘NPRED’

November 27, 2021

Title Predictor Identifier: Nonparametric Prediction

Version 1.0.6

Author Ashish Sharma [aut] (<https://orcid.org/0000-0002-6758-0519>),
Raj Mehrotra [aut],
Sanjeev Jha [aut],
Jingwan Li [aut],
Ze Jiang [aut, cre] (<https://orcid.org/0000-0002-3472-0829>)

Maintainer Ze Jiang <ze.jiang@unsw.edu.au>

Description Partial informational correlation (PIC) is used to identify the meaningful predic-
tors to the response from a large set of potential predictors. Details of methodolo-
gies used in the package can be found in Sharma, A., Mehro-
tra, R., Li, J., & Jha, S. (2016). <doi:10.1016/j.envsoft.2016.05.021>, and Mehro-

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 3.4.0)

URL https://github.com/zejiang-unsw/NPRED#readme

BugReports https://github.com/zejiang-unsw/NPRED/issues

Imports stats

Suggests zoo, SPEI, WASP, knitr, ggplot2, synthesis, testthat,
bookdown, rmarkdown

RoxygenNote 7.1.1

VignetteBuilder knitr

NeedsCompilation yes

Repository CRAN

Date/Publication 2021-11-27 08:00:05 UTC
R topics documented:

- `calc.PIC`  
- `calc.PW`  
- `calc.scaleSTDratio`  
- `data.gen.ar1`  
- `data.gen.ar4`  
- `data.gen.ar9`  
- `data1`  
- `data2`  
- `data3`  
- `knn`  
- `knnreg1cv`  
- `pic.calc`  
- `pw.calc`  
- `stepwise.PIC`  
- `Sydney`  

Index

- `calc.PIC`  

Description

Calculate PIC

Usage

`calc.PIC(x, y, z, nnmax = 10000, nvarmax = 100)`

Arguments

- `x` A vector of response.
- `y` A vector of new predictor.
- `z` A matrix of pre-existing predictors that could be NULL if no prior predictors exist.
- `nnmax` The maximum number of sample size.
- `nvarmax` The maximum number of potential predictors.

Value

A list of 2 elements: the partial mutual information (pmi), and partial informational correlation (pic).

References

calc.PW  

*Calculate Partial Weight*

**Description**

Calculate Partial Weight

**Usage**

`calc.PW(x, py, cpy, cpyPIC)`

**Arguments**

- `x` A vector of response.
- `py` A matrix containing possible predictors of `x`.
- `cpy` The column numbers of the meaningful predictors (cpy).
- `cpyPIC` Partial informational correlation (cpyPIC).

**Value**

A vector of partial weights (pw) of the same length of z.

**References**


---

calc.scaleSTDratio  

*Calculate the ratio of conditional error standard deviations*

**Description**

Calculate the ratio of conditional error standard deviations

**Usage**

`calc.scaleSTDratio(x, zin, zout)`

**Arguments**

- `x` A vector of response.
- `zin` A matrix containing the meaningful predictors selected from a large set of possible predictors (z).
- `zout` A matrix containing the remaining possible predictors after taking out the meaningful predictors (zin).
Value

The STD ratio.

References


data.gen.ar1

Generate predictor and response data.

Description

Generate predictor and response data.

Usage

data.gen.ar1(nobs, ndim = 9)

Arguments

nobs  The data length to be generated.
ndim  The number of potential predictors (default is 9).

Value

A list of 2 elements: a vector of response (x), and a matrix of potential predictors (dp) with each column containing one potential predictor.

Examples

# AR1 model from paper with 9 dummy variables
data.ar1 <- data.gen.ar1(500)
stepwise.PIC(data.ar1$x, data.ar1$dp)
**data.gen.ar4**  
Generate predictor and response data.

**Description**  
Generate predictor and response data.

**Usage**  
data.gen.ar4(nobs, ndim = 9)

**Arguments**  
- **nobs**: The data length to be generated.
- **ndim**: The number of potential predictors (default is 9).

**Value**  
A list of 2 elements: a vector of response (x), and a matrix of potential predictors (dp) with each column containing one potential predictor.

**Examples**

```r
# AR4 model from paper with total 9 dimensions
data.ar4 <- data.gen.ar4(500)
stepwise.PIC(data.ar4$x, data.ar4$dp)
```

**data.gen.ar9**  
Generate predictor and response data.

**Description**  
Generate predictor and response data.

**Usage**  
data.gen.ar9(nobs, ndim = 9)

**Arguments**  
- **nobs**: The data length to be generated.
- **ndim**: The number of potential predictors (default is 9).
Value

A list of 2 elements: a vector of response (x), and a matrix of potential predictors (dp) with each column containing one potential predictor.

Examples

```r
# AR9 model from paper with total 9 dimensions
data.ar9 <- data.gen.ar9(500)
stepwise.PIC(data.ar9$x, data.ar9$dp)
```

---

data1

Sample data: AR9 model: \( x(i)=0.3*x(i-1)-0.6*x(i-4)-0.5*x(i-9)+\text{eps} \)

Description

A dataset containing 500 rows (data length) and 16 columns. The first column is response data and the rest columns are possible predictors.

Usage

data(data1)

---

data2

Sample data: AR4 model: \( x(i)=0.6*x(i-1)-0.4*x(i-4)+\text{eps} \)

Description

A dataset containing 500 rows (data length) and 16 columns. The first column is response data and the rest columns are possible predictors.

Usage

data(data2)

---

data3

Sample data: AR1 model: \( x(i)=0.9*x(i-1)+0.866*\text{eps} \)

Description

A dataset containing 500 rows (data length) and 16 columns. The first column is response data and the rest columns are possible predictors.

Usage

data(data3)
knn

Modified k-nearest neighbour conditional bootstrap or regression function estimation with extrapolation

Description

Modified k-nearest neighbour conditional bootstrap or regression function estimation with extrapolation

Usage

knn(
  x, 
  z, 
  zout, 
  k = 0, 
  pw, 
  reg = TRUE, 
  nensemble = 100, 
  tailcorrection = TRUE, 
  tailprob = 0.25, 
  tailfac = 0.2, 
  extrap = TRUE 
)

Arguments

x  A vector of response.

z  A matrix of existing predictors.

zout A matrix of predictor values the response is to be estimated at.

k  The number of nearest neighbours used. The default value is 0, indicating Lall and Sharma default is used.

pw  A vector of partial weights of the same length of z.

reg A logical operator to inform whether a conditional expectation should be output or not nensemble. Used if reg=F and represents the number of realisations that are generated Value.

nensemble An integer the specifies the number of ensembles used. The default is 100.

tailcorrection A logical value, T (default) or F, that denotes whether a reduced value of k (number of nearest neighbours) should be used in the tails of any conditioning plane. Whether one is in the tails or not is determined based on the nearest neighbour response value.

tailprob  A scalar that denotes the p-value of the cdf (on either extreme) the tailcorrection takes effect. The default value is 0.25.
tailfac  A scalar that specifies the lowest fraction of the default k that can be used in the tails. Depending on how extreme one is in the tails, the actual k decreases linearly from k (for a p-value greater than tailprob) to tailfac*k proportional to the actual p-value of the nearest neighbour response, divided by tailprob. The default value is 0.2.

extrap  A logical value, T (default) or F, that denotes whether a kernel extrapolation method is used to predict x.

Value

A matrix of responses having same rows as zout if reg=T, or having nensemble columns is reg=F.

References


Examples

data(data1)  # AR9 model  x(i)=0.3*x(i-1)-0.6*x(i-4)-0.5*x(i-9)+eps
x <- data1[, 1]  # response
py <- data1[, -1]  # possible predictors
ans.ar9 <- stepwise.PIC(x, py)  # identify the meaningful predictors and estimate partial weights
z <- py[, ans.ar9$cpy]  # predictor matrix
pw <- ans.ar9$wt  # partial weights

# vector denoting where we want outputs, can be a matrix representing grid.
zout <- apply(z, 2, mean)

knn(x, z, zout, reg = TRUE, pw = pw)  # knn regression estimate using partial weights.

knn(x, z, zout, reg = FALSE, pw = pw)  # alternatively, knn conditional bootstrap (100 realisations).
# Mean of the conditional bootstrap estimate should be
# approximately the same as the regression estimate.

zout <- ts(data.gen.ar9(500, ndim = length(ans.ar9$cpy))$dp)  # new input
xhat1 <- xhat2 <- x
xhat1 <- NPRED::knn(x, z, zout, k = 5, reg = TRUE, extrap = FALSE)  # without extrapolation
xhat2 <- NPRED::knn(x, z, zout, k = 5, reg = TRUE, extrap = TRUE)  # with extrapolation

ts.plot(ts(x), ts(xhat1), ts(xhat2),
col = c("black", "red", "blue"), ylim = c(-5, 5),
lwd = c(2, 2, 1))
)
knnregl1cv

Leave one out cross validation.

Description

Leave one out cross validation.

Usage

knnregl1cv(x, z, k = 0, pw)

Arguments

x A vector of response.
z A matrix of predictors.
k The number of nearest neighbours used. The default is 0, indicating Lall and Sharma default is used.
pw A vector of partial weights of the same length of z.

Value

A vector of L1CV estimates of the response.

References


pic.calc

Calculate PIC

Description

Calculate PIC

Usage

pic.calc(X, Y, Z = NULL)
Arguments

X A vector of response.
Y A matrix of new predictors.
Z A matrix of pre-existing predictors that could be NULL if no prior predictors exist.

Value

A list of 2 elements: the partial mutual information (pmi), and partial informational correlation (pic).

References


Calculate stepwise PIC

**Description**

Calculate stepwise PIC

**Usage**

stepwise.PIC(x, py, nvarmax = 100, alpha = 0.1)

**Arguments**

- **x**: A vector of response.
- **py**: A matrix containing possible predictors of x.
- **nvarmax**: The maximum number of variables to be selected.
- **alpha**: The significance level used to judge whether the sample estimate in Equation

\[
PIC = \sqrt{1 - \exp(-2\hat{PIC})}
\]

is significant or not. A default alpha value is 0.1.

**Value**

A list of 2 elements: the column numbers of the meaningful predictors (cpy), and partial informational correlation (cpyPIC).

**References**


**Examples**

data(data1) # AR9 model  x(i)=0.3*x(i-1)-0.6*x(i-4)-0.5*x(i-9)+eps
x <- data1[, 1] # response
py <- data1[, -1] # possible predictors
stepwise.PIC(x, py)

data(data2) # AR4 model:  x(i)=0.6*x(i-1)-0.4*x(i-4)+eps
x <- data2[, 1] # response
py <- data2[, -1] # possible predictors
stepwise.PIC(x, py)

data(data3) # AR1 model  x(i)=0.9*x(i-1)+0.866*eps
x <- data3[, 1] # response
py <- data3[, -1] # possible predictors
stepwise.PIC(x, py)
Sydney

Sample data: Data over Sydney region

Description

A dataset containing Rainfall (15 stations), NCEP and CSIRO (7 atmospheric variables).

Usage

data(Sydney)
Index

* datasets
  data1, 6
  data2, 6
  data3, 6
  Sydney, 12

calc.PIC, 2
calc.PW, 3
calc.scaleSTDratio, 3

data.gen.ar1, 4
data.gen.ar4, 5
data.gen.ar9, 5
data1, 6
data2, 6
data3, 6

knn, 7
knnregl1cv, 9

pic.calc, 9
pw.calc, 10

stepwise.PIC, 11
Sydney, 12