Package 'RPCLR'

February 19, 2015

Type Package

Title RPCLR (Random-Penalized Conditional Logistic Regression)

Version 1.0

Date 2012-08-19

Author Raji Balasubramanian

Maintainer Raji Balasubramanian <rbalasub@schoolph.umass.edu>

Depends MASS, survival

Description This package implements the R-PCLR (Random-Penalized Conditional Logistic Regression) algorithm for obtaining variable importance. The algorithm is applicable for the analysis of high dimensional data from matched case-control studies.

License GPL-2

LazyLoad yes

Repository CRAN

Date/Publication 2012-08-19 16:38:48

NeedsCompilation no

R topics documented:

RPCLR-package 2
GenerateData 2
GetVarImp 4

Index **6**

RPCLR-package *Implements the R-PCLR algorithm to calculate variable importance*

Description

The R-PCLR algorithm can be used to estimate variable importance in settings of high dimensional data arising from matched case-control studies. The algorithm accounts for the correlation between observations belonging to the same matched stratum, while incorporating some of the powerful features of Random Forests for evaluating the significance of high dimensional feature sets.

Details

Package:	RPCLR
Type:	Package
Version:	1.0
Date:	2012-08-19
License:	GPL 2.0
LazyLoad:	yes

Author(s)

Raji Balasubramanian

Maintainer: Raji Balasubramanian <rbalasub@schoolph.umass.edu>

References

Balasubramanian, R., Houseman, E. A., Coull, B. A., Lev, M. H., Schwamm, L. H., Betensky, R. A. (2012). Variable importance in matched case-control studies in settings of high dimensional data, Submitted to Biostatistics.

GenerateData *Simulate a dataset from a 1:1 matched case control study*

Description

Simulate a dataset from a 1:1 matched case control study

Usage

```
GenerateData(numstrat, NumType.BM, NumType.NS, mu.diff, rho)
```

Arguments

numstrat	number of matched pairs
NumType.BM	number of features with non-zero difference in means between cases and controls (i.e. biomarkers)
NumType.NS	number of features with identical means between cases and controls (i.e. noise)
mu.diff	Difference in means between cases and controls for biomarkers
rho	correlation between matched pairs for biomarkers only

Details

Biomarkers and noise features are simulated as independent random variables following a Gaussian distribution with unit variance.

Value

Data	a numeric data matrix of n (number of subjects) rows and p (number of features) columns
Out	a response vector of length n of binary indicators of case/control status
Strat	a vector of length n of matched pair (stratum) indicators

Author(s)

Raji Balasubramanian

References

Balasubramanian, R., Houseman, E. A., Coull, B. A., Lev, M. H., Schwamm, L. H., Betensky, R. A. (2012). Variable importance in matched case-control studies in settings of high dimensional data, Submitted to Biostatistics.

See Also

GetVarImp

Examples

```
## Simulate Data
MyDat <- GenerateData(50, 3, 7, 0.5, 0.4)
Dat <- MyDat$Data
Out <- MyDat$Out
Strat <- MyDat$Strat
```

GetVarImp *To obtain variable importance scores using the R-PCLR algorithm.*

Description

This function outputs variable importance scores based on the R-PCLR algorithm. This is applicable to settings of binary response (case versus control) and can be used to analyze high dimensional data arising from matched case control studies.

Usage

```
GetVarImp(MyData, MyOut, MyStrat, mtry, numBS)
```

Arguments

MyData	a numeric data matrix of n (number of subjects) rows and p (number of features) columns
MyOut	a response vector of length n of binary indicators of case/control status
MyStrat	a vector of length n of matched pair (stratum) indicators
mtry	Number of covariates to be sampled randomly for inclusion in each model
numBS	Number of bootstrap replicates

Details

The function implements the R-PCLR algoritm. Details are found in the paper referenced below (Balasubramanian, R. et al., 2012). The algorithm utilizes a model-based approach that incorporates a penalized conditional likelihood, which allows adjustment for the matched design. The penalized conditional logistic regression model incorporates a ridge penalty and is implemented using the ridge() function within the survival library. The penalty parameter is set to the default option in the ridge() function. See Gray, R.J (1992) for details.

Value

A p x 1 vector of variable importance scores.

Author(s)

Raji Balasubramanian

References

Balasubramanian, R., Houseman, E. A., Coull, B. A., Lev, M. H., Schwamm, L. H., Betensky, R. A. (2012). Variable importance in matched case-control studies in settings of high dimensional data, Submitted to Biostatistics.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. Journal of the American Statistical Association, 87, 942-51.

See Also

GenerateData, clogit, ridge

Examples

```
## Simulate Data of 100 matched pairs, 3 biomarkers, 5 noise features
set.seed(1234)
MyDat <- GenerateData(50, 3, 5, 0.5, 0.4)
Dat <- MyDat$Data
Out <- MyDat$Out
Strat <- MyDat$Strat

## Get Variable Importance
MyResults <- GetVarImp(Dat, Out, Strat, mtry=3, numBS=25)

## Print results
hist(MyResults, breaks=6, col="orange", xlab="Importance score", ylab="Number of features", main="Histogram of R
output <- cbind(as.character(colnames(Dat)), format(MyResults, digits=3))
print(output)

## Sort from most important (highest importance score) to least important feature (lowest importance score)
ind <- sort(MyResults, index.return=TRUE, decreasing=TRUE)$ix
output[ind,]
```

Index

GenerateData, 2
GetVarImp, 4

RPCLR *(*RPCLR-package*)*, 2
RPCLR-package, 2