Package 'UplsLda'

April 27, 2021

Version 1.0

Date 2021-04-27

Title Improved partial least square linear discriminant analysis via clustering

Author Md. Shahjaman

Maintainer Shahjaman <shahjaman_brur@yahoo.com>

Depends R (>= 4.0),ROC,stats,MASS,caret,e1071,cluster

Imports ROC,caret

Suggests MASS

Description We developed an R package to perform binary classification using PLS dimension reduction and linear discriminant analysis applied on the PLS components via clustering.

License GPL (>= 2)

NeedsCompilation no

R topics documented:

UplsLda-package *Improved partial least square linear discriminant analysis via clustering*

Description

This package performs binary classification using the method described in Boulesteix (2004) which consists in PLS dimension reduction and linear discriminant analysis applied on the PLS components via clustering

Details

Package:	UplsLda
Type:	Package
Version:	1.0
Date:	2021-04-27
License:	GPL (>=2.0)

Author(s)

Md.Shahjaman and Md. Rabiul Auwul Maintainer: Md.Shahjaman shahjaman_brur@yahoo.com

Colon *Gene expression data from Alon et al. (1999)*

Description

Gene expression data (2000 genes for 62 samples) from the microarray experiments of Colon tissue samples of Alon et al. (1999).

Usage

```
data(Colon)
```

Details

This data set contains 62 samples with 2000 genes: 40 tumor tissues, coded 2 and 22 normal tissues, coded 1.

Value

A list with the following elements:

X	a (62 x 2000) matrix giving the expression levels of 2000 genes for the 62 Colon tissue samples. Each row corresponds to a patient, each column to a gene.
Y	a numeric vector of length 62 giving the type of tissue sample (tumor or normal).
gene.names	a vector containing the names of the 2000 genes for the gene expression matrix X.

Source

The data are described in Alon et al. (1999) and can be freely downloaded from http://microarray.princeton.edu/oncology/affydata/index.html.

References

Alon, U. and Barkai, N. and Notterman, D.A. and Gish, K. and Ybarra, S. and Mack, D. and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA,**96**(12), 6745–6750.

Examples

```
# load data set
data(Colon)

# how many samples and how many genes ?
dim(Colon$X)

# how many samples of class 1 and 2 respectively ?
sum(Colon$Y==1)
sum(Colon$Y==2)
```

Iplsda *Improved partial least square linear discriminant analysis*

Description

The function Iplsda performs binary classification using the method described in Boulesteix (2004) which consists in PLS dimension reduction and linear discriminant analysis applied on the PLS components via clustering.

Usage

```
Iplsda(Xtr, Xte, trlevel, televel, ncls = 2)
```

Arguments

Xtr	Training dataset
Xte	Test dataset
trlevel	Training data level
televel	Test data level
ncls	Number of possible cluster in the training dataset

Value

The following values are produced by Iplsda():

cla_acc	Accuracy of classical pls.lda function
prop_acc	Accuracy of proposed pls.lda function via clustering
pred_level	Predicted test level by the proposed pls.lda

Author(s)

Md.Shahjaman shahjaman_brur@yahoo.com

---

pls.lda *Classification with PLS Dimension Reduction and Linear Discriminant Analysis*

Description

The function pls.lda performs binary or multicategorical classification using the method described in Boulesteix (2004) which consists in PLS dimension reduction and linear discriminant analysis applied on the PLS components.

Usage

pls.lda(Xtrain, Ytrain, Xtest=NULL, ncomp, nruncv=0, alpha=2/3, priors=NULL)

Arguments

Xtrain	a (ntrain x p) data matrix containing the predictors for the training data set. Xtrain may be a matrix or a data frame. Each row is an observation and each column is a predictor variable.
Ytrain	a vector of length ntrain giving the classes of the ntrain observations. The classes must be coded as 1,...,K (K>=2).
Xtest	a (ntest x p) data matrix containing the predictors for the test data set. Xtest may also be a vector of length p (corresponding to only one test observation). If Xtest=NULL, the training data set is considered as test data set as well.
ncomp	if nruncv=0, ncomp is the number of latent components to be used for PLS dimension reduction. If nruncv>0, the cross-validation procedure described in Boulesteix (2004) is used to choose the best number of components from the vector of integers ncomp or from 1,...,ncomp if ncomp is of length 1.
nruncv	the number of cross-validation iterations to be performed for the choice of the number of latent components. If nruncv=0, cross-validation is not performed and ncomp latent components are used.
alpha	the proportion of observations to be included in the training set at each cross-validation iteration.
priors	The class priors to be used for linear discriminant analysis. If unspecified, the class proportions in the training set are used.

Details

The function pls.lda proceeds as follows to predict the class of the observations from the test data set. First, the SIMPLS algorithm is run on Xtrain and Ytrain to determine the new PLS components based on the training observations only. The new PLS components are then computed for the test data set. Classification is performed by applying classical linear discriminant analysis (LDA) to the new components. Of course, the LDA classifier is built using the training observations only.

Value

A list with the following components:

predclass	the vector containing the predicted classes of the ntest observations from Xtest.
ncomp	the number of latent components used for classification.
pls.out	an object containing the results from the call of the pls.regression function (from the plsgenomics package).
lda.out	an object containing the results from the call of the lda function (from the MASS package).
pred.lda.out	an object containing the results from the call of the predict.lda function (from the MASS package).

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/eng.html)

References

A. L. Boulesteix (2004). PLS dimension reduction for classification with microarray data, Statistical Applications in Genetics and Molecular Biology **3**, Issue 1, Article 33.

A. L. Boulesteix, K. Strimmer (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 7:32-44.

S. de Jong (1993). SIMPLS: an alternative approach to partial least squares regression, Chemometrics Intell. Lab. Syst. **18**, 251–263.

---

pls.lda.cv — *Determination of the number of latent components to be used for classification with PLS and LDA*

Description

The function pls.lda.cv determines the best number of latent components to be used for classification with PLS dimension reduction and linear discriminant analysis as described in Boulesteix (2004).

Usage

pls.lda.cv(Xtrain, Ytrain, ncomp, nruncv=20, alpha=2/3, priors=NULL)

Arguments

Xtrain	a (ntrain x p) data matrix containing the predictors for the training data set. Xtrain may be a matrix or a data frame. Each row is an observation and each column is a predictor variable.
Ytrain	a vector of length ntrain giving the classes of the ntrain observations. The classes must be coded as 1,...,K (K>=2).

ncomp	the vector of integers from which the best number of latent components has to be chosen by cross-validation. If ncomp is of length 1, the best number of components is chosen from 1,...,ncomp.
nruncv	the number of cross-validation iterations to be performed for the choice of the number of latent components.
alpha	the proportion of observations to be included in the training set at each cross-validation iteration.
priors	The class priors to be used for linear discriminant analysis. If unspecified, the class proportions in the training set are used.

Details

The cross-validation procedure described in Boulesteix (2004) is used to determine the best number of latent components to be used for classification. At each cross-validation run, Xtrain is split into a pseudo training set and a pseudo test set and the classification error rate is determined for each number of latent components. Finally, the function pls.lda.cv returns the number of latent components for which the mean classification rate over the nrun partitions is minimal.

Value

The number of latent components to be used for classification.

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/eng.html)

References

A. L. Boulesteix (2004). PLS dimension reduction for classification with microarray data, Statistical Applications in Genetics and Molecular Biology **3**, Issue 1, Article 33.

A. L. Boulesteix, K. Strimmer (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 7:32-44.

S. de Jong (1993). SIMPLS: an alternative approach to partial least squares regression, Chemometrics Intell. Lab. Syst. **18**, 251–263.

---

pls.lda.sample *partial least square sample*

Description

partial least square sample

Usage

```
pls.lda.sample(samp, X, Y, ncomp, priors = NULL)
```

Arguments

samp	Sample
X	Training dataset
Y	Training data level
ncomp	Number of component
priors	Prior probabilities

Value

errorcv	Description of 'comp1'

Author(s)

Md. Shahjaman

---

pls.regression *Multivariate Partial Least Squares Regression*

Description

The function pls.regression performs pls multivariate regression (with several response variables and several predictor variables) using de Jong's SIMPLS algorithm. This function is an adaptation of R. Wehrens' code from the package pls.pcr.

Usage

pls.regression(Xtrain, Ytrain, Xtest=NULL, ncomp=NULL, unit.weights=TRUE)

Arguments

Xtrain	a (ntrain x p) data matrix of predictors. Xtrain may be a matrix or a data frame. Each row corresponds to an observation and each column to a predictor variable.
Ytrain	a (ntrain x m) data matrix of responses. Ytrain may be a vector (if m=1), a matrix or a data frame. If Ytrain is a matrix or a data frame, each row corresponds to an observation and each column to a response variable. If Ytrain is a vector, it contains the unique response variable for each observation.
Xtest	a (ntest x p) matrix containing the predictors for the test data set. Xtest may also be a vector of length p (corresponding to only one test observation).
ncomp	the number of latent components to be used for regression. If ncomp is a vector of integers, the regression model is built successively with each number of components. If ncomp=NULL, the maximal number of components min(ntrain,p) is chosen.
unit.weights	if TRUE then the latent components will be constructed from weight vectors that are standardized to length 1, otherwise the weight vectors do not have length 1 but the latent components have norm 1.

Details

The columns of the data matrices Xtrain and Ytrain must not be centered to have mean zero, since centering is performed by the function pls.regression as a preliminary step before the SIMPLS algorithm is run.

In the original definition of SIMPLS by de Jong (1993), the weight vectors have length 1. If the weight vectors are standardized to have length 1, they satisfy a simple optimality criterion (de Jong, 1993). However, it is also usual (and computationally efficient) to standardize the latent components to have length 1.

In contrast to the original version found in the package pls.pcr, the prediction for the observations from Xtest is performed after centering the columns of Xtest by substracting the columns means calculated from Xtrain.

Value

A list with the following components:

B	the (p x m x length(ncomp)) matrix containing the regression coefficients. Each row corresponds to a predictor variable and each column to a response variable. The third dimension of the matrix B corresponds to the number of PLS components used to compute the regression coefficients. If ncomp has length 1, B is just a (p x m) matrix.
Ypred	the (ntest x m x length(ncomp)) containing the predicted values of the response variables for the observations from Xtest. The third dimension of the matrix Ypred corresponds to the number of PLS components used to compute the regression coefficients.
P	the (p x max(ncomp)) matrix containing the X-loadings.
Q	the (m x max(ncomp)) matrix containing the Y-loadings.
T	the (ntrain x max(ncomp)) matrix containing the X-scores (latent components)
R	the (p x max(ncomp)) matrix containing the weights used to construct the latent components.
meanX	the p-vector containing the means of the columns of Xtrain.

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/eng.html) and Korbinian Strimmer (http://strimmerlab.org/).

Adapted in part from pls.pcr code by R. Wehrens (in a former version of the 'pls' package https://CRAN.R-project.org/package=pls).

References

S. de Jong (1993). SIMPLS: an alternative approach to partial least squares regression, Chemometrics Intell. Lab. Syst. **18**, 251–263.

C. J. F. ter Braak and S. de Jong (1993). The objective function of partial least squares regression, Journal of Chemometrics **12**, 41–54.

pls.regression.cv *Determination of the number of latent components to be used in PLS regression*

Description

The function pls.regression.cv determines the best number of latent components to be used for PLS regression using the cross-validation approach described in Boulesteix and Strimmer (2005).

Usage

pls.regression.cv(Xtrain, Ytrain, ncomp, nruncv=20, alpha=2/3)

Arguments

Xtrain	a (ntrain x p) data matrix containing the predictors for the training data set. Xtrain may be a matrix or a data frame. Each row is an observation and each column is a predictor variable.
Ytrain	a (ntrain x m) data matrix of responses. Ytrain may be a vector (if m=1), a matrix or a data frame. If Ytrain is a matrix or a data frame, each row is an observation and each column is a response variable. If Ytrain is a vector, it contains the unique response variable for each observation.
ncomp	the vector of integers from which the best number of latent components has to be chosen by cross-validation. If ncomp is of length 1, the best number of components is chosen from 1,...,ncomp.
nruncv	the number of cross-validation iterations to be performed for the choice of the number of latent components.
alpha	the proportion of observations to be included in the training set at each cross-validation iteration.

Details

The cross-validation procedure described in Boulesteix and Strimmer (2005) is used to determine the best number of latent components to be used for classification. At each cross-validation run, Xtrain is split into a pseudo training set and a pseudo test set and the squared error is determined for each number of latent components. Finally, the function pls.regression.cv returns the number of latent components for which the mean squared error over the nrun partitions is minimal.

Value

The number of latent components to be used in PLS regression, as determined by cross-validation.

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/eng.html) and Korbinian Strimmer (http://strimmerlab.org/).

References

A. L. Boulesteix and K. Strimmer (2005). Predicting Transcription Factor Activities from Combined Analysis of Microarray and ChIP Data: A Partial Least Squares Approach.

A. L. Boulesteix, K. Strimmer (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 7:32-44.

S. de Jong (1993). SIMPLS: an alternative approach to partial least squares regression, Chemometrics Intell. Lab. Syst. **18**, 251–263.

---

standard.simpls *standard simpls*

Description

standard simpls

Usage

standard.simpls(Xtrain, Ytrain, Xtest = NULL, ncomp = NULL)

Arguments

Xtrain	Training data
Ytrain	Training data level
Xtest	Test data
ncomp	Number of component

Author(s)

Md. Shahjaman

---

transformy *transformy*

Description

transformy

Usage

transformy(y)

Arguments

y	data level

Author(s)

Md.Shahjaman

unitr.simpls *unitr simpls*

Description

unitr simpls

Usage

```
unitr.simpls(Xtrain, Ytrain, Xtest = NULL, ncomp = NULL)
```

Arguments

Xtrain	Training dataset
Ytrain	Training data level
Xtest	Test dataset
ncomp	Number of component

Author(s)

Md. Shahjaman

Index