
`generate.learningsets`

Generating learning sets

Description

This function generates a design matrix giving the indices of observations forming the learning data set for several iterations.

Usage

```
generate.learningsets(n,method,fold=NULL,niter=NULL,nlearn=NULL)
```

Arguments

<code>n</code>	The total number of observations in the available data set.
<code>method</code>	One of "LOOCV" (leave-one-out cross-validation), "CV" (cross-validation), "MCCV" (Monte-Carlo cross-validation, also called subsampling), "bootstrap" (bootstrap sampling - with replacement).
<code>fold</code>	Gives the number of CV-groups. Used only when <code>method="CV"</code> .
<code>niter</code>	Number of iterations.
<code>nlearn</code>	Number of observations in the learning sets. Used only for <code>method="MCCV"</code> and <code>method="bootstrap"</code> . When <code>method="bootstrap"</code> , the default is <code>nlearn=n</code> .

Details

When `method="CV"`, `niter` gives the number of times the whole CV-procedure is repeated. The output matrix has then `foldxniter` rows. When `method="MCCV"` or `method="bootstrap"`, `niter` is simply the number of considered learning sets.

Note that `method="CV", fold=n` is equivalent to `method="LOOCV"`.

Value

A matrix giving the indices (from 1 to `n`) of the observations included in the learning sets. Each row corresponds to a learning set. The order of the columns is not important. The number of rows is equal to `n` when `method="LOOCV"`, `niter` when `method="MCCV"` or `method="bootstrap"`, `fold` when `method="CV"` and `niter` is null, and `fold x niter` when `method="CV"` and `niter` is non-null.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmr.net/boulesteix>)

References

Boulesteix AL, Stobl C, Augustin T, Daumer M, 2007. Evaluating microarray-based classifiers: an overview. Technical Report 5 (<http://epub.ub.uni-muenchen.de/2065/1/tr005.pdf>).

See Also

`testclass`.

Examples

```
# load MAclinical library
# library(MAclinical)

# LOOCV
generate.learningsets(n=40,method="LOOCV")

# CV
generate.learningsets(n=40,method="CV",fold=5)
generate.learningsets(n=40,method="CV",fold=5,niter=3)

# MCCV
generate.learningsets(n=40,method="MCCV",niter=3,nlearn=30)

# bootstrap
generate.learningsets(n=40,method="bootstrap",niter=3)
```

<code>logistic_z</code>	<i>Class prediction based on logistic regression using clinical parameters only</i>
-------------------------	---

Description

This function builds a prediction rule based on the learning data (clinical predictors only) and applies it to the test data. It uses the function `glm`.

Usage

```
logistic_z(Xlearn=NULL,Zlearn,Ylearn,Xtest=NULL,Ztest)
```

Arguments

<code>Xlearn</code>	A <code>nlearn</code> x <code>p</code> matrix giving the microarray predictors for the learning data set. This argument is ignored.
<code>Zlearn</code>	A <code>nlearn</code> x <code>q</code> matrix giving the clinical predictors for the learning data set.
<code>Ylearn</code>	A numeric vector of length <code>nlearn</code> giving the class membership of the learning observations, coded as 0,1.

Xtest A $n_{\text{test}} \times p$ matrix giving the microarray predictors for the test data set. This argument is ignored.

Ztest A $n_{\text{test}} \times q$ matrix giving the clinical predictors for the test data set.

Details

See Boulesteix et al (2008).

Value

A list with the element:

prediction A numeric vector of length `nrow(Xtest)` giving the predicted class for each observation from the test data set.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

See Also

[testclass](#), [testclass_simul](#), [simulate](#), [plsr_x_pv](#), [plsr_xz_pv](#), [plsr_x](#), [plsr_xz](#), [rf_z](#), [svm_x](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating zlearn, ylearn, ztest
zlearn<-matrix(rnorm(120),30,4)
ylearn<-sample(0:1,30,replace=TRUE)
ztest<-matrix(rnorm(80),20,4)

my.prediction<-logistic_z(Zlearn=zlearn,Ylearn=ylearn,Ztest=ztest)
my.prediction
```

`plsr_x`

Classification based on PLS dimension reduction and random forests using microarray data only

Description

This function builds a prediction rule based on the learning data (microarray predictors only) and applies it to the test data. The classifier consists of two steps: PLS dimension reduction (without pre-validation step) for summarizing microarray data, and random forests applied to the obtained PLS components. See Boulesteix et al (2008) for more details.

The function `plsr_x` uses the functions `cforest` and `varimp` from the package `party` and the function `pls.regression` from the package `plsgenomics`.

Usage

```
plsr_x(Xlearn, Zlearn=NULL, Ylearn, Xtest, Ztest=NULL, ncomp=0:3, ordered=NULL, nbgene=NULL, ...)
```

Arguments

<code>Xlearn</code>	A <code>nlearn</code> x <code>p</code> matrix giving the microarray predictors for the learning data set.
<code>Zlearn</code>	A <code>nlearn</code> x <code>q</code> matrix giving the clinical predictors for the learning data set. This argument is ignored.
<code>Ylearn</code>	A numeric vector of length <code>nlearn</code> giving the class membership of the learning observations, coded as 0,...,K-1 (where K is the number of classes).
<code>Xtest</code>	A <code>ntest</code> x <code>p</code> matrix giving the microarray predictors for the test data set.
<code>Ztest</code>	A <code>ntest</code> x <code>q</code> matrix giving the clinical predictors for the test data set. This argument is ignored.
<code>ncomp</code>	A numeric vector giving the candidate numbers of PLS components. All numbers must be >0 .
<code>ordered</code>	A vector of length <code>p</code> giving the order of the microarray predictors in terms of relevance for prediction. For instance, if the three first elements of <code>ordered</code> are 30,2,2400, it means that the most relevant genes are the genes in the 30th, 2nd and 2400th columns of the gene expression data matrix <code>Xlearn</code> . Note: if <code>ordered=NULL</code> , the columns of <code>Xlearn</code> and <code>Xtest</code> are assumed to be already ordered.
<code>nbgene</code>	The number of genes to be selected for use in dimension reduction. Default is <code>nbgene=NULL</code> , in which case all genes are used.
<code>...</code>	Other arguments to be passed to the function <code>cforest_control</code> from the <code>party</code> package.

Details

See Boulesteix et al (2008).

Value

A list with the elements:

<code>prediction</code>	A numeric vector of length <code>nrow(Xtest)</code> giving the predicted class for each observation from the test data set.
<code>importance</code>	The variable importance information output by the function <code>varimp</code> from the package <code>party</code> for the corresponding forest.
<code>bestncomp</code>	The best number of PLS components, as obtained using the model selection method based on the out-of-bag error.
<code>OOB</code>	A numeric vector of length <code>ncomp</code> giving the out-of-bag error of the forest constructed with the corresponding number of PLS components.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

See Also

[testclass](#), [testclass_simul](#), [simulate](#), [plsrf_x_pv](#), [plsrf_xz](#), [plsrf_xz_pv](#), [rf_z](#), [logistic_z](#), [svm_x](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating xlearn, zlearn, ylearn, xtest, ztest
xlearn<-matrix(rnorm(3000),30,100)
zlearn<-matrix(rnorm(120),30,4)
ylearn<-sample(0:1,30,replace=TRUE)
xtest<-matrix(rnorm(2000),20,100)
ztest<-matrix(rnorm(80),20,4)

my.prediction1<-plsrf_x(Xlearn=xlearn,Ylearn=ylearn,Xtest=xtest)

ordered<-sample(100)
my.prediction2<-plsrf_x(Xlearn=xlearn,Ylearn=ylearn,Xtest=xtest,ordered=ordered,nbgene=20)
```

`plsrfr_x_pv`

Classification based on pre-validated PLS dimension reduction and random forests using microarray data only

Description

This function builds a prediction rule based on the learning data (microarray predictors only) and applies it to the test data. The classifier consists of two steps: PLS dimension reduction with pre-validation step for summarizing microarray data, and random forests applied to the obtained PLS components. See Boulesteix et al (2008) for more details.

The function `plsrfr_x_pv` uses the functions `cforest` and `varimp` from the package `party` and the function `pls.regression` from the package `plsgenomics`.

Usage

```
plsrfr_x_pv(Xlearn,Zlearn=NULL,Ylearn,Xtest,Ztest=NULL,ncomp=0:3,  
ordered=NULL,nbgene=NULL,fold=10,...)
```

Arguments

<code>Xlearn</code>	A <code>nlearn</code> x <code>p</code> matrix giving the microarray predictors for the learning data set.
<code>Zlearn</code>	A <code>nlearn</code> x <code>q</code> matrix giving the clinical predictors for the learning data set. This argument is ignored.
<code>Ylearn</code>	A numeric vector of length <code>nlearn</code> giving the class membership of the learning observations, coded as <code>0,...,K-1</code> (where <code>K</code> is the number of classes).
<code>Xtest</code>	A <code>ntest</code> x <code>p</code> matrix giving the microarray predictors for the test data set.
<code>Ztest</code>	A <code>ntest</code> x <code>q</code> matrix giving the clinical predictors for the test data set. This argument is ignored.
<code>ncomp</code>	A numeric vector giving the candidate numbers of pre-validated PLS components. All numbers must be <code>>0</code> .
<code>ordered</code>	A vector of length <code>p</code> giving the order of the microarray predictors in terms of relevance for prediction. For instance, if the three first elements of <code>ordered</code> are <code>30,2,2400</code> , it means that the most relevant genes are the genes in the 30th, 2nd and 2400th columns of the gene expression data matrix <code>Xlearn</code> . Note: if <code>ordered=NULL</code> , the columns of <code>Xlearn</code> and <code>Xtest</code> are assumed to be already ordered.
<code>nbgene</code>	The number of genes to be selected for use in dimension reduction. Default is <code>nbgene=NULL</code> , in which case all genes are used.
<code>fold</code>	The number of folds for the pre-validation step. See Boulesteix et al (2008) for more details. The default is <code>fold=10</code> .
<code>...</code>	Other arguments to be passed to the function <code>cforest_control</code> from the <code>party</code> package.

Details

See Boulesteix et al (2008).

Value

A list with the elements:

<code>prediction</code>	A numeric vector of length <code>nrow(Xtest)</code> giving the predicted class for each observation from the test data set.
<code>importance</code>	The variable importance information output by the function <code>varimp</code> from the package <code>party</code> for the corresponding forest.
<code>bestncomp</code>	The best number of pre-validated PLS components, as obtained using the model selection method based on the out-of-bag error.
<code>OoB</code>	A numeric vector of length <code>ncomp</code> giving the out-of-bag error of the forest constructed with the corresponding number of pre-validated PLS components.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmr.net/boulesteix>)

References

- Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.
- Tibshirani R, Efron B, 2002. Pre-validation and inference in microarrays. *Stat. Appl. Genet. Mol. Biol.* 1:1.

See Also

[testclass](#), [testclass_simul](#), [simulate](#), [plsrf_x](#), [plsrf_xz](#), [plsrf_xz_pv](#), [rf_z](#), [logistic_z](#), [svm_x](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating xlearn, zlearn, ylearn, xtest, ztest
xlearn<-matrix(rnorm(3000),30,100)
ylearn<-sample(0:1,30,replace=TRUE)
xtest<-matrix(rnorm(2000),20,100)

my.prediction1<-plsrf_x_pv(Xlearn=xlearn,Ylearn=ylearn,Xtest=xtest)

ordered<-sample(100)
my.prediction2<-plsrf_x(Xlearn=xlearn,Ylearn=ylearn,Xtest=xtest,ordered=ordered,nbgene=20)
my.prediction3<-plsrf_x_pv(Xlearn=xlearn,Ylearn=ylearn,Xtest=xtest,fold=30)
```

`plsrf_xz`

Classification based on PLS dimension reduction and random forests using both clinical and microarray predictors

Description

This function builds a prediction rule based on the learning data (both clinical and microarray predictors) and applies it to the test data. The classifier consists of two steps: PLS dimension reduction (without pre-validation step) for summarizing microarray data, and random forests applied to both PLS components and clinical predictors. See Boulesteix et al (2008) for more details.

The function `plsrf_xz` uses the functions `cforest` and `varimp` from the package `party` and the function `pls.regression` from the package `plsgenomics`.

Usage

```
plsrf_xz(Xlearn,Zlearn,Ylearn,Xtest,Ztest,ncomp=0:3,ordered=NULL,nbgene=NULL,...)
```

Arguments

<code>Xlearn</code>	A <code>nlearn</code> x <code>p</code> matrix giving the microarray predictors for the learning data set.
<code>Zlearn</code>	A <code>nlearn</code> x <code>q</code> matrix giving the clinical predictors for the learning data set.
<code>Ylearn</code>	A numeric vector of length <code>nlearn</code> giving the class membership of the learning observations, coded as <code>0,...,K-1</code> (where <code>K</code> is the number of classes).
<code>Xtest</code>	A <code>ntest</code> x <code>p</code> matrix giving the microarray predictors for the test data set.
<code>Ztest</code>	A <code>ntest</code> x <code>q</code> matrix giving the clinical predictors for the test data set.
<code>ncomp</code>	A numeric vector giving the candidate numbers of PLS components. All numbers must be ≥ 0 . The number 0 corresponds to prediction based on clinical parameters only.
<code>ordered</code>	A vector of length <code>p</code> giving the order of the microarray predictors in terms of relevance for prediction. For instance, if the three first elements of <code>ordered</code> are <code>30,2,2400</code> , it means that the most relevant genes are the genes in the 30th, 2nd and 2400th columns of the gene expression data matrix <code>Xlearn</code> . Note: if <code>ordered=NULL</code> , the columns of <code>Xlearn</code> and <code>Xtest</code> are assumed to be already ordered.
<code>nbgene</code>	The number of genes to be selected for use in dimension reduction. Default is <code>nbgene=NULL</code> , in which case all genes are used.
<code>...</code>	Other arguments to be passed to the function <code>cforest_control</code> from the <code>party</code> package.

Details

See Boulesteix et al (2008).

Value

A list with the elements:

<code>prediction</code>	A numeric vector of length <code>nrow(Xtest)</code> giving the predicted class for each observation from the test data set.
<code>importance</code>	The variable importance information output by the function <code>varimp</code> from the package <code>party</code> for the corresponding forest.
<code>bestncomp</code>	The best number of PLS components, as obtained using the model selection method based on the out-of-bag error.
<code>OOB</code>	A numeric vector of length <code>ncomp</code> giving the out-of-bag error of the forest constructed with the corresponding number of PLS components.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmsr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. A Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

See Also

[testclass](#), [testclass_simul](#), [simulate](#), [plsrf_x_pv](#), [plsrf_xz_pv](#), [plsrf_x](#), [rf_z](#), [logistic_z](#), [svm_x](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating xlearn, zlearn, ylearn, xtest, ztest
xlearn<-matrix(rnorm(3000),30,100)
zlearn<-matrix(rnorm(120),30,4)
ylearn<-sample(0:1,30,replace=TRUE)
xtest<-matrix(rnorm(2000),20,100)
ztest<-matrix(rnorm(80),20,4)

my.prediction1<-plsrf_xz(Xlearn=xlearn,Zlearn=zlearn,Ylearn=ylearn,Xtest=xtest,Ztest=ztest)

ordered<-sample(100)
my.prediction2<-plsrf_xz(Xlearn=xlearn,Zlearn=zlearn,Ylearn=ylearn,
Xtest=xtest,Ztest=ztest,ordered=ordered,nbgene=20)
```

`plsrf_xz_pv`

Classification based on pre-validated PLS dimension reduction and random forests using both clinical and microarray predictors

Description

This function builds a prediction rule based on the learning data (both clinical and microarray predictors) and applies it to the test data. The classifier consists of two steps: PLS dimension reduction involving a pre-validation step for summarizing microarray data, and random forests applied to both PLS components and clinical predictors. See Boulesteix et al (2008) for more details.

The function `plsrf_xz_pv` uses the functions `cforest` and `varimp` from the package `party` and the function `pls.regression` from the package `plsgenomics`.

Usage

```
plsrf_xz_pv(Xlearn,Zlearn,Ylearn,Xtest,Ztest,ncomp=0:3,  
ordered=NULL,nbgene=NULL,fold=10,...)
```

Arguments

<code>Xlearn</code>	A <code>nlearn</code> x <code>p</code> matrix giving the microarray predictors for the learning data set.
<code>Zlearn</code>	A <code>nlearn</code> x <code>q</code> matrix giving the clinical predictors for the learning data set.
<code>Ylearn</code>	A numeric vector of length <code>nlearn</code> giving the class membership of the learning observations, coded as <code>0,...,K-1</code> (where <code>K</code> is the number of classes).
<code>Xtest</code>	A <code>ntest</code> x <code>p</code> matrix giving the microarray predictors for the test data set.
<code>Ztest</code>	A <code>ntest</code> x <code>q</code> matrix giving the clinical predictors for the test data set.
<code>ncomp</code>	A numeric vector giving the candidate numbers of pre-validated PLS components. All numbers must be ≥ 0 . The number 0 corresponds to prediction based on clinical parameters only.
<code>ordered</code>	A vector of length <code>p</code> giving the order of the microarray predictors in terms of relevance for prediction. For instance, if the three first elements of <code>ordered</code> are <code>30,2,2400</code> , it means that the most relevant genes are the genes in the 30th, 2nd and 2400th columns of the gene expression data matrix <code>Xlearn</code> . Note: if <code>ordered=NULL</code> , the columns of <code>Xlearn</code> and <code>Xtest</code> are assumed to be already ordered.
<code>nbgene</code>	The number of genes to be selected for use in dimension reduction. Default is <code>nbgene=NULL</code> , in which case all genes are used.
<code>fold</code>	The number of folds for the pre-validation step. See Boulesteix et al (2008) for more details. The default is <code>fold=10</code> .
<code>...</code>	Other arguments to be passed to the function <code>cforest_control</code> from the <code>party</code> package.

Details

See Boulesteix et al (2008).

Value

A list with the elements:

<code>prediction</code>	A numeric vector of length <code>nrow(Xtest)</code> giving the predicted class for each observation from the test data set.
<code>importance</code>	The variable importance information output by the function <code>varimp</code> from the package <code>party</code> for the corresponding forest.
<code>bestncomp</code>	The best number of pre-validated PLS components, as obtained using the model selection method based on the out-of-bag error.
<code>OOB</code>	A numeric vector of length <code>ncomp</code> giving the out-of-bag error of the forest constructed with the corresponding number of pre-validated PLS components.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmr.net/boulesteix>)

References

- Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.
- Tibshirani R, Efron B, 2002. Pre-validation and inference in microarrays. *Stat. Appl. Genet. Mol. Biol.* 1:1.

See Also

[testclass](#), [testclass_simul](#), [simulate](#), [plsrf_x](#), [plsrf_x_pv](#), [plsrf_xz](#), [rf_z](#), [logistic_z](#), [svm_x](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating xlearn, zlearn, ylearn, xtest, ztest
xlearn<-matrix(rnorm(3000),30,100)
zlearn<-matrix(rnorm(120),30,4)
ylearn<-sample(0:1,30,replace=TRUE)
xtest<-matrix(rnorm(2000),20,100)
ztest<-matrix(rnorm(80),20,4)

my.prediction1<-plsrf_xz_pv(Xlearn=xlearn,Zlearn=zlearn,Ylearn=ylearn,
Xtest=xtest,Ztest=ztest)

ordered<-sample(100)
my.prediction2<-plsrf_xz_pv(Xlearn=xlearn,Zlearn=zlearn,Ylearn=ylearn,
```

```
Xtest=xtest,Ztest=ztest,ordered=ordered,nbgene=20)
my.prediction3<-plsrf_xz_pv(Xlearn=xlearn,Zlearn=zlearn,Ylearn=ylearn,
Xtest=xtest,Ztest=ztest,fold=30)
```

<code>rf_z</code>	<i>Class prediction based on random forests using clinical parameters only</i>
-------------------	--

Description

This function builds a prediction rule based on the learning data (clinical predictors only) and applies it to the test data. It uses the function `cforest` from the package `party`. See Boulesteix et al (2008) for more details.

Usage

```
rf_z(Xlearn=NULL,Zlearn,Ylearn,Xtest=NULL,Ztest,...)
```

Arguments

<code>Xlearn</code>	A <code>nlearn</code> x <code>p</code> matrix giving the microarray predictors for the learning data set. This argument is ignored.
<code>Zlearn</code>	A <code>nlearn</code> x <code>q</code> matrix giving the clinical predictors for the learning data set.
<code>Ylearn</code>	A numeric vector of length <code>nlearn</code> giving the class membership of the learning observations, coded as <code>0,...,K-1</code> (where <code>K</code> is the number of classes).
<code>Xtest</code>	A <code>ntest</code> x <code>p</code> matrix giving the microarray predictors for the test data set. This argument is ignored.
<code>Ztest</code>	A <code>ntest</code> x <code>q</code> matrix giving the clinical predictors for the test data set.
<code>...</code>	Other arguments to be passed to the function <code>cforest_control</code> from the <code>party</code> package.

Details

See Boulesteix et al (2008).

Value

A list with the elements:

<code>prediction</code>	A numeric vector of length <code>nrow(Xtest)</code> giving the predicted class for each observation from the test data set.
<code>importance</code>	The variable importance information output by the function <code>varimp</code> from the package <code>party</code> .
<code>OoB</code>	The out-of-bag error of the constructed forest.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmsr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

See Also

[testclass](#), [testclass_simul](#), [simulate](#), [plsr_x_pv](#), [plsr_xz_pv](#), [plsr_x](#), [plsr_xz](#), [logistic_z](#), [svm_x](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating zlearn, ylearn, ztest
zlearn<-matrix(rnorm(120),30,4)
ylearn<-sample(0:1,30,replace=TRUE)
ztest<-matrix(rnorm(80),20,4)

my.prediction<-rf_z(Zlearn=zlearn,Ylearn=ylearn,Ztest=ztest)
my.prediction
```

simulate

Simulating data

Description

This function simulates a list of data sets as described in Boulesteix et al (2008), section 3.1.

Usage

```
simuldata_list(niter=50,n=500,p=1000,psig=50,q=5,muX=0,muZ=0)
simuldatacluster_list(niter=50,n=500,p=1000,psig=50,q=5,muX=0,muZ=0)
```

Arguments

niter	The number of data sets to be simulated.
n	The number of observations.
p	The number of microarray variables (genes).
psig	The number of significant microarray variables (must be <p).
q	The number of clinical variables.

muX The class mean difference for the **psig** relevant genes.
muZ The class mean difference for the **q** clinical variables.

Details

With the function `simuldata_cluster`, observations with `y=1` are assumed to come from two different subgroups, 1a and 1b, each with probability 0.5. Relevant genes are generated such that they separate class 1a from the rest, whereas clinical variables separate class 1b from the rest.

Value

A `niter`-list of simulated data sets. Each data set is given as a list with three elements:

`y` the `n`-vector of class memberships, coded as 0,1.
`x` the `n x p` matrix of gene expressions levels. Each row corresponds to an observation, each column to a variable (gene).
`z` the `n x q` matrix of clinical variables. Each row corresponds to an observation, each column to a clinical variable.

Author(s)

Anne-Laure Boulesteix (<http://www.slcmsr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

See Also

[testclass](#), [testclass_simul](#), [plsrf_x_pv](#), [plsrf_xz_pv](#), [plsrf_x](#), [plsrf_xz](#), [logistic_z](#), [rf_z](#), [svm_x](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating 3 simulated data sets
my.data<-simuldata_list(niter=3,n=100,p=150,psig=10,q=5,muX=2,muZ=1)
length(my.data)
dim(my.data[[1]]$x)
dim(my.data[[1]]$z)
length(my.data[[1]]$y)
```

<code>svm_x</code>	<i>Classification based on support vector machines using microarray data only</i>
--------------------	---

Description

This function builds a prediction rule based on the learning data (microarray predictors only) and applies it to the test data. Prediction is performed based on support vector machines. The function `svm_x` uses the function `svm` from the package `e1071`.

Usage

```
svm_x(Xlearn,Zlearn=NULL,Ylearn,Xtest,Ztest=NULL,ordered=NULL,nbgene=NULL,...)
```

Arguments

<code>Xlearn</code>	A <code>nlearn</code> x <code>p</code> matrix giving the microarray predictors for the learning data set.
<code>Zlearn</code>	A <code>nlearn</code> x <code>q</code> matrix giving the clinical predictors for the learning data set. This argument is ignored.
<code>Ylearn</code>	A numeric vector of length <code>nlearn</code> giving the class membership of the learning observations, coded as 0,1.
<code>Xtest</code>	A <code>ntest</code> x <code>p</code> matrix giving the microarray predictors for the test data set.
<code>Ztest</code>	A <code>ntest</code> x <code>q</code> matrix giving the clinical predictors for the test data set. This argument is ignored.
<code>ordered</code>	A vector of length <code>p</code> giving the order of the microarray predictors in terms of relevance for prediction. For instance, if the three first elements of <code>ordered</code> are 30,2,2400, it means that the most relevant genes are the genes in the 30th, 2nd and 2400th columns of the gene expression data matrix <code>Xlearn</code> . Note: if <code>ordered=NULL</code> , the columns of <code>Xlearn</code> and <code>Xtest</code> are assumed to be already ordered.
<code>nbgene</code>	The number of genes to be selected for use in dimension reduction. Default is <code>nbgene=NULL</code> , in which case all genes are used.
<code>...</code>	Other arguments to be passed to the function <code>svm</code> from the <code>e1071</code> package.

Details

This function is included in the package for comparison.

Value

A list with the element

<code>prediction</code>	A numeric vector of length <code>nrow(Xtest)</code> giving the predicted class for each observation from the test data set.
-------------------------	---

Author(s)

Anne-Laure Boulesteix (<http://www.slcmsr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

See Also

[testclass](#), [testclass_simul](#), [simulate](#), [plsr_x_pv](#), [plsr_xz](#), [plsr_xz_pv](#), [rf_z](#), [logistic_z](#).

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating xlearn, zlearn, ylearn, xtest, ztest
xlearn<-matrix(rnorm(3000),30,100)
ylearn<-sample(0:1,30,replace=TRUE)
xtest<-matrix(rnorm(2000),20,100)

my.prediction1<-svm_x(Xlearn=xlearn,Ylearn=ylearn,Xtest=xtest)

ordered<-sample(100)
my.prediction2<-svm_x(Xlearn=xlearn,Ylearn=ylearn,Xtest=xtest,ordered=ordered,nbgene=20)
```

testclass	<i>Evaluating a classification method based on several learning sub-sets</i>
-----------	--

Description

This function evaluates classifiers built using microarray data and/or clinical predictors, based on several pairs of learning and test data sets.

Usage

```
testclass(x=NULL,y,z=NULL,learningsets,classifier,ncomp=0:3,
varsel=NULL,nbgene=NULL,fold=10,...)
```

Arguments

<code>x</code>	A $n \times p$ matrix giving the gene expression levels of p genes (columns) for n patients (rows).
<code>y</code>	A numeric vector of length n giving the class membership of the n patients, coded as $0, \dots, K-1$ (where K is the number of classes).
<code>z</code>	A $n \times q$ data frame giving the q clinical predictors for the n patients. Nominal variables should be given as factors, variables with an at least ordinal scale should be given as numeric.
<code>learningsets</code>	A matrix with <code>niter</code> rows giving the indices of the arrays to be included in the learning sets for the <code>niter</code> iterations, as generated by the function <code>generate.learningsets</code> . The i -th row gives the indices of the arrays included in the learning set for the i -th iteration. For instance, in LOOCV, the i -th row of the matrix <code>learningsets</code> contains all the integers from 1 to n except i . Note that an observation may be included twice or more in the same learning set (for instance in bootstrap sampling).
<code>classifier</code>	The function used to construct a classifier. The function must have the same structure as <code>plsrf_xz_pv</code> .
<code>ncomp</code>	The candidate numbers of PLS components (if PLS dimension reduction is used).
<code>varsel</code>	A <code>niter</code> \times p matrix giving the indices of the genes ordered by the chosen gene selection criterion. For example, the element in the first row and the first column is the index of the gene that is ranked best using the first learning set.
<code>nbgene</code>	The number of genes to use for classifier construction. Default is <code>nbgene=NULL</code> , corresponding to all genes.
<code>fold</code>	The number of folds for the pre-validation step. See Boulesteix et al (2008) for more details. Default is <code>fold=10</code> .
<code>...</code>	Other arguments to be passed to the function <code>cforest_control</code> from the <code>party</code> package or to the function <code>svm</code> from the package <code>e1071</code> , depending on the specified classifier.

Details

For an overview of different methods used to generate the learning sets defined by `generate.learningsets`, see Boulesteix et al (2007). These methods include (repeated) cross-validation, subsampling, bootstrap sampling.

Value

<code>error</code>	A numeric vector of length <code>niter</code> giving the misclassification rate for each iteration.
<code>bestncomp</code>	A numeric vector of length <code>niter</code> giving the best number of (pre-validated) PLS components, as obtained using the model selection method based on the out-of-bag error by Boulesteix et al (returned only for the classifiers <code>plsrf_xz_pv</code> , <code>plsrf_xz</code> , <code>plsrf_x_pv</code> , <code>plsrf_x</code>).

OOB A list of length `niter`, whose elements are numeric vectors of the same length as `ncomp` giving the out-of-bag error of the forest constructed with the corresponding number of (pre-validated) PLS components (returned only for the classifiers `plsrf_xz_pv`, `plsrf_xz`, `plsrf_x_pv`, `plsrf_x`, `rf_z`. For `rf_z`, no model selection is performed: OOB is just the out-of-bag error of the constructed forest.)

Author(s)

Anne-Laure Boulesteix (<http://www.slcmsr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

Boulesteix AL, Strobl C, Augustin T, Daumer D, 2007. Evaluating microarray-based classifiers: an overview. Submitted.

See Also

`testclass_simul`, `simulate`, `generate.learningsets`, `plsrf_xz_pv`, `plsrf_x_pv`, `plsrf_xz`, `plsrf_x`, `rf_z`, `svm_x`, `logistic_z`.

Examples

```
# load MAclinical library
# library(MAclinical)

# Generate data
x<-matrix(rnorm(20000),100,200)
z<-matrix(rnorm(500),100,5)
y<-sample(0:1,100,replace=TRUE)

# Generate learningsets (5-fold CV)
my.learningsets<-generate.learningsets(n=100,method="CV",fold=5)

# Evaluate accuracy of the PLS-PV-RF method
my.eval<-testclass(x=x,y=y,z=z,learningsets=my.learningsets,classifier=plsrf_xz_pv,ncomp=5,
varsel=NULL,nbgene=NULL,fold=10)

# With variable selection
my.varsel<-matrix(0,5,200)
for (i in 1:5)
{
  my.varsel[i,]<-order(abs(studentt.stat(X=x[my.learningsets[i,]],
L=y[my.learningsets[i,]]+1)),decreasing=TRUE)
}

my.eval<-testclass(x=x,y=y,z=z,learningsets=my.learningsets,classifier=plsrf_xz_pv,ncomp=5,
varsel=my.varsel,nbgene=15,fold=10)
```

Description

This function evaluates classifiers built using microarray data and/or clinical predictors, based on simulated data generated using the functions `simuldata_list` and `simuldatacluster_list` (see [simulate](#)).

Usage

```
testclass_simul(datalist, nlearn=100, classifier, ncomp=0:3, nbgene=NULL,
  varsel=NULL, fold=10, ...)
```

Arguments

<code>datalist</code>	A list of niter simulated data sets as generated by the functions <code>simuldata_list</code> and <code>simuldatacluster_list</code> (see simulate).
<code>nlearn</code>	The number of observations to be included in the learning data set. It must be smaller than the total number of observations of the data sets.
<code>classifier</code>	The function used to construct a classifier. The function must have the same structure as plsrfr_xz_pv .
<code>ncomp</code>	The candidate numbers of PLS components (if PLS dimension reduction is used).
<code>nbgene</code>	The number of genes to use for classifier construction. Default is <code>nbgene=NULL</code> , corresponding to all genes.
<code>varsel</code>	A niter x p matrix giving the indices of the genes ordered by the chosen gene selection criterion. For example, the element in the first row and the first column is the index of the gene that is ranked best using in the first simulation iteration.
<code>fold</code>	The number of folds for the pre-validation step, if any. See Boulesteix et al (2008) for more details. Default is <code>fold=10</code> .
<code>...</code>	Other arguments to be passed to the function <code>cforest_control</code> from the <code>party</code> package or to the function <code>svm</code> from the package <code>e1071</code> , depending on the specified classifier.

Details

See Boulesteix et al (2008).

Value

<code>error</code>	A numeric vector of length <code>niter</code> giving the misclassification rate for each iteration.
<code>bestncomp</code>	A numeric vector of length <code>niter</code> giving the best number of (pre-validated) PLS components, as obtained using the model selection method based on the out-of-bag error by Boulesteix et al (returned only for the classifiers <code>plsrf_xz_pv</code> , <code>plsrf_xz</code> , <code>plsrf_x_pv</code> , <code>plsrf_x</code>).
<code>O0B</code>	A list of length <code>niter</code> , whose elements are numeric vectors of the same length as <code>ncomp</code> giving the out-of-bag error of the forest constructed with the corresponding number of (pre-validated) PLS components (returned only for the classifiers <code>plsrf_xz_pv</code> , <code>plsrf_xz</code> , <code>plsrf_x_pv</code> , <code>plsrf_x</code> , <code>rf_z</code> . For <code>rf_z</code> , no model selection is performed: <code>O0B</code> is just the out-of-bag error of the constructed forest.)

Author(s)

Anne-Laure Boulesteix (<http://www.slcmsr.net/boulesteix>)

References

Boulesteix AL, Porzelius C, Daumer M, 2008. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. Submitted.

See Also

`testclass`, `plsrf_xz_pv`, `simulate`, `plsrf_xz_pv`, `plsrf_x_pv`, `plsrf_xz`, `plsrf_x`, `rf_z`, `svm_x`, `logistic_z`.

Examples

```
# load MAclinical library
# library(MAclinical)

# Generating 3 simulated data sets
my.data<-simuldata_list(niter=3,n=100,p=150,psig=10,q=5,muX=2,muZ=1)

# Perform prediction of the 60 last observations using the first 40 observations,
# based on PLS (without pre-validation) and random forests

testclass_simul(my.data,nlearn=40,classifier=plsrf_xz)
```