Package ‘CMA’

March 17, 2020

Type Package

Title Synthesis of microarray-based classification

Version 1.44.0

Date 2009-09-14

Author Martin Slawski <ms@cs.uni-sb.de>, Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>, Christoph Bernau <berna@ibe.med.uni-muenchen.de>

Maintainer Roman Hornung <hornung@ibe.med.uni-muenchen.de>

Depends R (>= 2.10), methods, stats, Biobase

Suggests MASS, class, nnet, glmnet, e1071, randomForest, plsgenomics, gbm, mgcv, corpcor, limma, st, mvtnorm

Description This package provides a comprehensive collection of various microarray-based classification algorithms
both from Machine Learning and Statistics.
Variable Selection, Hyperparameter tuning, Evaluation and Comparison can be performed combined or stepwise in a user-friendly environment.

Collate classes.r GenerateLearningsets.r GeneSelection.r tune.r
classification.r evaluation.r join.r compare.r Planarplot.r
compBoostCMA.r dldaCMA.r ElasticNetCMA.r ldaCMA.r fdaCMA.r
flexdaCMA.r gbmCMA.r knnCMA.r LassoCMA.r ldaCMA.r nnetCMA.r
pknnCMA.r plrCMA.r pls_ldaCMA.r pls_rfCMA.r pmrCMA.r qdaCMA.r
rfCMA.r scdaCMA.r weighted_mcr.r wmc.r shrinkldaCMA.r svmCMA.r
filter.r internals.r

License GPL (>= 2)

biocViews Classification, DecisionTree

git_url https://git.bioconductor.org/packages/CMA

git_branch RELEASE_3_10

git_last_commit 0908cc0

git_last_commit_date 2019-10-29

Date/Publication 2020-03-16

R topics documented:

CMA-package ................................................................. 3
Barplot ................................................................. 4
R topics documented:

- best .......................................................... 5
- boxplot ......................................................... 6
- classification ............................................... 7
- classification-methods .................................. 9
- cloutput-class ............................................. 9
- clvarseloutput-class ..................................... 10
- compare ...................................................... 11
- compare-methods ......................................... 13
- compBoostCMA ............................................. 14
- compBoostCMA-methods ................................ 16
- dldaCMA ...................................................... 16
- dldaCMA-methods ......................................... 18
- ElasticNetCMA .............................................. 19
- ElasticNetCMA-methods ................................ 21
- evaloutput-class ......................................... 21
- evaluation .................................................. 22
- evaluation-methods ....................................... 24
- fdaCMA ......................................................... 24
- fdaCMA-methods ........................................... 26
- filter .......................................................... 27
- flexdaCMA ..................................................... 27
- flexdaCMA-methods ....................................... 29
- ftable .......................................................... 30
- gbmCMA ......................................................... 30
- gbmCMA-methods .......................................... 32
- GenerateLearningsets .................................. 33
- genesel-class .............................................. 34
- GeneSelection ............................................... 35
- GeneSelection-methods ................................ 38
- golub .......................................................... 38
- internals ..................................................... 39
- join ........................................................... 39
- join-methods ............................................... 40
- khan ............................................................ 40
- knnCMA ......................................................... 41
- knnCMA-methods .......................................... 43
- LassoCMA ..................................................... 43
- LassoCMA-methods ....................................... 45
- ldaCMA ......................................................... 45
- ldaCMA-methods .......................................... 47
- learningsets-class ...................................... 47
- metCMA ......................................................... 48
- metCMA-methods .......................................... 50
- obsinfo ......................................................... 51
- pknnCMA ....................................................... 52
- pknnCMA-methods ........................................ 53
- Planarplot ................................................... 53
- Planarplot-methods ...................................... 54
- plot ........................................................... 55
- plot-tuningresult ......................................... 56
- plrCMA ........................................................ 58
- plrCMA-methods .......................................... 59
The aim of the package is to provide a user-friendly environment for the evaluation of classification methods using gene expression data. A strong focus is on combined variable selection, hyperparameter tuning, evaluation, visualization and comparison of (up to now) 21 classification methods from three main fields: Discriminant Analysis, Neural Networks and Machine Learning. Although the package has been created with the intention to be used for Microarray data, it can as well be used in various \((p > n)\)-scenarios.
Most Important Steps for the workflow are:

1. Generate evaluation datasets using `GenerateLearningsets`
2. (Optionally): Perform variable selection using `GeneSelection`
3. (Optionally): Perform hyperparameter tuning using `tune`
5. Repeat 2.-4. based on 1. for several methods: `compBoostCMA`, `dldaCMA`, `ElasticNetCMA`, `fdaCMA`, `flexdaCMA`, `gbmCMA`, `knnCMA`, `LassoCMA`, `LASSO`, `pknCMA`, `plCMA`, `pls_lldaCMA`, `pls_lrCMA`, `pls_rfcma`, `pnnCMA`, `qdaCMA`, `rfCMA`, `scdaCMA`, `shrinkldaCMA`, `svmCMA`
6. Evaluate the results from 5. using `evaluation` and make a comparison by calling `compare`

**Author(s)**

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>
Maintainer: Christoph Bernau <bernau@ibe.med.uni-muenchen.de>.

**References**


---

**Barplot**

**Barplot of variable importance**

**Description**

This method can be seen as a visual pendant to `toplist`. The plot visualizes variable importance by a barplot. The height of the barplots correspond to variable importance. What variable importance exactly means depends on the method chosen when calling `GeneSelection`, s. `genesel`.

**Arguments**

- **x**: An object of class `genesel`
- **top**: Number of top genes whose variable importance should be displayed. Defaults to 10.
- **iter**: Iteration number (`learningset`) for which variable importance should be displayed.
- **...**: Further graphical options passed to `barplot`.
**Value**

No return.

**Note**

Note the following

- If `scheme = "multiclass"`, only one plot will be made. Otherwise, one plot will be made for each binary scenario (depending on whether "scheme" is "one-vs-all" or "pairwise").
- Variable importance do not make sense for variable selection (ranking) methods that are essentially discrete, such as the Wilcoxon-Rank sum statistic or the Kruskal-Wallis statistic.
- For the methods "lasso","elasticnet","boosting" the number of nonzero coefficients can be very small, resulting in bars of height zero if top has been chosen too large.

**Author(s)**

Martin Slawski <ms@cs.uni-sb.de>

Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

**References**


**See Also**

genesel, GeneSelection, toplist

---

**Description**

In this package hyperparameter tuning is performed by an inner cross-validation step for each learningset. A grid of values is tried and evaluated in terms of the misclassification rate, the results are saved in an object of class `tuningresult`. This method displays (separately for each learningset) the hyperparameter/ hyperparameter combination that showed the best results. Note that this must not be unique; in this case, only one combination is displayed.

**Usage**

`best(object, ...)`

**Arguments**

- `object` An object of class `tuningresult`.
- `...` Currently unused argument.

**Value**

A list with elements equal to the number of different learningsets. Each element contains the best hyperparameter combination and the corresponding misclassification rate.
Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

tune

boxplot

Make a boxplot of the classifier evaluation

Description

This method displays the slot scores of performance scores of an object of class evaloutput.

Arguments

x          An object of class evaloutput.
...        Further graphical parameters passed to the classical boxplot function.

Value

The only return is a boxplot.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References


See Also

evaluation
classification

General method for classification with various methods

Description

Most general function in the package, providing an interface to perform variable selection, hyper-parameter tuning and classification in one step. Alternatively, the first two steps can be performed separately and can then be plugged into this function.

For S4 method information, s. classification-methods.

Usage

classification(X, y, f, learningsets, genesel, genesellist = list(), nbgene, classifier, tuneres, tuninglist = list(), trace = TRUE, models=FALSE,...)

Arguments

X Gene expression data. Can be one of the following:
  • A matrix. Rows correspond to observations, columns to variables.
  • A data.frame, when f is not missing (s. below).
  • An object of class ExpressionSet.

y Class labels. Can be one of the following:
  • A numeric vector.
  • A factor.
  • A character if X is an ExpressionSet that specifies the phenotype variable.
  • missing, if X is a data.frame and a proper formula f is provided.

WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learningsets An object of class learningsets. May be missing, then the complete datasets is used as learning set.

genesel Optional (but usually recommended) object of class genesel containing variable importance information for the argument learningsets

genesellist In the case that the argument genesel is missing, this is an argument list passed to GeneSelection. If both genesel and genesellist are missing, no variable selection is performed.

nbgene Number of best genes to be kept for classification, based on either genesel or the call to GeneSelection using genesellist. In the case that both are missing, this argument is not necessary. note:

  • If the gene selection method has been one of "lasso","elasticnet","boosting", nbgene will be reset to min(s,nbgene) where s is the number of nonzero coefficients.
  • if the gene selection scheme has been "one-vs-all","pairwise" for the multiclass case, there exist several rankings. The top nbgene will be kept of each of them, so the number of effective used genes will sometimes be much larger.
classification

classifier  Name of function ending with CMA indicating the classifier to be used.

tuneres  Analogous to the argument genesel - object of class tuningresult containing information about the best hyperparameter choice for the argument learningsets.

tuninglist  Analogous to the argument genesel-list. In the case that the argument tuneres is missing, this in argument list passed to tune. If both tuneres and tuninglist are missing, no variable selection is performed. warning: Note that if a user-defined hyperparameter grid is passed, this will result in a list within a list: tuninglist = list(grids=list(argname = c())), s. example. warning: Contrary to tune, if tuninglist is an empty list (default), no hyperparameter tuning will be performed at all. To use pre-defined hyperparameter grids, the argument is tuninglist = list(grids = list()).

trace  Should progress be traced ? Default is TRUE.

models  a logical value indicating whether the model object shall be returned

...  Further arguments passed to the function classifier.

Details

For details about hyperparameter tuning, consult tune.

Value

A list of objects of class cloutput and clvarseloutput, respectively; its length equals the number of different learningsets. The single elements of the list can conveniently be combined using the join function. The results can be analyzed and evaluated by various measures using the method evaluation.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

References


See Also

GeneSelection, tune, evaluation, compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA

Examples

### a simple k-nearest neighbour example
### datasets
## Not run: plot(x)
data(golub)
golubY <- golub[,1]
golubX <- as.matrix(golub[,-1])
### learningsets
set.seed(111)
lset <- GenerateLearningsets(y=golubY, method = "CV", fold=5, strat =TRUE)
### 1. GeneSelection
selttest <- GeneSelection(golubX, golubY, learningsets = lset, method = "t.test")
### 2. tuning
tunek <- tune(golubX, golubY, learningsets = lset, genesel = selttest, nbgene = 20, classifier = knnCMA)
### 3. classification
knn1 <- classification(golubX, golubY, learningsets = lset, genesel = selttest,
tuneres = tunek, nbgene = 20, classifier = knnCMA)
### steps 1.-3. combined into one step:
knn2 <- classification(golubX, golubY, learningsets = lset,
geneseellist = list(method = "t.test"), classifier = knnCMA,
tuninglist = list(grids = list(k = c(1:8))), nbgene = 20)
### show and analyze results:
knnjoin <- join(knn2)
show(knn2)
eval <- evaluation(knn2, measure = "misclassification")
show(eval)
summary(eval)
boxplot(eval)
## End(Not run)

classification-methods

General method for classification with various methods

Description

Perform classification for the following signatures:

Methods

X = "matrix", y = "numeric", f = "missing" signature 1
X = "matrix", y = "factor", f = "missing" signature 2
X = "data.frame", y = "missing", f = "formula" signature 3
X = "ExpressionSet", y = "character", f = "missing" signature 4

For further argument and output information, consult classification.

cloutput-class

"cloutput"

Description

Object returned by one of the classifiers (functions ending with CMA)
clvarseloutput-class

Slots

learnind: Vector of indices that indicates which observations were used in the learning set.
y: Actual (true) class labels of predicted observations.
yhat: Predicted class labels by the classifier.
prob: A numeric matrix whose rows equals the number of predicted observations (length of y/yhat) and whose columns equal the number of different classes in the learning set. Rows add up to one. Entry j,k of this matrix contains the probability for the j-th predicted observation to belong to class k. Can be a matrix of NAs, if the classifier used does not provide any probabilities
method: Name of the classifier used.
mode: character, one of "binary" (if the number of classes in the learning set is two) or multiclass (if it is more than two).
model: List containing the constructed classifiers.

Methods

show Use show(cloutput-object) for brief information
ftable Use ftable(cloutput-object) to obtain a confusion matrix/cross-tabulation of y vs. yhat, s. ftable,cloutput-method.
plot Use plot(cloutput-object) to generate a probability plot of the matrix prob described above, s. plot,cloutput-method
roc Use roc(cloutput-object) to compute the empirical ROC curve and the Area Under the Curve (AUC) based on the predicted probabilities, s. roc,cloutput-method

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

clvarseloutput compBoostCMA, dldaCMA, ElasticNetCMA, fdacMA, flexdaCMA, gbmCMA, knnCMA, ldacMA, LassoCMA, nnetCMA, pknnCMA, plrcMA, pls_ldacMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdacMA, rfCMA, scdacMA, shrinkldaCMA, svmCMA

clvarseloutput-class "clvarseloutput"

Description

Object returned by all classifiers that can perform variable selection or compute variable importance. These are:

- Random Forest, s. rfCMA,
- Componentwise Boosting, s. compBoostCMA,
- LASSO-logistic regression, s. LassoCMA,
- ElasticNet-logistic regression, s. ElasticNetCMA

Objects of class clvarseloutput extend both the class cloutput and varsel, s. below.
compare

Slots

learnind: Vector of indices that indicates which observations where used in the learning set.
y: Actual (true) class labels of predicted observations.
yhat: Predicted class labels by the classifier.
prob: A numeric matrix whose rows equals the number of predicted observations (length of y/yhat) and whose columns equal the number of different classes in the learning set. Rows add up to one. Entry j,k of this matrix contains the probability for the j-th predicted observation to belong to class k. Can be a matrix of NAs, if the classifier used does not provide any probabilities.
method: Name of the classifier used.
mode: character, one of "binary" (if the number of classes in the learning set is two) or multiclass (if it is more than two).
varsel: numeric vector of variable importance measures (for Random Forest) or absolute values of regression coefficients (for the other three methods mentionned above) (from which the majority will be zero).

Extends

Class "cloutput", directly. Class "varseloutput", directly.

Methods

show Use show(cloutput-object) for brief information
ftable Use ftable(cloutput-object) to obtain a confusion matrix/cross-tabulation of y vs. yhat, s. ftable,cloutput-method.
plot Use plot(cloutput-object) to generate a probability plot of the matrix prob described above, s. plot,cloutput-method
roc Use roc(cloutput-object) to compute the empirical ROC curve and the Area Under the Curve (AUC) based on the predicted probabilities, s. roc,cloutput-method

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

rfCMA, compBoostCMA, LassoCMA, ElasticNetCMA

compare Compare different classifiers

Description

Classifiers can be evaluated separately using the method evaluation. Normally, several classifiers are used for the same dataset and their performance is compared. This comparison procedure is essentially facilitated by this method. For S4 method information, s. compare-methods
Usage

```r
compare(clresultlist, measure = c("misclassification", "sensitivity", "specificity", "average probability", "brier score", "auc"), aggfun = meanrm, plot = FALSE, ...)
```

Arguments

- `clresultlist` A list of lists (!) of objects of class `cloutput` or `clvarseloutput`. Each inner list is usually returned by `classification`. Additionally, the different list elements of the outer list should have been created by different classifiers, s. also example below.
- `measure` A character vector containing one or more of the elements listed below. By default, all measures are computed, using `evaluation` with `scheme = "iterationwise"`. Note that "sensitivity", "specificity", "auc" cannot be computed for the multiclass case.
  - "misclassification" The misclassification rate.
  - "sensitivity" The sensitivity or 1-false negative rate. Can only be computed for binary classification.
  - "specificity" The specificity or 1-false positive rate. Can only be computed for binary classification.
  - "average probability" The average probability assigned to the correct class. Requirement is that the used classifier provides probability estimations. The optimum performance is 1.
  - "brier score" The Brier Score is generally defined as \( \sum_{i} \sum_{k} (I(y_i=k)-P(k))^2 \), with \( I() \) denoting the indicator function and \( P(k) \) the estimated probability for class \( k \). The optimum performance is 0.
  - "auc" The Area under the Curve (AUC) belonging to the empirical ROC curve computed from the estimated probabilities and the true class labels. Can only be computed for binary classification and if "scheme = iterationwise", s. below. S. also roc, cloutput-method.
- `aggfun` Function that determines how performance among different iterations are aggregated. Default is `meanrm`, which computes the mean using `na.rm=T`. Other possible choices are quantiles.
- `plot` Should the performance of different classifiers be visualized by a joint boxplot? Default is `FALSE`.
- `...` Further arguments passed to `boxplot` in the case that `plot = TRUE`.

Value

A `data.frame` with rows corresponding to the compared classifiers and columns to the performance measures, aggregated by `aggfun`, s. above.

Note

If more than one measure is computed and `plot = TRUE`, one separate plot is created for each of them.
**compare-methods**

**Author(s)**

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

**References**

Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97, 77-87


**See Also**

classification, evaluation

**Examples**

```r
## Not run:
### compare the performance of several discriminant analysis methods
### for the Khan dataset:
data(khan)
khanX <- as.matrix(khan[,,-1])
khanY <- khan[,1]
set.seed(27611)
fiveCV10iter <- GenerateLearningsets(y=khanY, method = "CV", fold = 5, niter = 2, strat = TRUE)
### candidate methods: DLDA, LDA, QDA, pls_LDA, sclda
class_dlda <- classification(X = khanX, y=khanY, learningsets = fiveCV10iter, classifier = dldaCMA)
### perform GeneSelection for LDA, FDA, QDA (using F-Tests):
genesel_da <- GeneSelection(X=khanX, y=khanY, learningsets = fiveCV10iter, method = "f.test")
###
class_lda <- classification(X = khanX, y=khanY, learningsets = fiveCV10iter, classifier = ldaCMA, genesel = genesel_da)
class_qda <- classification(X = khanX, y=khanY, learningsets = fiveCV10iter, classifier = qdaCMA, genesel = genesel_da)
### We now make a comparison concerning the performance (sev. measures):
### first, collect in a list:
dalike <- list(class_dlda, class_lda, class_qda)
### use pre-defined compare function:
comparison <- compare(dalike, plot = TRUE, measure = c("misclassification", "brier score", "average probability"))
print(comparison)
## End(Not run)
```

**Description**

Compare different classifiers for the following signatures:
Methods

clresultlist = "list" signature 1

For further argument and output information, consult compare

compBoostCMA Componentwise Boosting

Description

Roughly speaking, Boosting combines 'weak learners' in a weighted manner in a stronger ensemble.

'Weak learners' here consist of linear functions in one component (variable), as proposed by Buehlmann and Yu (2003).

It also generates sparsity and can as well be as used for variable selection alone. (s. GeneSelection).

For S4 method information, see compBoostCMA-methods.

Usage

compBoostCMA(X, y, f, learnind, loss = c("binomial", "exp", "quadratic"), mstop = 100, nu = 0.1, models=FALSE, ...)

Arguments

X Gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when f is not missing (s. below).
  - An object of class ExpressionSet.

y Class labels. Can be one of the following:
  - A numeric vector.
  - A factor.
  - A character if X is an ExpressionSet that specifies the phenotype variable.
  - missing, if X is a data.frame and a proper formula f is provided.

WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learnind An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

loss Character specifying the loss function - one of "binomial" (LogitBoost), "exp" (AdaBoost), "quadratic"(L2Boost).

mstop Number of boosting iterations, i.e. number of updates to perform. The default (100) does not necessarily produce good results, therefore usage of tune for this argument is highly recommended.

nu Shrinkage factor applied to the update steps, defaults to 0.1. In most cases, it suffices to set nu to a very low value and to concentrate on the optimization of mstop.

models a logical value indicating whether the model object shall be returned

... Currently unused arguments.
**Details**

The method is partly based on code from the package `mboost` from T. Hothorn and P. Buehlmann. The algorithm for the multiclass case is described in Lutz and Buehlmann (2006) as ‘rowwise updating’.

**Value**

An object of class `clvarseloutput`.

**Author(s)**

Martin Slawski <ms@cs.uni-sb.de>

Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

**References**


**See Also**

dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls ldaCMA, pls lrCMA, pls rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA

**Examples**

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[-1])
### select learning set
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run componentwise (logit)-boosting (not tuned)
result <- compBoostCMA(X=golubX, y=golubY, learnind=learnind, mstop = 500)
### show results
show(result)
ftable(result)
plot(result)
### multiclass example:
```
### load Khan data

data(khan)

### extract class labels

khanY <- khan[,1]

### extract gene expression

khanX <- as.matrix(khan[,-1])

### select learning set

set.seed(111)

learnind <- sample(length(khanY), size=floor(ratio*length(khanY)))

### run componentwise multivariate (logit)-boosting (not tuned)

result <- compBoostCMA(X=khanX, y=khanY, learnind=learnind, mstop = 1000)

### show results

show(result)

ftable(result)

plot(result)

---

**compBoostCMA-methods**

*Componentwise Boosting*

**Description**

Roughly speaking, Boosting combines 'weak learners' in a weighted manner in a stronger ensemble.

'Weak learners' here consist of linear functions in one component (variable), as proposed by Buehlmann and Yu (2003).

It also generates sparsity and can as well be as used for variable selection alone. (s. GeneSelection.)

**Methods**

- X = "matrix", y = "numeric", f = "missing" signature 1
- X = "matrix", y = "factor", f = "missing" signature 2
- X = "data.frame", y = "missing", f = "formula" signature 3
- X = "ExpressionSet", y = "character", f = "missing" signature 4

For further argument and output information, consult compBoostCMA.

---

**dldaCMA**

*Diagonal Discriminant Analysis*

**Description**

Performs a diagonal discriminant analysis under the assumption of a multivariate normal distribution in each classes (with equal, diagonally structured) covariance matrices. The method is also known under the name 'naive Bayes' classifier.

For S4 method information, see dldaCMA-methods.

**Usage**

dldaCMA(X, y, f, learnind, models=FALSE, ...)

Arguments

\(X\) Gene expression data. Can be one of the following:
- A matrix. Rows correspond to observations, columns to variables.
- A data.frame, when \(f\) is not missing (s. below).
- An object of class ExpressionSet.

\(y\) Class labels. Can be one of the following:
- A numeric vector.
- A factor.
- A character if \(X\) is an ExpressionSet that specifies the phenotype variable.
- missing, if \(X\) is a data.frame and a proper formula \(f\) is provided.

**WARNING**: The class labels will be re-coded to range from 0 to \(K-1\), where \(K\) is the total number of different classes in the learning set.

\(f\) A two-sided formula, if \(X\) is a data.frame. The left part correspond to class labels, the right to variables.

learnind An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

models a logical value indicating whether the model object shall be returned

... Currently unused argument.

Value

An object of class cloutput.

Note

As opposed to linear or quadratic discriminant analysis, variable selection is not strictly necessary.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

Discriminant Analysis and Statistical Pattern Recognition.
*Wiley, New York*

See Also

compBoostCMA, ElasticNetCMA, FdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls ldaCMA, pls lrCMA, pls rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA
Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[,-1])
### select learning set
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run DLDA
dldaresult <- dldaCMA(X=golubX, y=golubY, learnind=learnind)
### show results
show(dldaresult)
ftable(dldaresult)
plot(dldaresult)
### multiclass example:
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression
khanX <- as.matrix(khan[,-1])
### select learning set
set.seed(111)
learnind <- sample(length(khanY), size=floor(ratio*length(khanY)))
### run LDA
ldaresult <- dldaCMA(X=khanX, y=khanY, learnind=learnind)
### show results
show(ldaresult)
ftable(ldaresult)
plot(ldaresult)
```

dldaCMA-methods  
Diagonal Discriminant Analysis

Description

Performs a diagonal discriminant analysis under the assumption of a multivariate normal distribution in each classes (with equal, diagonally structured) covariance matrices. The method is also known under the name ‘naive Bayes’ classifier.

Methods

- `X = "matrix", y = "numeric", f = "missing"` signature 1
- `X = "matrix", y = "factor", f = "missing"` signature 2
- `X = "data.frame", y = "missing", f = "formula"` signature 3
- `X = "ExpressionSet", y = "character", f = "missing"` signature 4

For further argument and output information, consult `dldaCMA`.
Description

Zou and Hastie (2004) proposed a combined L1/L2 penalty for regularization and variable selection. The Elastic Net penalty encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The computation is done with the function glmpath from the package of the same name. The method can be used for variable selection alone, s. GeneSelection. For S4 method information, see ElasticNetCMA-methods.

Usage

ElasticNetCMA(X, y, f, learnind, norm.fraction = 0.1, alpha=0.5, models=FALSE, ...)

Arguments

X  Gene expression data. Can be one of the following:
   • A matrix. Rows correspond to observations, columns to variables.
   • A data.frame, when f is not missing (s. below).
   • An object of class ExpressionSet. note: by default, the predictors are scaled to have unit variance and zero mean. Can be changed by passing standardize = FALSE via the ... argument.

y  Class labels. Can be one of the following:
   • A numeric vector.
   • A factor.
   • A character if X is an ExpressionSet that specifies the phenotype variable.
   • missing, if X is a data.frame and a proper formula f is provided.

   WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f  A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learnind  An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

norm.fraction  L1 Shrinkage intensity, expressed as the fraction of the coefficient L1 norm compared to the maximum possible L1 norm (corresponds to fraction = 1). Lower values correspond to higher shrinkage. Note that the default (0.1) need not produce good results, i.e. tuning of this parameter is recommended.

alpha  The elasticnet mixing parameter, with 0<alpha<= 1. The penalty is defined as (1-alpha)/2||beta||_2^2+alpha||beta||_1. alpha=1 is the lasso penalty; Currently ‘alpha<0.01’ not reliable, unless you supply your own lambda sequence.

models  a logical value indicating whether the model object shall be returned

...  Further arguments passed to the function glmpath from the package of the same name.
Value

An object of class clvarseloutput.

Note

For a strongly related method, see LassoCMA.
Up to now, this method can only be applied to binary classification.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernaub@ibe.med.uni-muenchen.de>

References


See Also

compBoostCMA, dldaCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA,
svmCMA

Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[-1])
### select learning set
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run ElasticNet - penalized logistic regression (no tuning)
result <- ElasticNetCMA(X=golubX, y=golubY, learnind=learnind, norm.fraction = 0.2, alpha=0.5)
show(result)
ftable(result)
plot(result)
```
**ElasticNetCMA-methods**

Classfication and variable selection by the ElasticNet

**Description**

Zou and Hastie (2004) proposed a combined L1/L2 penalty for regularization and variable selection. The Elastic Net penalty encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The computation is done with the function `glmnet` from the package of the same name.

**Methods**

- `X = "matrix", y = "numeric", f = "missing"` signature 1
- `X = "matrix", y = "factor", f = "missing"` signature 2
- `X = "data.frame", y = "missing", f = "formula"` signature 3
- `X = "ExpressionSet", y = "character", f = "missing"` signature 4

For references, further argument and output information, consult `ElasticNetCMA-evaloutput-class` "evaloutput"

**Description**

Object returned by the method `evaluation`.

**Slots**

- `score`: A numeric vector of performance scores whose length depends on "scheme", s. below. It equals the number of iterations (number of different datasets) if "scheme = iterationwise" and the number of all observations in the complete dataset otherwise. As not necessarily all observations must be predicted at least one time, score can also contain NAs for those observations not classified at all.
- `measure`: performance measure used, s. `evaluation`.
- `scheme`: scheme used, s. `evaluation`
- `method`: name of the classifier that has been evaluated.

**Methods**

- `show` Use `show(evaloutput-object)` for brief information.
- `summary` Use `summary(evaloutput-object)` to apply the classic `summary()` function to the slot `score`, s. `summary,evaloutput-method`
- `boxplot` Use `boxplot(evaloutput-object)` to display a boxplot of the slot score, s. `boxplot,evaloutput-method`
- `obsinfo` Use `obsinfo(evaloutput-object,threshold)` to display all observations consistently correctly or incorrectly classified (depending on the value of the argument `threshold`), s. `obsinfo`
Description
The performance of classifiers can be evaluated by six different measures and two different schemes
that are described more precisely below.
For S4 method information, s. evaluation-methods.

Usage
evaluation(clresult, cltrain = NULL, cost = NULL, y = NULL, measure = c("misclassification", "sensitivity", "specificity", "average probability", "brier score", "auc", "0.632", "0.632+"),
scheme = c("iterationwise", "observationwise", "classwise"))

Arguments
clresult A list of objects of class cloutput or clvarseloutput
cltrain An object of class cloutput in which the whole dataset was used as learning set. Only used if method = "0.632" or method = "0.632+" in order to obtain an estimation for the resubstitution error rate.
cost An optional cost matrix used if measure = "misclassification". If it is not specified (default), the cost is the usual indicator loss. Otherwise, entry i,j of cost quantifies the loss when the true class is class i-1 and the predicted class is j-1, provided the conventional coding 0,...,K-1 in the case of K classes is used. Usually, the matrix contains only non-negative entries with zeros on the diagonal, but this is not obligatory. Make sure that the dimension of the matrix matches the number of classes.
y A vector containing the true class labels. Only needed if scheme = "classwise".
measure Performance measure to be used:
"misclassification" The misclassification rate.
"sensitivity" The sensitivity or 1-false negative rate. Can only be computed for binary classification.
"specificity" The specificity or 1-false positive rate. Can only be computed for binary classification.
"average probability" The average probability assigned to the correct class. Requirement is that the used classifier provides probability estimations. The optimum performance is 1.
"brier score" The Brier Score is generally defined as \( \sum_{i} \sum_{k} (I(y_i=k)-P(k))^2 \), with \( I() \) denoting the indicator function and \( P(k) \) the estimated probability for class k. The optimum performance is 0.
"auc" The Area under the Curve (AUC) belonging to the empirical ROC curve computed from the estimated probabilities and the true class labels. Can only be computed for binary classification and if "scheme = iterationwise", s. below. S. also roc, cloutput-method.

"0.632" The 0.632 estimator (s. reference) for the misclassification rate (applied iteration- or) observationwise, if bootstrap learning sets have been used. Note that cltrain must be provided.

"0.632+" The 0.632+ estimator (s. reference) for the misclassification rate (applied iteration- or) observationwise, if bootstrap learning sets have been used. Note that cltrain must be provided.

scheme

"iterationwise" The performance measures listed above are computed for each different iteration, i.e. each different learningset.

"observationwise" The performance measures listed above (except for "auc") are computed separately for each observation classified one or several times, depending on the learningset scheme.

"classwise" The performance measures (exceptions: "auc", "0.632", "0.632+") are computed separately for each class, averaged over both iterations and observations.

Value

An object of class evaloutput.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

References


See Also

evaloutput, classification, compare

Examples

### simple linear discriminant analysis example using bootstrap datasets:
### datasets:
data(golub)
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,2:11])
### generate 25 bootstrap datasets
set.seed(333)
bootds <- GenerateLearningsets(y = golubY, method = "bootstrap", ntrain = 30, niter = 10, strat = TRUE)
### run classification()

```r
ldalist <- classification(X=golubX, y=golubY, learningsets = bootds, classifier=ldaCMA)
### Evaluation:
eval_iter <- evaluation(ldalist, scheme = "iter")
eval_obs <- evaluation(ldalist, scheme = "obs")
show(eval_iter)
show(eval_obs)
summary(eval_iter)
summary(eval_obs)
### auc with boxplot
eval_auc <- evaluation(ldalist, scheme = "iter", measure = "auc")
boxplot(eval_auc)
### which observations have often been misclassified ?
obsinfo(eval_obs, threshold = 0.75)
```

---

**Evaluation-methods**  
**Evaluation of classifiers**

### Description

Evaluate classifiers for the following signatures:

### Methods

**clresult = "list"** signature 1

For further argument and output information, consult `evaluation`.

---

**fdaCMA**  
**Fisher's Linear Discriminant Analysis**

### Description

Fisher’s Linear Discriminant Analysis constructs a subspace of ’optimal projections’ in which classification is performed. The directions of optimal projections are computed by the function `cancor` from the package `stats`. For an exhaustive treatment, see e.g. Ripley (1996).

For S4 method information, see `fdaCMA-methods`.

### Usage

```r
fdaCMA(X, y, f, learnind, comp = 1, plot = FALSE, models=FALSE)
```

### Arguments

- **X**: Gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when `f` is not missing (s. below).
  - An object of class `ExpressionSet`.
- **y**: Class labels. Can be one of the following:
  - A numeric vector.
fdacMA

• A factor.
• A character if X is an ExpressionSet that specifies the phenotype variable.
• missing, if X is a data.frame and a proper formula f is provided. WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.
learnind An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.
comp Number of discriminant coordinates (projections) to compute. Default is one, must be smaller than or equal to K-1, where K is the number of classes.
plot Should the projections onto the space spanned by the optimal projection directions be plotted ? Default is FALSE.
models a logical value indicating whether the model object shall be returned

Value
An object of class cloutput.

Note
Excessive variable selection has usually to performed before fdacMA can be applied in the p > n setting. Not reducing the number of variables can result in an error message.

Author(s)
Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References
Ripley, B.D. (1996)
Pattern Recognition and Neural Networks.
Cambridge University Press

See Also
compBoostCMA, dldaCMA, ElasticNetCMA, fdacMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA

Examples
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,2:11])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run FDA
fdaresult <- fdaCMA(X=golubX, y=golubY, learnind=learnind, comp = 1, plot = TRUE)
### show results
show(fdaresult)
ftable(fdaresult)
plot(fdaresult)
### multiclass example:
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression from first 10 genes
khanX <- as.matrix(khan[,2:11])
### select learningset
set.seed(111)
learnind <- sample(length(khanY), size=floor(ratio*length(khanY)))
### run FDA
fdaresult <- fdaCMA(X=khanX, y=khanY, learnind=learnind, comp = 2, plot = TRUE)
### show results
show(fdaresult)
ftable(fdaresult)
plot(fdaresult)

Fisher’s Linear Discriminant Analysis

Fisher’s Linear Discriminant Analysis constructs a subspace of ‘optimal projections’ in which classification is performed. The directions of optimal projections are computed by the function `cancor` from the package `stats`. For an exhaustive treatment, see e.g. Ripley (1996).

Methods

- `X = "matrix", y = "numeric", f = "missing"` signature 1
- `X = "matrix", y = "factor", f = "missing"` signature 2
- `X = "data.frame", y = "missing", f = "formula"` signature 3
- `X = "ExpressionSet", y = "character", f = "missing"` signature 4

For references, further argument and output information, consult `fdaCMA`. 

**Filter functions for Gene Selection**

**Description**

The functions listed above are usually not called by the user but via `GeneSelection`.

**Usage**

```r
\texttt{ttest(X, y, learnind, ...)}
\texttt{welchtest(X, y, learnind, ...)}
\texttt{ftest(X, y, learnind, ...)}
\texttt{kruskaltest(X, y, learnind, ...)}
\texttt{limmatest(X, y, learnind, ...)}
\texttt{golubcrit(X, y, learnind, ...)}
\texttt{rfe(X, y, learnind, ...)}
\texttt{shrinkcat(X, y, learnind, ...)}
```

**Arguments**

- `X` A numeric matrix of gene expression values.
- `y` A numeric vector of class labels.
- `learnind` An index vector specifying the observations that belong to the learning set.
- `...` Currently unused argument.

**Value**

An object of class `varseloutput`.

**References**


---

**Flexible Discriminant Analysis**

**Description**

This method is experimental.

It is easy to show that, after appropriate scaling of the predictor matrix `X`, Fisher’s Linear Discriminant Analysis is equivalent to Discriminant Analysis in the space of the fitted values from the linear regression of the `nlearn x K` indicator matrix of the class labels on `X`. This gives rise to 'nonlinear discriminant analysis’ methods that expand `X` in a suitable, more flexible basis. In order to avoid overfitting, penalization is used. In the implemented version, the linear model is replaced by a generalized additive one, using the package mgcv.

For S4 method information, s. `flexdaCMA-methods`.
Usage

`flexdaCMA(X, y, f, learnind, comp = 1, plot = FALSE, models=FALSE, ...)`

Arguments

- **X**: Gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when `f` is not missing (s. below).
  - An object of class ExpressionSet.

- **y**: Class labels. Can be one of the following:
  - A numeric vector.
  - A factor.
  - A character if `X` is an ExpressionSet that specifies the phenotype variable.
  - missing, if `X` is a data.frame and a proper formula `f` is provided.

**WARNING**: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

- **f**: A two-sided formula, if `X` is a data.frame. The left part correspond to class labels, the right to variables.

- **learnind**: An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

- **comp**: Number of discriminant coordinates (projections) to compute. Default is one, must be smaller than or equal to K-1, where K is the number of classes.

- **plot**: Should the projections onto the space spanned by the optimal projection directions be plotted? Default is FALSE.

- **models**: a logical value indicating whether the model object shall be returned

... Further arguments passed to the function `gam` from the package `mgcv`.

Value

An object of class `cloutput`.

Note

Excessive variable selection has usually to performed before `flexdaCMA` can be applied in the p > n setting. Recall that the original predictor dimension is even enlarged, therefore, it should be applied only with very few variables.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

Ripley, B.D. (1996)
Pattern Recognition and Neural Networks.
*Cambridge University Press*
See Also
compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA

Examples

### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 5 genes
golubX <- as.matrix(golub[,2:6])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run flexible Discriminant Analysis
result <- flexdaCMA(X=golubX, y=golubY, learnind=learnind, comp = 1)
### show results
show(result)
ftable(result)
plot(result)

flexdaCMA-methods

Flexible Discriminant Analysis

Description

This method is experimental.

It is easy to show that, after appropriate scaling of the predictor matrix \( X \), Fisher’s Linear Discriminant Analysis is equivalent to Discriminant Analysis in the space of the fitted values from the linear regression of the \( n_{\text{learn}} \times K \) indicator matrix of the class labels on \( X \). This gives rise to ‘nonlinear discriminant analysis’ methods that expand \( X \) in a suitable, more flexible basis. In order to avoid overfitting, penalization is used. In the implemented version, the linear model is replaced by a generalized additive one, using the package mgcv.

Methods

\( X = \"\text{matrix}\", y = \"\text{numeric}\", f = \"\text{missing}\" \) signature 1
\( X = \"\text{matrix}\", y = \"\text{factor}\", f = \"\text{missing}\" \) signature 2
\( X = \"\text{data.frame}\", y = \"\text{missing}\", f = \"\text{formula}\" \) signature 3
\( X = \"\text{ExpressionSet}\", y = \"\text{character}\", f = \"\text{missing}\" \) signature 4

For further argument and output information, consult flexdaCMA.
Cross-tabulation of predicted and true class labels

Description

An object of class cloutput contains (among others) the slot y and yhat. The former contains the true, the last the predicted class labels. Both are cross-tabulated in order to obtain a so-called confusion matrix. Counts out of the diagonal are misclassifications.

Arguments

- **x**
  - An object of class cloutput
- ... Currently unused argument.

Value

No return.

Author(s)

Martin Slawski <martin.slawski@campus.lmu.de>
Anne-Laure Boulesteix [http://www.slcmsr.net/boulesteix](http://www.slcmsr.net/boulesteix)

See Also

For more advanced evaluation: [evaluation](#)

Tree-based Gradient Boosting

Description

Roughly speaking, Boosting combines 'weak learners' in a weighted manner in a stronger ensemble. This method calls the function gbm.fit from the package gbm. The 'weak learners' are simple trees that need only very few splits (default: 1).

For S4 method information, see [gbmCMA-methods](#).

Usage

gbmCMA(X, y, f, learnind, models=FALSE,...)

---

ftable

Cross-tabulation of predicted and true class labels
gbmCMA

Arguments

\(X\)  
Gene expression data. Can be one of the following:
- A matrix. Rows correspond to observations, columns to variables.
- A data.frame, when \(f\) is not missing (s. below).
- An object of class ExpressionSet.

\(y\)  
Class labels. Can be one of the following:
- A numeric vector.
- A factor.
- A character if \(X\) is an ExpressionSet that specifies the phenotype variable.
- missing, if \(X\) is a data.frame and a proper formula \(f\) is provided.

**WARNING**: The class labels will be re-coded to range from 0 to \(K-1\), where \(K\) is the total number of different classes in the learning set.

\(f\)  
A two-sided formula, if \(X\) is a data.frame. The left part correspond to class labels, the right to variables.

\(learnind\)  
An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

\(models\)  
a logical value indicating whether the model object shall be returned

...  
Further arguments passed to the function gbm.fit from the package of the same name. Worth mentioning are

\(ntrees\)  
Number of trees to fit (size of the ensemble), defaults to 100. This parameter should be optimized using \texttt{tune}.

\(shrinkage\)  
The learning rate (default is 0.001). Usually fixed to a very low value.

\(distribution\)  
Loss function to be used. Default is "bernoulli", i.e. LogitBoost, a (less robust) alternative is "adaboost".

\(interaction.depth\)  
Number of splits used by the 'weak learner' (single decision tree). Default is 1.

Value

An onject of class \texttt{cloutput}.

Note

Up to now, this method can only be applied to binary classification.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

The state of boosting.
*Computing Science and Statistics, 31*:172-181

See Also

compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plRcMA, pls_1ldaCMA, pls_1rCMA, pls_rfCMA, pnnCMA, qdaCMA, rFcMA, scdaCMA, shrinkldaCMA, svmCMA

Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[,-1])
### select learning set
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run tree-based gradient boosting (no tuning)
gbmresult <- gbmCMA(X=golubX, y=golubY, learnind=learnind, n.trees = 500)
show(gbmresult)
ftable(gbmresult)
plot(gbmresult)
```

---

**gbmCMA-methods**  
**Tree-based Gradient Boosting**

### Description

Roughly speaking, Boosting combines ‘weak learners’ in a weighted manner in a stronger ensemble. This method calls the function `gbm.fit` from the package `gbm`. The ‘weak learners’ are simple trees that need only very few splits (default: 1).

### Methods

- `X = "matrix", y = "numeric", f = "missing"`  signature 1
- `X = "matrix", y = "factor", f = "missing"`  signature 2
- `X = "data.frame", y = "missing", f = "formula"`  signature 3
- `X = "ExpressionSet", y = "character", f = "missing"`  signature 4

For further argument and output information, consult gbmCMA.
GenerateLearningsets

Repeated Divisions into learn- and test sets

Description

Due to very small sample sizes, the classical division learnset/testset does not give accurate information about the classification performance. Therefore, several different divisions should be used and aggregated. The implemented methods are discussed in Braga-Neto and Dougherty (2003) and Molinaro et al. (2005) whose terminology is adopted.

This function is usually the basis for all deeper analyses.

Usage

GenerateLearningsets(n, y, method = c("LOOCV", "CV", "MCCV", "bootstrap"),
fold = NULL, niter = NULL, ntrain = NULL, strat = FALSE)

Arguments

n
The total number of observations in the available data set. May be missing if y is provided instead.

y
A vector of class labels, either numeric or a factor. Must be given if strat=TRUE or n is not specified.

method
Which kind of scheme should be used to generate divisions into learning sets and test sets? Can be one of the following:
"LOOCV" Leaving-One-Out Cross Validation.
"CV" (Ordinary) Cross-Validation. Note that fold must as well be specified.
"MCCV" Monte-Carlo Cross Validation, i.e. random divisions into learning sets with ntrain(s.below) observations and tests sets with ntrain observations.
"bootstrap" Learning sets are generated by drawing n times with replacement from all observations. Those not drawn not all form the test set.

cold
Gives the number of CV-groups. Used only when method="CV"

niter
Number of iterations (s.details).

ntrain
Number of observations in the learning sets. Used only when method="MCCV".

strat
Logical. Should stratified sampling be performed, i.e. the proportion of observations from each class in the learning sets be the same as in the whole data set?

Details

- When method="CV", niter gives the number of times the whole CV-procedure is repeated. The output matrix has then foldx niter 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

An object of class learningsets
Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

References

Is cross-validation valid for small-sample microarray classification?
*Bioinformatics*, 20(3), 374-380

Prediction error estimation: a comparison of resampling methods.
*Bioinformatics*, 21(15), 3301-3307


See Also

`learningsets`, `GeneSelection`, `tune`, `classification`

Examples

```r
# LOOCV
loo <- GenerateLearningsets(n=40, method="LOOCV")
show(loo)

# five-fold-CV
CV5 <- GenerateLearningsets(n=40, method="CV", fold=5)
show(CV5)

# MCCV
mccv <- GenerateLearningsets(n=40, method = "MCCV", niter=3, ntrain=30)
show(mccv)

# Bootstrap
boot <- GenerateLearningsets(n=40, method="bootstrap", niter=3)

# stratified five-fold-CV
set.seed(113)
classlabels <- sample(1:3, size = 50, replace = TRUE, prob = c(0.3, 0.5, 0.2))
CV5strat <- GenerateLearningsets(y = classlabels, method="CV", fold=5, strat = TRUE)
show(CV5strat)
```

description

Object returned from a call to `GeneSelection`
Slots

rankings: A list of matrices. For the two-class case and the multi-class case where a genuine multi-class method has been used for variable selection, the length of the list is one. Otherwise, it is named according to the different binary scenarios (e.g. 1 vs 3). Each list element is a matrix with rows corresponding to iterations (different learningsets) and columns to variables. Each row thus contains an index vector representing the order of the variables with respect to their variable importance (s. slot importance)

importance: A list of matrices, with the same structure as described for the slot rankings. Each row of these matrices are ordered according to rankings and contain the variable importance measure (absolute value of test statistic or regression coefficient).

method: Name of the method used for variable selection, s. GeneSelection.

scheme: The scheme used in the case of a non-binary response, one of "pairwise", "one-vs-all" or "multiclass".

Methods

show Use show(genesel-object) for brief information

toplist Use toplist(genesel-object,k=10,iter=1) to display the top first 10 variables and their variable importance for the first iteration (first learningset), s.toplist.

plot Use plot(genesel-object,k=10,iter=1) to display a barplot of the variable importance of the top first 10 variables, s.plot,genesel-method

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

GeneSelection

Description

For different learning data sets as defined by the argument learningsets, this method ranks the genes from the most relevant to the less relevant using one of various 'filter' criteria or provides a sparse collection of variables (Lasso, ElasticNet, Boosting). The results are typically used for variable selection for the classification procedure that follows. For S4 class information, s. GeneSelection-methods.

Usage

GeneSelection(X, y, f, learningsets, method = c("t.test", "welch.test", "wilcox.test", "f.test", "k.test", "golub", "shrinkcat"), scheme, trace = TRUE, ...)

GeneSelection  General method for variable selection with various methods

Genesel
Arguments

X  Gene expression data. Can be one of the following:
   • A matrix. Rows correspond to observations, columns to variables.
   • A data.frame, when f is not missing (s. below).
   • An object of class ExpressionSet.

y  Class labels. Can be one of the following:
   • A numeric vector.
   • A factor.
   • A character if X is an ExpressionSet.
     • missing, if X is a data.frame and a proper formula f is provided.

f  A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learningsets  An object of class learningsets. May be missing, then the complete datasets is used as learning set.

method  A character specifying the method to be used:
   t.test  two-sample t.test (equal variances for both classes assumed).
   welch.test  Welch modification of the t.test (unequal variances for both classes).
   wilcox.test  Wilcoxon rank sum test.
   f.test  F test belonging to the linear hypothesis that the mean is the same for all classes. Usually used for the multiclass scheme, is equivalent to method = t.test in the two-class case.
   kruskal.test  Multi-class generalization of the Wilcoxon rank sum test and the nonparametric pendant to the F test, respectively.
   limma  ’Moderated t’ statistic for the two-class case and ’moderated F’ statistic for the multiclass case, described in Smyth (2003). Requires the package limma.
   rfe  One-step Recursive Feature Elimination, based on the Support Vector Machine. The method is decribed in Guyon et al. (2002). Requires the package e1071. Take care that appropriate hyperparameters are passed by the ... argument.
   rf  Random Forest Variable Importance Measure. Requires the package randomForest
   lasso  L1 penalized logistic regression leads to sparsity with respect to the variables used. Calls the function LassoCMA, which requires the package glmpath. warning: Take care that appropriate hyperparameters are passed by the ... argument.
   elasticnet  Penalized logistic regression with both L1 and L2 penalty, claimed by Zhou and Hastie (2004) to select ‘variable groups’. Calls the function ElasticNetCMA, which requires the package glmpath. warning: Take care that appropriate hyperparameters are passed by the ... argument.
   boosting  Componentwise boosting (Buehlmann and Yu, 2003) has been shown to mimic the LASSO (Efron et al., 2004; Buehlmann and Yu, 2006). Calls the function compBoostCMA Take care that appropriate hyperparameters are passed by the ... argument.
   golub  The (theoretically unfounded) variable selection criterion used by Golub et al. (1999), s. golub.
   shrinkcat  The correlation-adjusted t-score from Zuber and Strimmer (2009)
GeneSelection

scheme The scheme to be used in the case of a non-binary response. Must be one of "pairwise", "one-vs-all" or "multiclass". The last case only makes sense if method is one of f.test, limma, rf, boosting, which can directly be applied to the multi class case.

trace Should the progress be traced? Default is TRUE.

... Further arguments passed to the function performing variable selection, s. method.

Value An object of class genesel.

Notemost of the methods described above are only apt for the binary classification case. The only ones that can be used without restriction in the multiclass case are

- f.test
- kruskal.test
- rf
- boosting

For the rest, pairwise or one-vs-all schemes are used.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

References


See Also

filter, GenerateLearningsets, tune, classification

Examples

```r
# load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,-1])
### Generate five different learningsets
set.seed(111)
five <- GenerateLearningsets(y=golubY, method = "CV", fold = 5, strat = TRUE)
### simple t-test:
selttest <- GeneSelection(golubX, golubY, learningsets = five, method = "t.test")
### show result:
show(selttest)
toplist(selttest, k = 10, iter = 1)
plot(selttest, iter = 1)
```

GeneSelection-methods  General method for variable selection with various methods

Description

Performs gene selection for the following signatures:

Methods

- \( \text{X} = \text{"matrix"}, \text{y} = \text{"numeric"}, \text{f} = \text{"missing"} \) signature 1
- \( \text{X} = \text{"matrix"}, \text{y} = \text{"factor"}, \text{f} = \text{"missing"} \) signature 2
- \( \text{X} = \text{"data.frame"}, \text{y} = \text{"missing"}, \text{f} = \text{"formula"} \) signature 3
- \( \text{X} = \text{"ExpressionSet"}, \text{y} = \text{"character"}, \text{f} = \text{"missing"} \) signature 4

For further argument and output information, consult GeneSelection.

Golub

ALL/AML dataset of Golub et al. (1999)

Description

s. below

Usage

data(golub)
Format

A data frame with 38 observations and 3052 variables. The first column (named golub.cl) contains the tumor classes (ALL = acute lymphatic leukaemia, AML = acute myeloid leukaemia).
golub.cl: a factor with levels ALL, AML.
X2-X3051: Gene expression values.

Source

Adopted from the dataset in the package multtest.

References

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P.,
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.
Science 286, 531-537.

Examples

data(golub)

---

**Internals**

**Internal functions**

**Description**

Not intended to be called directly by the user.

**join**

*Combine list elements returned by the method classification*

**Description**

The method classification returns a list of class cloutput or clvarseloutput. It is often more convenient to work with an object of class cloutput instead with a whole list, e.g. because the convenience method defined for that class can be used.

For S4 method information, s. join-methods

**Usage**

join(cloutputlist)

**Arguments**

cloutputlist A list of objects of classes cloutput or clvarseloutput, usually that returned by a call to the method classification. The only requirement for a successful join is that the used dataset and classifier are the same for each list element.
Value
An object of class `cloutput`. **warning:** If the elements of cloutputlist have originally been of class `clvarseloutput`, the slot `varsel` will be dropped!

Note
The result of the `join` method is incompatible with the methods `evaluation`, `compare`. These require the lists returned by `classification`.

See Also
`classification`, `evaluation`

---

**join-methods**

*Combine list elements returned by the method classification*

### Description
The list of objects of class `cloutput` can be unified into one object for the following signatures:

### Methods

`cloutputlist = "list"` signature 1

For further argument and output information, consult `join`.

### khan

*Small blue round cell tumor dataset of Khan et al. (2001)*

### Description
s. below

### Usage

data(khan)

### Format
A data frame with 63 observations on the following 2309 variables. The first column (named `khanY`) contains the tumor classes (BL = Burkitt Lymphoma, EWS = Ewing Sarcoma, NB = Neuro Blastoma, RMS = Rhabdomyosarcoma). 

`khanY`: a factor with levels BL EWS NB RMS X2-X2309: Gene expression values.

### Source
Adopted from the dataset in the package `pamr`.
References

Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.

Examples

data(khan)

knnCMA

Description

Ordinary k nearest neighbours algorithm from the very fast implementation in the package class.
For S4 method information, see knnCMA-methods.

Usage

knnCMA(X, y, f, learnind, models=FALSE, ...)

Arguments

X Gene expression data. Can be one of the following:
  • A matrix. Rows correspond to observations, columns to variables.
  • A data.frame, when f is not missing (s. below).
  • An object of class ExpressionSet.

y Class labels. Can be one of the following:
  • A numeric vector.
  • A factor.
  • A character if X is an ExpressionSet that specifies the phenotype variable.
  • missing, if X is a data.frame and a proper formula f is provided.

  WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learnind An index vector specifying the observations that belong to the learning set. Must not be missing for this method.

models a logical value indicating whether the model object shall be returned

... Further arguments to be passed to knn from the package class, in particular the number of nearest neighbours to use (argument k).

Value

An object of class cloutput.
Note

Class probabilities are not returned. For a probabilistic variant of knn, see pknnCMA.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

Ripley, B.D. (1996)
Pattern Recognition and Neural Networks.
Cambridge University Press

See Also

cmpBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, ldaCMA, LassoCMA, nnetCMA,
pknnCMA, plrCMA, pls ldaCMA, pls lrCMA, pls rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA,
svmCMA

Examples

### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,-1])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run k-nearest neighbours
result <- knnCMA(X=golubX, y=golubY, learnind=learnind, k = 3)
### show results
show(result)
ftable(result)
### multiclass example:
### load Khan data
data(khan)
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression
khanX <- as.matrix(khan[,-1])
### select learningset
set.seed(111)
learnind <- sample(length(khanY), size=floor(ratio*length(khanY)))
### run knn
result <- knnCMA(X=khanX, y=khanY, learnind=learnind, k = 5)
### show results
show(result)
ftable(result)
**knnCMA-methods**

**Nearest Neighbours**

**Description**

Ordinary k nearest neighbours algorithm from the very fast implementation in the package `class`.

**Methods**

- \( X = \text{"matrix"}, y = \text{"numeric"}, f = \text{"missing"} \) signature 1
- \( X = \text{"matrix"}, y = \text{"factor"}, f = \text{"missing"} \) signature 2
- \( X = \text{"data.frame"}, y = \text{"missing"}, f = \text{"formula"} \) signature 3
- \( X = \text{"ExpressionSet"}, y = \text{"character"}, f = \text{"missing"} \) signature 4

For further argument and output information, consult `knnCMA`.

---

**LassoCMA**

**L1 penalized logistic regression**

**Description**

The Lasso (Tibshirani, 1996) is one of the most popular tools for simultaneous shrinkage and variable selection. Recently, Friedman, Hastie and Tibshirani (2008) have developed and algorithm to compute the entire solution path of the Lasso for an arbitrary generalized linear model, implemented in the package `glmnet`. The method can be used for variable selection alone, s. `GeneSelection`. For S4 method information, see `LassoCMA-methods`.

**Usage**

LassoCMA(X, y, f, learnind, norm.fraction = 0.1, models=FALSE,...)

**Arguments**

- **X**
  - Gene expression data. Can be one of the following:
    - A matrix. Rows correspond to observations, columns to variables.
    - A data.frame, when f is not missing (s. below).
    - An object of class ExpressionSet. note: by default, the predictors are scaled to have unit variance and zero mean. Can be changed by passing standardize = FALSE via the ... argument.

- **y**
  - Class labels. Can be one of the following:
    - A numeric vector.
    - A factor.
    - A character if \( X \) is an ExpressionSet that specifies the phenotype variable.
    - missing, if \( X \) is a data.frame and a proper formula f is provided.

**WARNING:** The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.
f A two-sided formula, if \(X\) is a \texttt{data.frame}. The left part correspond to class
labels, the right to variables.

\texttt{learnind} An index vector specifying the observations that belong to the learning set. May
be \texttt{missing}; in that case, the learning set consists of all observations and pre-
dictions are made on the learning set.

\texttt{norm.fraction} L1 Shrinkage intensity, expressed as the fraction of the coefficient L1 norm
compared to the maximum possible L1 norm (corresponds to fraction = 1).
Lower values correspond to higher shrinkage. Note that the default (0.1) need
not produce good results, i.e. tuning of this parameter is recommended.

\texttt{models} a logical value indicating whether the model object shall be returned

\texttt{...} Further arguments passed to the function \texttt{glmpath} from the package of the same
name.

\textbf{Value}

An object of class \texttt{clvarseloutput}.

\textbf{Note}

For a strongly related method, s. \texttt{ElasticNetCMA}.
Up to now, this method can only be applied to binary classification.

\textbf{Author(s)}

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

\textbf{References}

Tibshirani, R. (1996)
Regression shrinkage and selection via the lasso.
\textit{Journal of the Royal Statistical Society B}, 58(1), 267-288

Paths for Generalized Linear Models via Coordinate Descent
\url{http://www-stat.stanford.edu/~hastie/Papers/glmnet.pdf}

\textbf{See Also}

\texttt{compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, nnetCMA,
pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA,
svmCMA}

\textbf{Examples}

### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[,-1])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run L1 penalized logistic regression (no tuning)
lassoresult <- LassoCMA(x=golubX, y=golubY, learnind=learnind, norm.fraction = 0.2)
show(lassoresult)
ftable(lassoresult)
plot(lassoresult)

LassoCMA-methods

L1 penalized logistic regression

Description

The Lasso (Tibshirani, 1996) is one of the most popular tools for simultaneous shrinkage and variable selection. Recently, Friedman, Hastie and Tibshirani (2008) have developed an algorithm to compute the entire solution path of the Lasso for an arbitrary generalized linear model, implemented in the package `glmnet`. The method can be used for variable selection alone, s. GeneSelection

Methods

\[
\text{X} = \text{"matrix"}, \text{y} = \text{"numeric"}, \text{f} = \text{"missing"} \quad \text{signature 1}
\]
\[
\text{X} = \text{"matrix"}, \text{y} = \text{"factor"}, \text{f} = \text{"missing"} \quad \text{signature 2}
\]
\[
\text{X} = \text{"data.frame"}, \text{y} = \text{"missing"}, \text{f} = \text{"formula"} \quad \text{signature 3}
\]
\[
\text{X} = \text{"ExpressionSet"}, \text{y} = \text{"character"}, \text{f} = \text{"missing"} \quad \text{signature 4}
\]

For references, further argument and output information, consult LassoCMA.

ldaCMA

Linear Discriminant Analysis

Description

Performs a linear discriminant analysis under the assumption of a multivariate normal distribution in each class (with equal, but generally structured) covariance matrices. The function lda from the package MASS is called for computation.

For S4 method information, see ldaCMA-methods.

Usage

\[
\text{ldaCMA(X, y, f, learnind, models=FALSE, \ldots)}
\]

Arguments

\[X\]
Gene expression data. Can be one of the following:
- A matrix. Rows correspond to observations, columns to variables.
- A data.frame, when f is not missing (s. below).
- An object of class ExpressionSet.

\[y\]
Class labels. Can be one of the following:
• A numeric vector.
• A factor.
• A character if \( X \) is an ExpressionSet that specifies the phenotype variable.
• missing, if \( X \) is a data.frame and a proper formula \( f \) is provided.

**WARNING**: The class labels will be re-coded to range from 0 to \( K-1 \), where \( K \) is the total number of different classes in the learning set.

\( f \)
A two-sided formula, if \( X \) is a data.frame. The left part correspond to class labels, the right to variables.

\( \text{learnind} \)
An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

\( \text{models} \)
a logical value indicating whether the model object shall be returned

... Further arguments to be passed to \texttt{lda} from the package MASS

**Value**
An object of class \texttt{cloutput}.

**Note**
Excessive variable selection has usually to performed before \texttt{ldaCMA} can be applied in the \( p > n \) setting. Not reducing the number of variables can result in an error message.

**Author(s)**
Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

**References**
Discriminant Analysis and Statistical Pattern Recognition.
*Wiley, New York*

**See Also**
\texttt{compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rFCMA, scdaCMA, shrinkldaCMA, svmCMA}

**Examples**
```r
## Not run:
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,2:11])
### select learningset
ratio <- 2/3
```
```r
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run LDA
ldaresult <- ldaCMA(X=golubX, y=golubY, learnind=learnind)
### show results
show(ldaresult)
ftable(ldaresult)
plot(ldaresult)
### multiclass example:
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression from first 10 genes
khanX <- as.matrix(khan[,2:11])
### select learning set
set.seed(111)
learnind <- sample(length(khanY), size=floor(ratio*length(khanY)))
### run LDA
ldaresult <- ldaCMA(X=khanX, y=khanY, learnind=learnind)
### show results
show(ldaresult)
ftable(ldaresult)
plot(ldaresult)
```

---

### ldaCMA-methods

**Linear Discriminant Analysis**

**Description**

Performs a linear discriminant analysis for the following signatures:

**Methods**

- **X** = "matrix", **y** = "numeric", **f** = "missing" signature 1
- **X** = "matrix", **y** = "factor", **f** = "missing" signature 2
- **X** = "data.frame", **y** = "missing", **f** = "formula" signature 3
- **X** = "ExpressionSet", **y** = "character", **f** = "missing" signature 4

For further argument and output information, consult `ldaCMA`.

---

### learningsets-class

"learningsets"

**Description**

An object returned from `GenerateLearningsets` which is usually passed as arguments to `GeneSelection`, `tune` and `classification`.

Slots

learnmatrix: A matrix of dimension $n_{iter} \times n_{train}$. Each row contains the indices of those observations representing the learning set for one iteration. If method = CV, zeros appear due to rounding issues.

method: The method used to generate the learnmatrix. See GenerateLearningsets.

ntrain: Number of observations in one learning set. If method = CV, this number is not attained for all iterations, due to rounding issues.

iter: Number of iterations (different learning sets) that are stored in learnmatrix.

Methods

• showUse show(learningsets-object) for brief information.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

See Also

GenerateLearningsets, GeneSelection, tune, classification

nnetCMA

Description

This method provides access to the function nnet in the package of the same name that trains Feed-forward Neural Networks with one hidden layer.
For S4 method information, see nnetCMA-methods

Usage

nnetCMA(X, y, f, learnind, eigengenes = FALSE, models=FALSE,...)

Arguments

X Gene expression data. Can be one of the following:
• A matrix. Rows correspond to observations, columns to variables.
• A data.frame, when f is not missing (s. below).
• An object of class ExpressionSet.

y Class labels. Can be one of the following:
• A numeric vector.
• A factor.
• A character if X is an ExpressionSet that specifies the phenotype variable.
• missing, if X is a data.frame and a proper formula f is provided.
**WARNING:** The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

- `f` A two-sided formula, if `X` is a `data.frame`. The left part correspond to class labels, the right to variables.

- `learnind` An index vector specifying the observations that belong to the learning set. May be `missing`; in that case, the learning set consists of all observations and predictions are made on the learning set.

- `eigengenes` Should the training be performed be in the space of eigengenes obtained from a singular value decomposition of the Gene expression data matrix? Default is `FALSE`; in this case, variable selection is necessary to reduce the number of weights that have to be optimized.

- `models` a logical value indicating whether the model object shall be returned

Further arguments passed to the function `nnet` from the package of the same name.

Important parameters are:

- "size", i.e. the number of units in the hidden layer
- "decay" for weight decay.

### Value

An object of class `cloutput`.

### Note

- Excessive variable selection is usually necessary if `eigengenes = FALSE`
- Different runs of this method on the same dataset not necessarily produce the same results due to the fact that optimization for Feed-Forward Neural Networks is rather difficult and depends on the choice of (normally randomly chosen) starting values for the network weights.

### Author(s)

- Martin Slawski `<ms@cs.uni-sb.de>`
- Anne-Laure Boulesteix `<boulesteix@ibe.med.uni-muenchen.de>`
- Christoph Bernau `<bernau@ibe.med.uni-muenchen.de>`

### References

Ripley, B.D. (1996)
Pattern Recognition and Neural Networks.
*Cambridge University Press*

### See Also

- `compBoostCMA`, `dldaCMA`, `ElasticNetCMA`, `fdaCMA`, `flexdaCMA`, `gbmCMA`, `knnCMA`, `LassoCMA`, `nnetCMA`, `pknnCMA`, `plrCMA`, `pls_ladCMA`, `pls_lrcMA`, `pls_rfCMA`, `pnnCMA`, `qdaCMA`, `rfCMA`, `scdaCMA`, `shrinkldaCMA`, `svmCMA`
nnetCMA-methods

Feed-Forward Neural Networks

Description

This method provides access to the function `nnet` in the package of the same name that trains Feed-forward Neural Networks with one hidden layer.

Methods

```r
X = "matrix", y = "numeric", f = "missing" signature 1
X = "matrix", y = "factor", f = "missing" signature 2
X = "data.frame", y = "missing", f = "formula" signature 3
X = "ExpressionSet", y = "character", f = "missing" signature 4
```
For further argument and output information, consult `nnetCMA`.

Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,2:11])
### select learning set
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run nnet (not tuned)
nnetresult <- nnetCMA(X=golubX, y=golubY, learnind=learnind, size = 3, decay = 0.01)
### show results
show(nnetresult)
ftable(nnetresult)
plot(nnetresult)
### in the space of eigengenes (not tuned)
golubXfull <- as.matrix(golubX[,,-1])
nnetresult <- nnetCMA(X=golubXfull, y=golubY, learnind = learnind, eigengenes = TRUE, size = 3, decay = 0.01)
### show results
show(nnetresult)
ftable(nnetresult)
plot(nnetresult)
```
Classifiability of observations

Description

Some observations are harder to classify than others. It is frequently of interest to know which observations are consistently misclassified; these are candidates for outliers or wrong class labels.

Arguments

- **object**: An object of class `evaluation`, generated with `scheme = "observationwise"
- **threshold**: threshold value of (observation-wise) performance measure, s. `evaluation` that has to be exceeded in order to speak of consistent misclassification. If `measure = "average probability"`, then values below `threshold` are regarded as consistent misclassification. Note that the default values 1 is not sensible in that case
- **show**: Should the information be printed? Default is `TRUE`.

Details

As not all observation must have been classified at least once, observations not classified at all are also shown.

Value

A list with two components

- **misclassification**: A data.frame containing the indices of consistently misclassified observations and the corresponding performance measure.
- **notclassified**: The indices of those observations not classified at all, s. details.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References


See Also

`evaluation`
pknnCMA  
Probabilistic Nearest Neighbours

Description

Nearest neighbour variant that replaces the simple voting scheme by a weighted one (based on euclidean distances). This is also used to compute class probabilities.

For S4 class information, see pknnCMA-methods.

Usage

pknnCMA(X, y, f, learnind, beta = 1, k = 1, models=FALSE, ...)

Arguments

X  Gene expression data. Can be one of the following:
   • A matrix. Rows correspond to observations, columns to variables.
   • A data.frame, when f is not missing (s. below).
   • An object of class ExpressionSet.

y  Class labels. Can be one of the following:
   • A numeric vector.
   • A factor.
   • A character if X is an ExpressionSet that specifies the phenotype variable.
   • missing, if X is a data.frame and a proper formula f is provided.

   WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f  A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learnind  An index vector specifying the observations that belong to the learning set. Must not be missing for this method.

beta  Slope parameter for the logistic function which is used for the computation of class probabilities. The default value (1) need not produce reasonable results and can produce warnings.

k  Number of nearest neighbours to use.

models  a logical value indicating whether the model object shall be returned

...  Currently unused argument.

Details

The algorithm is as follows:

• Determine the k nearest neighbours
• For each class represented among these, compute the average euclidean distance.
• The negative distances are plugged into the logistic function with parameter beta.
• Classify into the class with highest probability.
pknnCMA-methods

Value

An object of class `cloutput`.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

`compBoostCMA`, `dldaCMA`, `ElasticNetCMA`, `fdaCMA`, `flexdaCMA`, `gbmCMA`, `knnCMA`, `ldaCMA`, `LassoCMA`, `nnetCMA`, `plrcMA`, `pls_ldaCMA`, `pls_lrCMA`, `pls_rfCMA`, `pnnCMA`, `qdaCMA`, `rfCMA`, `scdaCMA`, `shrinkldaCMA`, `svmCMA`

Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,-1])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run probabilistic k-nearest neighbours
result <- pknnCMA(X=golubX, y=golubY, learnind=learnind, k = 3)
### show results
show(result)
ftable(result)
plot(result)
```

---

**pknnCMA-methods**

**Probabilistic nearest neighbours**

**Description**

Nearest neighbour variant that replaces the simple voting scheme by a weighted one (based on euclidean distances). This is also used to compute class probabilities.

**Methods**

- `X = "matrix", y = "numeric", f = "missing"` signature 1
- `X = "matrix", y = "factor", f = "missing"` signature 2
- `X = "data.frame", y = "missing", f = "formula"` signature 3
- `X = "ExpressionSet", y = "character", f = "missing"` signature 4

For further argument and output information, consult `pknnCMA`. 
Planarplot

**Visualize Separability of different classes**

**Description**

Given two variables, the methods trains a classifier (argument classifier) based on these two variables and plots the resulting class regions, learning- and test observations in the plane.

Appropriate variables are usually found by `GeneSelection`.

For S4 method information, s. `Planarplot-methods`.

**Usage**

```r
Planarplot(X, y, f, learnind, predind, classifier, gridsize = 100, ...)
```

**Arguments**

- **X** Gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when f is not missing (s. below).
  - An object of class ExpressionSet.

- **y** Class labels. Can be one of the following:
  - A numeric vector.
  - A factor.
  - A character if X is an ExpressionSet that specifies the phenotype variable.
  - missing, if X is a data.frame and a proper formula f is provided.

- **f** A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

- **learnind** An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

- **predind** A vector containing exactly two indices that denote the two variables used for classification.

- **classifier** Name of function ending with CMA indicating the classifier to be used.

- **gridsize** The gridsize used for two-dimensional plotting.
  For both variables specified in predind, an equidistant grid of size gridsize is created. The resulting two grids are then combined to obtain gridsize*2 points in the real plane which are used to draw the class regions. Defaults to 100 which is usually a reasonable choice, but takes some time.

- **...** Further argument passed to classifier.

**Value**

No return.
Author(s)

Martin Slawski <ms@cs.uni-sb.de>

Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>. Idea is from the MLInterfaces package, contributed by Jess Mar, Robert Gentleman and Vince Carey.

See Also

GeneSelection, compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA

Examples

### simple linear discrimination for the golub data:

data(golub)
golubY <- golub[,1]
golubX <- as.matrix(golub[,-1])
golubn <- nrow(golubX)
set.seed(111)
learnind <- sample(golubn, size=floor(2/3*golubn))
Planarplot(X=golubX, y=golubY, learnind=learnind, predind=c(2,4),
           classifier=ldaCMA)
plot

Probability plot

Description
A popular way of visualizing the output of classifier is to plot, separately for each class, the predicted probability of each predicted observations for the respective class. For this purpose, the plot area is divided into $K$ parts, where $K$ is the number of classes. Predicted observations are assigned, according to their true class, to one of those parts. Then, for each part and each predicted observation, the predicted probabilities are plotted, displayed by coloured dots, where each colour corresponds to one class.

Arguments

x
An object of class cloutput whose slot probmatrix does not contain any missing value, i.e. probability estimations are provided by the classifier.

main
A title for the plot (character).

Value
No return.

Note
The plot usually only makes sense if a sufficiently large numbers of observations has been classified. This is usually achieved by running the classifier on several learning sets with the method classification. The output can then be processed via join to obtain an object of class cloutput to which this method can be applied.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References


See Also

cloutput
plot tuningresult

Visualize results of tuning

Description

After hyperparameter tuning using \texttt{tune} it is useful to see which choice of hyperparameters is suitable and how good the performance is.

Arguments

\begin{itemize}
  \item \texttt{x} An object of class \texttt{tuningresult}.
  \item \texttt{iter} Iteration number (\texttt{learningset}) for which tuning results should be displayed.
  \item \texttt{which} Character vector (maximum length is two) naming the arguments for which tuning results should be display. Default is \texttt{NULL}; if the number of tuned hyperparameter is less or equal than two, then the results for these hyperparameters will be plotted. If this number is two, then a \texttt{contour} plot will be made, otherwise a simple line segment plot. If the number of tuned hyperparameters exceeds two, then which may not be \texttt{NULL}.
  \item \ldots Further graphical options passed either to \texttt{plot} or \texttt{contour}.
\end{itemize}

Value

no return.

Note

Frequently, several hyperparameter (combinations) perform "best", s. also the remark in \texttt{best}.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References


See Also

tune, tuningresult
L2 penalized logistic regression

Description

High dimensional logistic regression combined with an L2-type (Ridge-)penalty. Multiclass case is also possible. For S4 method information, see plrCMA-methods.

Usage

plrCMA(X, y, f, learnind, lambda = 0.01, scale = TRUE, models=FALSE,...)

Arguments

- **X**: Gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when f is not missing (s. below).
  - An object of class ExpressionSet.

- **y**: Class labels. Can be one of the following:
  - A numeric vector.
  - A factor.
  - A character if X is an ExpressionSet that specifies the phenotype variable.
  - missing, if X is a data.frame and a proper formula f is provided.

  **WARNING**: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

- **f**: A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

- **learnind**: An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

- **lambda**: Parameter governing the amount of penalization. This hyperparameter should be tuned.

- **scale**: Scale the predictors as specified by X to have unit variance and zero mean.

- **models**: a logical value indicating whether the model object shall be returned

... Currently unused argument.

Value

An object of class cloutput.

Author(s)

Special thanks go to
Ji Zhu (University of Ann Arbor, Michigan)
Trevor Hastie (Stanford University)
who provided the basic code that was then adapted by
Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>.
References

*Biostatistics* 5:427-443.

See Also

compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, kNNCMA, ldaCMA, LassoCMA,
nnetCMA, pknnCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scCMA, shrinkldaCMA,
svmCMA

Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,1:-1])
### select learningset
ratio <- 2/3
set.seed(11)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run penalized logistic regression (no tuning)
plrresult <- plrCMA(X=golubX, y=golubY, learnind=learnind)
### show results
show(plrresult)
ftable(plrresult)
plot(plrresult)
### multiclass example:
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression from first 10 genes
khanX <- as.matrix(khan[,1:-1])
### select learningset
set.seed(11)
learnind <- sample(length(khanY), size=floor(ratio*length(khanY)))
### run penalized logistic regression (no tuning)
plrresult <- plrCMA(X=khanX, y=khanY, learnind=learnind)
### show results
show(plrresult)
ftable(plrresult)
plot(plrresult)
```

**plrCMA-methods**

L2 penalized logistic regression

Description

High dimensional logistic regression combined with an L2-type (Ridge-)penalty. Multiclass case is also possible.
### Methods

- **X = "matrix", y = "numeric", f = "missing"** signature 1
- **X = "matrix", y = "factor", f = "missing"** signature 2
- **X = "data.frame", y = "missing", f = "formula"** signature 3
- **X = "ExpressionSet", y = "character", f = "missing"** signature 4

For further argument and output information, consult `plrCMA`.

### Description

This method constructs a classifier that extracts Partial Least Squares components that are plugged into Linear Discriminant Analysis. The Partial Least Squares components are computed by the package `plsgenomics`.

For S4 method information, see `pls_ldaCMA-methods`.

### Usage

```r
pls_ldaCMA(X, y, f, learnind, comp = 2, plot = FALSE, models=FALSE)
```

### Arguments

- **X**
  - Gene expression data. Can be one of the following:
    - A matrix. Rows correspond to observations, columns to variables.
    - A data.frame, when f is not missing (s. below).
    - An object of class ExpressionSet.

- **y**
  - Class labels. Can be one of the following:
    - A numeric vector.
    - A factor.
    - A character if X is an ExpressionSet that specifies the phenotype variable.
    - missing, if X is a data.frame and a proper formula f is provided.

  **WARNING**: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

- **f**
  - A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

- **learnind**
  - An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

- **comp**
  - Number of Partial Least Squares components to extract. Default is 2 which can be suboptimal, depending on the particular dataset. Can be optimized using `tune`.

- **plot**
  - If comp <= 2, should the classification space of the Partial Least Squares components be plotted? Default is `FALSE`.

- **models**
  - a logical value indicating whether the model object shall be returned
Value

An object of class cloutput.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

Tumor classification by partial least squares using microarray gene expression data.
*Bioinformatics* 18, 39-50

Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.
*Briefings in Bioinformatics* 7:32-44.

See Also

compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA

Examples

```r
## Not run:
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression
khanX <- as.matrix(khan[-1])
### select learningset
set.seed(111)
learnind <- sample(length(khanY), size=floor(2/3*length(khanY)))
### run Shrunken Centroids classifier, without tuning
plsresult <- pls_ldaCMA(X=khanX, y=khanY, learnind=learnind, comp = 4)
### show results
show(plsresult)
ftable(plsresult)
plot(plsresult)
## End(Not run)
```

---

**pls_ldaCMA-methods**  
*Partial Least Squares combined with Linear Discriminant Analysis*

**Description**

This method constructs a classifier that extracts Partial Least Squares components that are plugged into Linear Discriminant Analysis. The Partial Least Squares components are computed by the package plsgenomics.
Methods

X = "matrix", y = "numeric", f = "missing" signature 1
X = "matrix", y = "factor", f = "missing" signature 2
X = "data.frame", y = "missing", f = "formula" signature 3
X = "ExpressionSet", y = "character", f = "missing" signature 4

For further argument and output information, consult pls_1daCMA.

Description

This method constructs a classifier that extracts Partial Least Squares components that form the covariates in a binary logistic regression model. The Partial Least Squares components are computed by the package plsgenomics.

For S4 method information, see pls_lrCMA-methods.

Usage

pls_lrCMA(X, y, f, learnind, comp = 2, lambda = 1e-4, plot = FALSE, models = FALSE)

Arguments

X Gene expression data. Can be one of the following:
  • A matrix. Rows correspond to observations, columns to variables.
  • A data.frame, when f is not missing (s. below).
  • An object of class ExpressionSet.

y Class labels. Can be one of the following:
  • A numeric vector.
  • A factor.
  • A character if X is an ExpressionSet that specifies the phenotype variable.
  • missing, if X is a data.frame and a proper formula f is provided.

WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learnind An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

comp Number of Partial Least Squares components to extract. Default is 2 which can be suboptimal, depending on the particular dataset. Can be optimized using tune.

lambda Parameter controlling the amount of L2 penalization for logistic regression, usually taken to be a small value in order to stabilize estimation in the case of separable data.

plot If comp <= 2, should the classification space of the Partial Least Squares components be plotted? Default is FALSE.

models a logical value indicating whether the model object shall be returned.
Value

An object of class cloutput.

Note

Up to now, only the two-class case is supported.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.
Briefings in Bioinformatics 7:32-44.

See Also

compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_lldaCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA, svmCMA

Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[,-1])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run PLS, combined with logistic regression
result <- pls_lrCMA(X=golubX, y=golubY, learnind=learnind)
### show results
show(result)
ftable(result)
plot(result)
```

Description

This method constructs a classifier that extracts Partial Least Squares components that form the covariates in a binary logistic regression model. The Partial Least Squares components are computed by the package plsgenomics.
Methods

X = "matrix", y = "numeric", f = "missing" signature 1
X = "matrix", y = "factor", f = "missing" signature 2
X = "data.frame", y = "missing", f = "formula" signature 3
X = "ExpressionSet", y = "character", f = "missing" signature 4

For further argument and output information, consult \texttt{pls_lrCMA}.

\begin{tabular}{ll}
\textbf{pls_rfCMA} & \textit{Partial Least Squares followed by random forests} \\
\end{tabular}

Description

This method constructs a classifier that extracts Partial Least Squares components used to generate Random Forests, s. \texttt{rfCMA}.

For \texttt{S4} method information, see \texttt{pls_rfCMA-methods}.

Usage

\texttt{pls_rfCMA(X, y, f, learnind, comp = 2 * nlevels(as.factor(y)), seed = 111, models=FALSE, ...)}

Arguments

\begin{itemize}
\item \textbf{X} Gene expression data. Can be one of the following:
  \begin{itemize}
  \item A \texttt{matrix}. Rows correspond to observations, columns to variables.
  \item A \texttt{data.frame}, when \texttt{f} is \textit{not} missing (s. below).
  \item An object of class \texttt{ExpressionSet}.
  \end{itemize}
\item \textbf{y} Class labels. Can be one of the following:
  \begin{itemize}
  \item A \texttt{numeric} vector.
  \item A \texttt{factor}.
  \item A \texttt{character} if \texttt{X} is an \texttt{ExpressionSet} that specifies the phenotype variable.
  \item \texttt{missing}, if \texttt{X} is a \texttt{data.frame} and a proper formula \texttt{f} is provided.
  \end{itemize}
\begin{warning}
\textbf{WARNING}: The class labels will be re-coded to range from 0 to K–1, where K is the total number of different classes in the learning set.
\end{warning}
\item \textbf{f} A two-sided formula, if \texttt{X} is a \texttt{data.frame}. The left part correspond to class labels, the right to variables.
\item \textbf{learnind} An index vector specifying the observations that belong to the learning set. May be \texttt{missing}; in that case, the learning set consists of all observations and predictions are made on the learning set.
\item \textbf{comp} Number of Partial Least Squares components to extract. Default ist two times the number of different classes.
\item \textbf{seed} Fix Random number generator seed to \texttt{seed}. This is useful to guarantee reproducibility of the results, due to the random component in the random Forest.
\item \textbf{models} a logical value indicating whether the model object shall be returned
\item \textbf{...} Further arguments to be passed to \texttt{randomForests} from the package of the same name.
\end{itemize}
Value

An object of class \texttt{cloutput}.

Author(s)

Martin Slawski \textless{}ms@cs.uni-sb.de\textgreater{}
Anne-Laure Boulesteix \textless{}boulesteix@ibe.med.uni-muenchen.de\textgreater{}

References

Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.
\textit{Briefings in Bioinformatics} 7:32-44.

See Also

\texttt{compBoostCMA}, \texttt{dldaCMA}, \texttt{ElasticNetCMA}, \texttt{fdacMA}, \texttt{flexdaCMA}, \texttt{gbmCMA}, \texttt{knnCMA}, \texttt{l1dCMA}, \texttt{LassoCMA}, \texttt{nnetCMA}, \texttt{pknCMA}, \texttt{p1rCMA}, \texttt{pl51dCMA}, \texttt{pl51rCMA}, \texttt{pnnCMA}, \texttt{qdaCMA}, \texttt{rPMA}, \texttt{scdCMA}, \texttt{shrinkldaCMA}, \texttt{svmCMA}

Examples

\begin{verbatim}
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[-1])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run PLS, combined with Random Forest
#result <- pls_rfCMA(X=golubX, y=golubY, learnind=learnind)
### show results
#show(result)
#ftable(result)
#plot(result)
\end{verbatim}
Methods

X = "matrix", y = "numeric", f = "missing" signature 1
X = "matrix", y = "factor", f = "missing" signature 2
X = "data.frame", y = "missing", f = "formula" signature 3
X = "ExpressionSet", y = "character", f = "missing" signature 4

For further argument and output information, consult `pls_rfcMA`.

---

Probabilistic Neural Networks

Probabilistic Neural Networks is the term Specht (1990) used for a Gaussian kernel estimator for the conditional class densities.

For S4 method information, see `pnnCMA-methods`.

Usage

```
pnnCMA(X, y, f, learnind, sigma = 1, models = FALSE)
```

Arguments

- **X** Gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when f is not missing (s. below).
  - An object of class ExpressionSet.
    Each variable (gene) will be scaled for unit variance and zero mean.
- **y** Class labels. Can be one of the following:
  - A numeric vector.
  - A factor.
  - A character if X is an ExpressionSet that specifies the phenotype variable.
    • missing, if X is a data.frame and a proper formula f is provided.

  **WARNING**: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.
- **f** A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.
- **learnind** An index vector specifying the observations that belong to the learning set. For this method, this must not be missing.
- **sigma** Standard deviation of the Gaussian Kernel used.
  This hyperparameter should be tuned, s. `tune`. The default is 1, but this generally does not lead to good results. Actually, this method reacts very sensitively to the value of sigma. Take care if warnings appear related to the particular choice.
- **models** a logical value indicating whether the model object shall be returned
Value

An object of class \texttt{cloutput}.

Note

There is actually no strong relation of this method to Feed-Forward Neural Networks, s. \texttt{nnetCMA}.

Author(s)

Martin Slawski \texttt{<ms@cs.uni-sb.de>}
Anne-Laure Boulesteix \texttt{<boulesteix@ibe.med.uni-muenchen.de>}

References


See Also

\texttt{compBoostCMA}, \texttt{dldaCMA}, \texttt{ElasticNetCMA}, \texttt{fdacMA}, \texttt{flexdaCMA}, \texttt{gbmCMA}, \texttt{knnCMA}, \texttt{ldaCMA}, \texttt{LassoCMA}, \texttt{nnetCMA}, \texttt{pknnCMA}, \texttt{plrcMA}, \texttt{pls_ldaCMA}, \texttt{pls_lrCMA}, \texttt{pls_rfCMA}, \texttt{qdacMA}, \texttt{rfCMA}, \texttt{scdaCMA}, \texttt{shrinkldaCMA}, \texttt{svmCMA}

Examples

### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 10 genes
golubX <- as.matrix(golub[,2:11])
### select learning set
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run PNN
pnnresult <- pnnCMA(X=golubX, y=golubY, learnind=learnind, sigma = 3)
### show results
show(pnnresult)
ftable(pnnresult)
plot(pnnresult)
Methods

- \( X = "\text{matrix}" , \ y = "\text{numeric}" , \ f = "\text{missing}" \) signature 1
- \( X = "\text{matrix}" , \ y = "\text{factor}" , \ f = "\text{missing}" \) signature 2
- \( X = "\text{data.frame}" , \ y = "\text{missing}" , \ f = "\text{formula}" \) signature 3
- \( X = "\text{ExpressionSet}" , \ y = "\text{character}" , \ f = "\text{missing}" \) signature 4

For references, further argument and output information, consult \texttt{pnnCMA}.

---

### prediction

**General method for predicting classes of new observations**

### Description

This method constructs the given classifier using the specified training data, gene selection and tuning results. Subsequently, class labels are predicted for new observations.

For S4 method information, see \texttt{classification-methods}.

### Usage

```r
prediction(X.tr,y.tr,X.new,f,classifier,genesel,models=F,nbgene,tuneres,...)
```

### Arguments

- **\( X.tr \)**: Training gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when \( f \) is not missing (s. below).
  - An object of class \texttt{ExpressionSet}.

- **\( X.new \)**: gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when \( f \) is not missing (s. below).
  - An object of class \texttt{ExpressionSet}.

- **\( y.tr \)**: Class labels of training observation. Can be one of the following:
  - A numeric vector.
  - A factor.
  - A character if \( X \) is an \texttt{ExpressionSet} that specifies the phenotype variable.
  - missing, if \( X \) is a data.frame and a proper formula \( f \) is provided.

**WARNING:** The class labels will be re-coded for classifier construction to range from 0 to \( K-1 \), where \( K \) is the total number of different classes in the learning set.

- **\( f \)**: A two-sided formula, if \( X \) is a data.frame. The left part correspond to class labels, the right to variables.

- **\( genesel \)**: Optional (but usually recommended) object of class \texttt{genesel} containing variable importance information for the argument learningsets. In this case the object contains a single variable selection. Appropriate \texttt{genesel}-objects can be obtained using the function \texttt{genesel} without learningset and setting \( X=X.tr \) and \( y=y.tr \) (i.e. corresponding to the training data of this function).
nbgene  Number of best genes to be kept for classification, based on either genesel or the call to \texttt{GeneSelection} using genesellist. In the case that both are missing, this argument is not necessary. \textbf{note:}
- If the gene selection method has been one of "lasso", "elasticnet", "boosting", \texttt{nbgene} will be reset to $\min(s, \texttt{nbgene})$ where $s$ is the number of nonzero coefficients.
- if the gene selection scheme has been "one-vs-all", "pairwise" for the multiclass case, there exist several rankings. The top \texttt{nbgene} will be kept of \textit{each} of them, so the number of effective used genes will sometimes be much larger.

\textbf{classifier}  Name of function ending with \texttt{CMA} indicating the classifier to be used.

\textbf{tuneres}  Analogous to the argument genesel - object of class \texttt{tuningresult} containing information about the best hyperparameter choice for the argument learningsets. Appropriate tuning-objects can be obtained using the function \texttt{tune} without learningsets and setting parameters $X=X.tr \ y=y.tr$ and genesel=genesel (i.e. using the same training data and gene selection as in this function)

\textbf{models}  a logical value indicating whether the model object shall be returned

...  Further arguments passed to the function \texttt{classifier}.

\textbf{Details}

This function builds the specified classifier and predicts the class labels of new observations. Hence, its usage differs from those of most other prediction functions in R.

\textbf{Value}

A object of class \texttt{predoutput-class}: Predicted classes can be seen by \texttt{show(predoutput)}

\textbf{Author(s)}

Christoph Bernau <bernau@ibe.med.uni-muenchen.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

\textbf{References}


\textbf{See Also}

\texttt{GeneSelection, tune, evaluation, compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, lassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rFCMA, pnnCMA, qdaCMA, rfCMA, scdCMA, shrinkldaCMA, svmCMAclassification}

\textbf{Examples}

### a simple k-nearest neighbour example
### datasets
## Not run: plot(x)
data(golub)
golubY <- golub[,1]
golubX <- as.matrix(golub[,1])
### Splitting data into training and test set

\[ X_{\text{tr}} \leftarrow \text{golubX}[1:30] \]
\[ X_{\text{new}} \leftarrow \text{golubX}[31:39] \]
\[ y_{\text{tr}} \leftarrow \text{golubY}[1:30] \]

1. **GeneSelection**
   
   \[ \text{selftest} \leftarrow \text{GeneSelection}(X = X_{\text{tr}}, y = y_{\text{tr}}, \text{method} = \text{"t.test"}) \]

2. **tuning**
   
   \[ \text{tunek} \leftarrow \text{tune}(X_{\text{tr}}, y_{\text{tr}}, \text{genesel} = \text{selftest}, \text{nbgene} = 20, \text{classifier} = \text{knnCMA}) \]

3. **classification**
   
   \[ \text{pred} \leftarrow \text{prediction}(X_{\text{tr}} = X_{\text{tr}}, y_{\text{tr}} = y_{\text{tr}}, X_{\text{new}} = X_{\text{new}}, \text{genesel} = \text{selftest}, \text{tuneres} = \text{tunek}, \text{nbgene} = 20, \text{classifier} = \text{knnCMA}) \]

### show and analyze results:

\[ \text{show(pred)} \]

## End(Not run)

---

### prediction-methods

**General method for predicting class lables of new observations**

**Description**

Perform prediction signatures:

**Methods**

- \( X_{\text{tr}} = \text{"matrix"}, X_{\text{new}} = \text{"matrix"}, y_{\text{tr}} = \text{any}, f = \text{"missing"} \) signature 1
- \( X_{\text{tr}} = \text{"data.frame"}, X_{\text{new}} = \text{"data.frame"}, y_{\text{tr}} = \text{"missing"}, f = \text{"formula"} \) signature 2
- \( X_{\text{tr}} = \text{"ExpressionSet"}, X_{\text{new}} = \text{"ExpressionSet"}, y_{\text{tr}} = \text{"character"}, f = \text{"missing"} \) signature 3

For further argument and output information, consult `classification`.

---

### predoutput-class

"predoutput"

**Description**

Object returned by the function `prediction`

**Slots**

- \( X_{\text{new}} \): Gene Expression matrix of new observations
- \( y_{\text{hat}} \): Predicted class labels for the new data.
- \( \text{model} \): List containing the constructed classifier.

**Methods**

- `show` Returns predicted class labels for the new data.
**qdaCMA**  

**Quadratic Discriminant Analysis**

### Description

Performs a quadratic discriminant analysis under the assumption of a multivariate normal distribution in each class without restriction concerning the covariance matrices. The function *qda* from the package MASS is called for computation.

For S4 method information, see *qdaCMA-methods*.

### Usage

```r
qdaCMA(X, y, f, learnind, models=FALSE, ...)
```

### Arguments

- **X** Gene expression data. Can be one of the following:  
  - A matrix. Rows correspond to observations, columns to variables.  
  - A data.frame, when `f` is not missing (s. below).  
  - An object of class ExpressionSet.

- **y** Class labels. Can be one of the following:  
  - A numeric vector.  
  - A factor.  
  - A character if `X` is an ExpressionSet that specifies the phenotype variable.  
  - missing, if `X` is a data.frame and a proper formula `f` is provided.

  **WARNING**: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

- **f** A two-sided formula, if `X` is a data.frame. The left part correspond to class labels, the right to variables.

- **learnind** An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

- **models** a logical value indicating whether the model object shall be returned

- **...** Further arguments to be passed to *qda* from the package MASS
Value

An object of class `cloutput`.

Note

Excessive variable selection has usually to performed before `qdaCMA` can be applied in the \( p > n \) setting. Not reducing the number of variables can result in an error message.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>

Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

Discriminant Analysis and Statistical Pattern Recognition.
Wiley, New York

See Also

`compBoostCMA`, `dldaCMA`, `ElasticNetCMA`, `fdaCMA`, `flexdaCMA`, `gbmCMA`, `knnCMA`, `LassoCMA`, `nnetCMA`, `pknnCMA`, `plrCMA`, `pls_ldaCMA`, `pls_lrCMA`, `pls_rfCMA`, `pnnCMA`, `rfCMA`, `scdaCMA`, `shrinkldaCMA`, `svmCMA`

Examples

```r
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression from first 3 genes
golubX <- as.matrix(golub[,2:4])
### select learningset
ratio <- 2/3
set.seed(112)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run QDA
qdaresult <- qdaCMA(X=golubX, y=golubY, learnind=learnind)
### show results
show(qdaresult)
ftable(qdaresult)
plot(qdaresult)
### multiclass example:
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression from first 4 genes
khanX <- as.matrix(khan[,2:5])
### select learningset
set.seed(111)
learnind <- sample(length(khanY), size=floor(ratio*length(khanY)))
### run QDA
```
qdaCMA-methods

qdaresult <- qdaCMA(X=khanX, y=khanY, learnind=learnind)

### show results
show(qdaresult)
ftable(qdaresult)
plot(qdaresult)

qdaCMA-methods  Quadratic Discriminant Analysis

Description

Performs a quadratic discriminant analysis under the assumption of a multivariate normal distribution in each classes without restriction concerning the covariance matrices. The function qda from the package MASS is called for computation.

Methods

X = "matrix", y = "numeric", f = "missing"  signature 1
X = "matrix", y = "factor", f = "missing"  signature 2
X = "data.frame", y = "missing", f = "formula"  signature 3
X = "ExpressionSet", y = "character", f = "missing"  signature 4
For further argument and output information, consult qdaCMA.

rfCMA  Classification based on Random Forests

Description

Random Forests were proposed by Breiman (2001) and are implemented in the package randomForest. In this package, they can as well be used to rank variables according to their importance, s. GeneSelection. For S4 method information, see rfCMA-methods

Usage

rfCMA(X, y, f, learnind, varimp = TRUE, seed = 111, models=FALSE,type=1,scale=FALSE,importance=TRUE)

Arguments

X  Gene expression data. Can be one of the following:
  • A matrix. Rows correspond to observations, columns to variables.
  • A data.frame, when f is not missing (s. below).
  • An object of class ExpressionSet.

y  Class labels. Can be one of the following:
  • A numeric vector.
  • A factor.
  • A character if X is an ExpressionSet that specifies the phenotype variable.
• missing, if \( X \) is a \texttt{data.frame} and a proper formula \( f \) is provided.

\textbf{WARNING}: The class labels will be re-coded to range from 0 to \( K-1 \), where \( K \) is the total number of different classes in the learning set.

\( f \) A two-sided formula, if \( X \) is a \texttt{data.frame}. The left part correspond to class labels, the right to variables.

\texttt{learnind} An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

\texttt{varimp} Should additional information for variable selection be provided? Defaults to \texttt{TRUE}.

\texttt{seed} Fix Random number generator seed to \texttt{seed}. This is useful to guarantee reproducibility of the results.

\texttt{models} a logical value indicating whether the model object shall be returned

\texttt{type} Parameter passed to function \texttt{importance}. Either 1 or 2, specifying the type of importance measure (1=mean decrease in accuracy, 2=mean decrease in node impurity).

\texttt{scale} Parameter passed to function \texttt{importance}. For permutation based measures, should the measures be divided by their standard errors?

\texttt{importance} Parameter passed to function \texttt{randomForest}. Should importance of predictors be assessed by permutation?

\texttt{...} Further arguments to be passed to \texttt{randomForest} from the package of the same name.

\textbf{Value}

If \texttt{varimp}, then an object of class \texttt{clvarseloutput} is returned, otherwise an object of class \texttt{cloutput}

\textbf{Author(s)}

Martin Slawski <ms@cs.uni-sb.de>

Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

\textbf{References}

Breiman, L. (2001)

Random Forest.

\textit{Machine Learning, 45}:5-32.

\textbf{See Also}

\texttt{compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, scdaCMA, shrinkldaCMA, svmCMA}
Examples

```r
### load Khan data
data(khan)
### extract class labels
khanY <- khan[,1]
### extract gene expression
khanX <- as.matrix(khan[,,-1])
### select learningset
set.seed(111)
learnind <- sample(length(khanY), size=floor(2/3*length(khanY)))
### run random Forest
#rfresult <- rfCMA(X=khanX, y=khanY, learnind=learnind, varimp = FALSE)
### show results
#show(rfresult)
#ftable(rfresult)
#plot(rfresult)
```

rfCMA-methods  Classification based on Random Forests

Description

Random Forests were proposed by Breiman (2001) and are implemented in the package randomForest. In this package, they can as well be used to rank variables according to their importance, s. GeneSelection.

Methods

- X = "matrix", y = "numeric", f = "missing" signature 1
- X = "matrix", y = "factor", f = "missing" signature 2
- X = "data.frame", y = "numeric", f = "formula" signature 3
- X = "ExpressionSet", y = "character", f = "missing" signature 4

For references, further argument and output information, consult rfCMA

roc  Receiver Operator Characteristic

Description

The empirical Receiver Operator Characteristic (ROC) is widely used for the evaluation of diagnostic tests, but also for the evaluation of classifiers. In this implementation, it can only be used for the binary classification case. The input are a numeric vector of class probabilities (which play the role of a test result) and the true class labels. Note that misclassification performance can (partly widely) differ from the Area under the ROC (AUC). This is due to the fact that misclassification rates are always computed for the threshold 'probability = 0.5'.

Arguments

- object  An object of cloutput.
- plot Should the ROC curve be plotted ? Default is TRUE.
- ... Argument to specify further graphical options.
Value

The empirical area under the curve (AUC).

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References


See Also

evaluation

scdaCMA | Shrunken Centroids Discriminant Analysis

Description

The nearest shrunken centroid classification algorithm is detailly described in Tibshirani et al. (2002).

It is widely known under the name PAM (prediction analysis for microarrays), which can also be found in the package *pamr*.

For S4 method information, see *scdaCMA-methods*.

Usage

scdaCMA(X, y, f, learnind, delta = 0.5, models=FALSE,...)

Arguments

X Gene expression data. Can be one of the following:
- A matrix. Rows correspond to observations, columns to variables.
- A data.frame, when f is not missing (s. below).
- An object of class ExpressionSet.

y Class labels. Can be one of the following:
- A numeric vector.
- A factor.
- A character if X is an ExpressionSet that specifies the phenotype variable.
- missing, if X is a data.frame and a proper formula f is provided.

**WARNING**: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.
An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

The shrinkage intensity for the class centroids - a hyperparameter that must be tuned. The default 0.5 not necessarily produces good results.

a logical value indicating whether the model object shall be returned

Currently unused argument.

An object of class cloutput.

The results can differ from those obtained by using the package pammr.

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>


### load Khan data
```r
data(khan)
```n
### extract class labels
```r
khanY <- khan[,1]
```n
### extract gene expression
```r
khanX <- as.matrix(khan[,,-1])
```n
### select learningset
```r
set.seed(111)
learnind <- sample(length(khanY), size=floor(2/3*length(khanY)))
```n
### run Shrunken Centroids classifier, without tuning
```r
scdaresult <- scdaCMA(X=khanX, y=khanY, learnind=learnind)
```n
### show results
```r
show(scdaresult)
ftable(scdaresult)
plot(scdaresult)
```
scdaCMA-methods  Shrunken Centroids Discriminant Analysis

Description
The nearest shrunken centroid classification algorithm is detailedly described in Tibshirani et al. (2002).
It is widely known under the name PAM (prediction analysis for microarrays), which can also be found in the package pamr.

Methods
X = "matrix", y = "numeric", f = "missing"  signature 1
X = "matrix", y = "factor", f = "missing"  signature 2
X = "data.frame", y = "missing", f = "formula"  signature 3
X = "ExpressionSet", y = "character", f = "missing"  signature 4
For references, further argument and output information, consult scdaCMA.

shrinkldaCMA  Shrinkage linear discriminant analysis

Description
Linear Discriminant Analysis combined with the James-Stein-Shrinkage approach of Schaefer and Strimmer (2005) for the covariance matrix.
Currently still an experimental version.
For S4 method information, see shrinkldaCMA-methods

Usage
shrinkldaCMA(X, y, f, learnind, models=FALSE, ...)

Arguments
X  Gene expression data. Can be one of the following:
   • A matrix. Rows correspond to observations, columns to variables.
   • A data.frame, when f is not missing (s. below).
   • An object of class ExpressionSet.
y  Class labels. Can be one of the following:
   • A numeric vector.
   • A factor.
   • A character if X is an ExpressionSet that specifies the phenotype variable.
   • missing, if X is a data.frame and a proper formula f is provided.
WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learnind An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

models a logical value indicating whether the model object shall be returned

Value An object of class cloutput.

Note This is still an experimental version.

Covariance shrinkage is performed by calling functions from the package corpcor. Variable selection is not necessary.

Author(s)
Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References

See Also
compBoostCMA, dlldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrCMA, pls_rfCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, svmCMA.

Examples
### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[,,-1])
### select learningset
ratio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(ratio*length(golubY)))
### run shrinkage-LDA
result <- shrinkldaCMA(X=golubX, y=golubY, learnind=learnind)
### shrinkldaCMA-methods  
**Shrinkage linear discriminant analysis**

**Description**
Linear Discriminant Analysis combined with the James-Stein-Shrinkage approach of Schaefer and Strimmer (2005) for the covariance matrix.
Currently still an experimental version. For S4 method information, see shrnkldaCMA-methods.

**Methods**

- \( \text{X} = \text{"matrix"}, \text{y} = \text{"numeric"}, \text{f} = \text{"missing"} \) signature 1
- \( \text{X} = \text{"matrix"}, \text{y} = \text{"factor"}, \text{f} = \text{"missing"} \) signature 2
- \( \text{X} = \text{"data.frame"}, \text{y} = \text{"missing"}, \text{f} = \text{"formula"} \) signature 3
- \( \text{X} = \text{"ExpressionSet"}, \text{y} = \text{"character"}, \text{f} = \text{"missing"} \) signature 4

For further argument and output information, consult shrnkldaCMA.

---

### summary
**Summarize classifier evaluation**

**Description**
This method principally does nothing more than applying the pre-implemented `summary()` function to the slot `score` of an object of class `evaloutput`. One then obtains the usual five-point-summary, consisting of minimum and maximum, lower and upper quartile and the median. Additionally, the mean is also shown.

**Arguments**

- `object`  
  An object of class `evaloutput`.

- `...`  
  Further arguments passed to the pre-implemented `summary` function.

**Value**
No return.

**Note**
That the results normally differ for different evaluation schemes ("iterationwise" or "observation-wise").
Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

evaluation, compare, obsinfo.

Description

Calls the function svm from the package e1071 that provides an interface to the award-winning LIBSVM routines. For S4 method information, see svmCMA-methods

Usage

svmCMA(X, y, f, learnind, probability, models=FALSE, seed=341, ...)

Arguments

X 
Gene expression data. Can be one of the following:
- A matrix. Rows correspond to observations, columns to variables.
- A data.frame, when f is not missing (s. below).
- An object of class ExpressionSet.

y 
Class labels. Can be one of the following:
- A numeric vector.
- A factor.
- A character if X is an ExpressionSet that specifies the phenotype variable.
- missing, if X is a data.frame and a proper formula f is provided.

WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

f 
A two-sided formula, if X is a data.frame. The left part correspond to class labels, the right to variables.

learnind 
An index vector specifying the observations that belong to the learning set. May be missing; in that case, the learning set consists of all observations and predictions are made on the learning set.

probability 
logical indicating whether the model should allow for probability predictions.

seed 
Fix random number generator for reproducibility.

models 
a logical value indicating whether the model object shall be returned

... 
Further arguments to be passed to svm from the package e1071

Value

An object of class cloutput.
Note

Contrary to the default settings in e1071:::svm, the used kernel is a linear kernel which has turned to be out a better default setting in the small sample, large number of predictors - situation, because additional nonlinearity is mostly not necessary there. It additionally avoids the tuning of a further kernel parameter gamma, s. help of the package e1071 for details. Nevertheless, hyperparameter tuning concerning the parameter cost must usually be performed to obtain reasonable results, s. tune.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
Christoph Bernau <bernaue@ibe.med.uni-muenchen.de>

References

A training algorithm for optimal margin classifiers.
Proceedings of the fifth annual workshop on Computational learning theory, pages 144-152, ACM Press.


Learning with kernels. MIT Press, Cambridge, MA.

See Also

compBoostCMA, dldaCMA, ElasticNetCMA, fdaCMA, flexdaCMA, gbmCMA, knnCMA, ldaCMA, LassoCMA, nnetCMA, pknnCMA, plrCMA, pls_ldaCMA, pls_lrcMA, pls_rFCMA, pnnCMA, qdaCMA, rfCMA, scdaCMA, shrinkldaCMA

Examples

### load Golub AML/ALL data
data(golub)
### extract class labels
golubY <- golub[,1]
### extract gene expression
golubX <- as.matrix(golub[-1])
### select learningset
tatio <- 2/3
set.seed(111)
learnind <- sample(length(golubY), size=floor(tatio*length(golubY)))
### run _untuned_linear SVM
svmresult <- svmCMA(X=golubX, y=golubY, learnind=learnind, probability=TRUE)
### show results
show(svmresult)
ftable(svmresult)
plot(svmresult)
svmCMA-methods

Description

Calls the function svm from the package e1071 that provides an interface to the award-winning LIBSVM routines.

Methods

X = "matrix", y = "numeric", f = "missing" signature 1
X = "matrix", y = "factor", f = "missing" signature 2
X = "data.frame", y = "missing", f = "formula" signature 3
X = "ExpressionSet", y = "character", f = "missing" signature 4

For further argument and output information, consult svmCMA.

toplist

Description

This is a convenient method to get quick access to the most important variables, based on the result of call to GeneSelection.

Usage

toplist(object, k = 10, iter = 1, show = TRUE, ...)

Arguments

object An object of genesel.

k Number of top genes for which information should be displayed. Defaults to 10.

iter iteration number (learningset) for which tuning results should be displayed.

show Should the results be printed ? Default is TRUE.

... Currently unused argument.

Value

The type of output depends on the gene selection scheme. For the multiclass case, if gene selection has been run with the "pairwise" or "one-vs-all" scheme, then the output will be a list of data.frames, each containing the gene indices plus variable importance for the top k genes. The list elements are named according to the binary scenarios (e.g., 1 vs. 3). Otherwise, a single data.frame is returned.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>

Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
References

See Also
genesel, GeneSelection.plot, genesel-method

---

**tune**

Hyperparameter tuning for classifiers

---

**Description**

Most classifiers implemented in this package depend on one or even several hyperparameters (s. details) that should be optimized to obtain good (and comparable !) results. As tuning scheme, we propose three fold Cross-Validation on each learningset (for fixed selected variables). Note that learningsets usually do not contain the complete dataset, so tuning involves a second level of splitting the dataset. Increasing the number of folds leads to larger datasets (and possibly to higher accuracy), but also to higher computing times.

For S4 method information, s. link{tune-methods}

**Usage**

tune(X, y, f, learningsets, genesel, genesellist = list(), nbgene, classifier, fold = 3, strat = FALSE, ...)

**Arguments**

- **X** Gene expression data. Can be one of the following:
  - A matrix. Rows correspond to observations, columns to variables.
  - A data.frame, when *f* is *not* missing (s. below).
  - An object of class ExpressionSet.

- **y** Class labels. Can be one of the following:
  - A numeric vector.
  - A factor.
  - A character if *X* is an ExpressionSet that specifies the phenotype variable.
  - missing, if *X* is a data.frame and a proper formula *f* is provided.

- **f** A two-sided formula, if *X* is a data.frame. The left part correspond to class labels, the right to variables.

- **learningsets** An object of class learningsets. May be missing, then the complete datasets is used as learning set.

- **genesel** Optional (but usually recommended) object of class genesel containing variable importance information for the argument learningsets

- **genesellist** In the case that the argument genesel is missing, this is an argument list passed to GeneSelection. If both genesel and genesellist are missing, no variable selection is performed.
**tune**

**nbgene** Number of best genes to be kept for classification, based on either genesel or the call to `GeneSelection` using genesellist. In the case that both are missing, this argument is not necessary. **note:**

- If the gene selection method has been one of "lasso", "elasticnet", "boosting", `nbgene` will be reset to `min(s, nbgene)` where `s` is the number of nonzero coefficients.
- If the gene selection scheme has been "one-vs-all", "pairwise" for the multiclass case, there exist several rankings. The top `nbgene` will be kept of each of them, so the number of effective used genes will sometimes be much larger.

**classifier** Name of function ending with `CMA` indicating the classifier to be used.

**fold** The number of cross-validation folds used within each `learningset`. Default is 3. Increasing `fold` will lead to higher computing times.

**strat** Should stratified cross-validation according to the class proportions in the complete dataset be used? Default is `FALSE`.

**grids** A named list. The names correspond to the arguments to be tuned, e.g. `k` (the number of nearest neighbours) for `knnCMA`, or `cost` for `svmCMA`. Each element is a numeric vector defining the grid of candidate values. Of course, several hyperparameters can be tuned simultaneously (though requiring much time). By default, `grids` is an empty list. In that case, a pre-defined list will be used, s. details.

**trace** Should progress be traced? Default is `TRUE`.

... Further arguments to be passed to `classifier`, of course not one of the arguments to be tuned (!).

**Details**

The following default settings are used, if the arguments `grids` is an empty list:

- `gbmCMA n.trees = c(50, 100, 200, 500, 1000)`
- `compBoostCMA mstop = c(50, 100, 200, 500, 1000)`
- `LassoCMA norm.fraction = seq(from=0.1, to=0.9, length=9)`
- `ElasticNetCMA norm.fraction = seq(from=0.1, to=0.9, length=5), alpha = 2^{-(5:1)}`
- `plrCMA lambda = 2^{(-4:4)}`
- `pls_ldaCMA comp = 1:10`
- `pls_lrCMA comp = 1:10`
- `pls_rfCMA comp = 1:10`
- `rfCMA mtry = ceiling(c(0.1, 0.25, 0.5, 1, 2)*sqrt(ncol(X))), nsize = c(1, 2, 3)`
- `knnCMA k=1:10`
- `pknnCMA k = 1:10`
- `scdaCMA delta = c(0.1, 0.25, 0.5, 1, 2, 5)`
- `pnnCMA sigma = c(2^{(-2:2)}), size = 1:5, decay = c(0, 0.2^{(-4:1)})`
- `svmCMA kernel = "linear" cost = c(0.1, 1, 5, 10, 50, 100, 500)`
- `svmCMA kernel = "radial" cost = c(0.1, 1, 5, 10, 50, 100, 500), gamma = 2^{(-2:2)}`
- `svmCMA kernel = "polynomial" cost = c(0.1, 1, 5, 10, 50, 100, 500), degree = 2:4`
Value

An object of class `tuningresult`

Note

The computation time can be enormously high. Note that for each different learningset, the classifier must be trained `fold` times number of possible different hyperparameter combinations times. E.g. if the number of the learningsets is fifty, `fold = 3` and two hyperparameters (each with 5 candidate values) are tuned, `50x3x25=3750` training iterations are necessary!

Author(s)

Martin Slawski `<ms@cs.uni-sb.de>`
Anne-Laure Boulesteix `<boulesteix@ibe.med.uni-muenchen.de>`
Christoph Bernau `<bernau@ibe.med.uni-muenchen.de>`

References


See Also

tuningresult, GeneSelection, classification

Examples

```r
## Not run:
### simple example for a one-dimensional grid, using compBoostCMA.
### dataset
data(golub)
golubY <- golub[,1]
golubX <- as.matrix(golub[-1])
### learningsets
set.seed(111)
lset <- GenerateLearningsets(y=golubY, method = "CV", fold=5, strat =TRUE)
### tuning after gene selection with the t.test
tuneres <- tune(X = golubX, y = golubY, learningsets = lset,
                 genesellist = list(method = "t.test"),
                 classifier=compBoostCMA, nbgene = 100,
                 grids = list(mstop = c(50, 100, 250, 500, 1000)))
### inspect results
show(tuneres)
best(tuneres)
plot(tuneres, iter = 3)
## End(Not run)
```
Hyperparameter tuning for classifiers

Description
Performs hyperparameter tuning for the following signatures:

Methods
- X = "matrix", y = "numeric", f = "missing"  signature 1
- X = "matrix", y = "factor", f = "missing"  signature 2
- X = "data.frame", y = "missing", f = "formula"  signature 3
- X = "ExpressionSet", y = "character", f = "missing"  signature 4

For further argument and output information, consult `tune`.

tuningresult-class  "tuningresult"

Description
Object returned by the function `tune`

Slots
- hypergrid: A data.frame representing the grid of values that were tried and evaluated. The number of columns equals the number of tuned hyperparameters and the number rows equals the number of all possible combinations of the discrete grids.
- tuneres: A list whose lengths equals the number of different learning sets for which tuning has been performed and whose elements are numeric vectors with length equal to the number of rows of hypergrid (as above), containing the misclassification rate belonging to the respective hyperparameter/hyperparameter combination. In order to get an overview about the best hyperparameter/hyperparameter combination, use the convenience method `best`.
- method: Name of the classifier that has been tuned.
- fold: Number of cross-validation fold used for tuning, s. argument of the same name in `tune`

Methods
- show Use `show(tuningresult-object)` for brief information.
- best Use `best(tuningresult-object)` to see which hyperparameter/hyperparameter combination has performed best in terms of the misclassification rate, s. `best,tuningresult-method`
- plot Use `plot(tuningresult-object, iter, which)` to display the performance of hyperparameter/hyperparameter combinations graphically, either as one-dimensional or as two-dimensional (contour) plot, s. `plot,tuningresult-method`

Author(s)
- Martin Slawski <ms@cs.uni-sb.de>
- Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>
See Also

`tune`

---

varseloutput-class "varseloutput"

Description

An object returned by the functions described in `filter`, usually not created directly by the user.

Slots

varsel: numeric vector of variable importance measures, e.g. absolute of genewise statistics.

Methods

No methods are currently defined.

Author(s)

Martin Slawski <ms@cs.uni-sb.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

`filter, clvarseloutput`

---

weighted.mcr Tuning / Selection bias correction

Description

Performs subsampling for several classifiers or a single classifiers with different tuning parameter values or numbers of selected genes. Eventually, a specific procedure for correcting for the tuning or selection bias, which is caused by optimal selection of classifiers or tuning parameters, is applied.

Usage

`weighted.mcr(classifiers, parameters, nbgenes, sel.method, X, y, portion, niter=100, shrinkage=F)`
Arguments

classifiers A character vector of the several CMA classifiers that shall be used. If the same classifier shall be used with different tuning parameters it must appear several times in this vector.

parameters A character containing the tuning parameter values corresponding to the classification methods in classifiers. Must have the same length as classifiers.

nbgenes A numeric vector indicating how many variables shall be selected by sel.method for the corresponding classifier. Must have the same length as classifiers.

sel.method The CMA-method (represented as a string) that shall be applied for variable selection. If this parameter is set to 'none' no variable selection is performed.

X The matrix of gene expression data. Can be one of the following. Rows correspond to observations, columns to variables.

y Class labels. Can be one of the following:
  • A numeric vector.
  • A factor.

WARNING: The class labels will be re-coded to range from 0 to K-1, where K is the total number of different classes in the learning set.

portion A numeric value which indicates the portion of observations that will be used for training the classifiers.

niter The number of subsampling iterations.

shrinkage A logical value indicating whether shrinkage (WMCS) shall be applied.

Details

The algorithm tries to avoid the additional computational costs of a nested cross validation by estimating the corrected misclassification rate of the best classifier by a weighted mean of all classifiers included in the subsampling approach.

Value

An object of class wmcr.result which provides the corrected and uncorrected misclassification rate of the best classifier as well as weights and misclassification rates for all classifiers used in the subsampling approach.

Author(s)

Christoph Bernau <bernau@ibe.med.uni-muenchen.de>

Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References


See Also

wmc, classification, GeneSelection, tune, evaluation,
Examples

```r
#inputs
classifiers<-rep('knnCMA',7)
bgenes<-rep(50,7)
parameters<-c('k=1','k=3','k=5','k=7','k=9','k=11','k=13')
portion<-0.8
niter<-100
data(golub)
x<-as.matrix(golub[, -1])
y<-golub[,1]
set.seed(1)

#function call
wmcr<-weighted.mcr(classifiers=classifiers,parameters=parameters,nbgenes=nbgens,sel.method=sel.method,X=X,y=y,portion=portion,niter=niter)
```

weighted.mcr-methods

General method for tuning / selection bias correction

Description

Perform tuning / selection bias correction in subsampling for the following signatures:

Methods

- classifiers="character",parameters="character",nbgenes="numeric",sel.method="character",X = "matrix", y = "numeric"
- classifiers="character",parameters="character",nbgenes="numeric",sel.method="character",X = "matrix", y = "factor"
- classifiers="character",parameters="character",nbgenes="missing",sel.method="character",X = "matrix", y = "factor"

For further argument and output information, consult weighted.mcr.

wmc

Tuning / Selection bias correction based on matrix of subsampling fold errors

Description

Perform tuning / selection bias correction for a matrix of subsampling fold errors.

Usage

`wmc(mcr.m, n.tr, n.ts, shrinkage=F)`

Arguments

- **mcr.m**: A matrix of resampling fold errors. Columns correspond the the fold errors of a single classifier.
- **n.tr**: Number of observations in the resampling training sets.
- **n.ts**: Number of observations in the resampling test sets.
- **shrinkage**: A logical value indicating whether shrinkage (WMCS) shall be applied.
Details

The algorithm tries to avoid the additional computational costs of a nested cross validation by estimating the corrected misclassification rate of the best classifier by a weighted mean of all classifiers included in the subsampling approach.

Value

A list containing the corrected misclassification rate, the index of the best method and a logical value indicating whether shrinkage has been applied.

Author(s)

Christoph Bernau <berna@ibe.med.uni-muenchen.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

References


See Also

weighted.mcr.classification, GeneSelection, tune, evaluation,
wmcr.result-class

Description

Object returned by function weighted.mcr.

Slots

- corrected.mcr: The corrected misclassification rate for the best method.
- best.method: The method which performed best in the subsampling approach.
- mcrs: Misclassification rates of all classifiers used in the subsampling approach.
- weights: The weights used for the different classifiers in the correction method.
- cov: Estimated covariance matrix for the misclassification rates of the different classifiers.
- uncorrected.mcr: The uncorrected misclassification rate of the best method.
- ranges: Minimum and maximal mean misclassification rates as well as the theoretical bound for nested cross validation (averaging over foldwise minima or maxima respectively).
- mcr.m: matrix of resampling fold errors, columns correspond to the fold errors of a single classifier.
- shrinkage: a logical value indicating whether shrinkage (WMCS) has been applied.

Methods

show Use show(wmcr.result-object) for brief information

Author(s)

Christoph Bernau <bernau@ibe.med.uni-muenchen.de>
Anne-Laure Boulesteix <boulesteix@ibe.med.uni-muenchen.de>

See Also

weighted.mcr
Index

*Topic **datasets**
golub, 38
khan, 40

*Topic **multivariate**
Barplot, 4
best, 5
boxplot, 6
classification, 7
classification-methods, 9
cloutput-class, 9
classvarseloutput-class, 10
CMA-package, 3
compare, 11
compare-methods, 13
compBoostCMA, 14
compBoostCMA-methods, 16
dldaCMA, 16
dldaCMA-methods, 18
ElasticNetCMA, 19
ElasticNetCMA-methods, 21
evaloutput-class, 21
evaluation, 22
evaluation-methods, 24
fdaCMA, 24
fdaCMA-methods, 26
filter, 27
flexdaCMA, 27
flexdaCMA-methods, 29
ftable, 30
gbmCMA, 30
gbmCMA-methods, 32
GenerateLearningsets, 33
genesel-class, 34
GeneSelection, 35
GeneSelection-methods, 38
internals, 39
join, 39
join-methods, 40
knnCMA, 41
knnCMA-methods, 43
lassoCMA, 43
LassoCMA-methods, 45
ldaCMA, 45
ldaCMA-methods, 47
learningset-class, 47
nnetCMA, 48
nnetCMA-methods, 50
obsinfo, 51
pknnCMA, 52
pknnCMA-methods, 53
Planarplot, 54
Planarplot-methods, 55
plot, 56
plot tuningresult, 57
p1rCMA, 58
p1rCMA-methods, 59
pls ldaCMA, 60
pls ldaCMA-methods, 61
pls lrCMA, 62
pls_lrCMA-methods, 63
pls rfCMA, 64
pls_rfCMA-methods, 65
pnnCMA, 66
pnnCMA-methods, 67
prediction, 68
prediction-methods, 70
predoutput-class, 70
qdaCMA, 71
qdaCMA-methods, 73
rfCMA, 73
rfCMA-methods, 75
roc, 75
scdaCMA, 76
scdaCMA-methods, 78
shrinkldaCMA, 78
shrinkldaCMA-methods, 80
summary, 80
svmCMA, 81
svmCMA-methods, 83
toplist, 83
tune, 84
tune-methods, 87
tuningresult-class, 87
varseloutput-class, 88

*Topic **tuning bias, selection bias, corrected misclassification**
<table>
<thead>
<tr>
<th>Package</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>compBoostCMA</td>
<td>matrix, factor, missing-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>matrix, numeric, missing-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>methods, 16</td>
</tr>
<tr>
<td>dldaCMA</td>
<td>4, 8, 10, 15, 16, 18, 20, 25, 29, 32, 42, 44, 46, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82</td>
</tr>
<tr>
<td>fdaCMA</td>
<td>methods, 26</td>
</tr>
<tr>
<td>ElasticNetCMA</td>
<td>4, 8, 10, 11, 15, 17, 19, 21, 25, 29, 32, 42, 44, 46, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82</td>
</tr>
<tr>
<td>fdaCMA</td>
<td>methods, 26</td>
</tr>
<tr>
<td>evaluation</td>
<td>methods, 24</td>
</tr>
<tr>
<td>evaloutput</td>
<td>methods, 24</td>
</tr>
<tr>
<td>evaout</td>
<td>evaluation-class, 21</td>
</tr>
<tr>
<td>mcrc</td>
<td>result-class, 92</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>data.frame, missing, formula-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>ExpressionSet, character, missing-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>matrix, factor, missing-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>matrix, numeric, missing-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>methods, 16</td>
</tr>
<tr>
<td>care.dev</td>
<td>(internals), 39</td>
</tr>
<tr>
<td>care.exp</td>
<td>(internals), 39</td>
</tr>
<tr>
<td>character.plot</td>
<td>(internals), 39</td>
</tr>
<tr>
<td>classification</td>
<td>4, 7, 9, 12, 13, 23, 34, 38–40, 47, 48, 49, 56, 69, 70, 72, 74, 77, 79, 80, 84, 85</td>
</tr>
<tr>
<td>classification</td>
<td>data.frame, missing, formula-method</td>
</tr>
<tr>
<td></td>
<td>(classification-methods), 9</td>
</tr>
<tr>
<td>classification</td>
<td>ExpressionSet, character, missing-method</td>
</tr>
<tr>
<td></td>
<td>(classification-methods), 9</td>
</tr>
<tr>
<td>classification</td>
<td>matrix, factor, missing-method</td>
</tr>
<tr>
<td></td>
<td>(classification-methods), 9</td>
</tr>
<tr>
<td>classification</td>
<td>matrix, numeric, missing-method</td>
</tr>
<tr>
<td></td>
<td>(classification-methods), 9</td>
</tr>
<tr>
<td>classification</td>
<td>methods, 9</td>
</tr>
<tr>
<td>cloutput</td>
<td>8, 11, 12, 17, 22, 25, 28, 30, 31, 39–41, 46, 49, 53, 56, 58, 61, 63, 65, 67, 72, 74, 75, 77, 79, 81</td>
</tr>
<tr>
<td>cloutput</td>
<td>(cloutput-class), 9</td>
</tr>
<tr>
<td>cloutput-class</td>
<td>9</td>
</tr>
<tr>
<td>clvarseloutput</td>
<td>8, 12, 15, 20, 21, 29, 40, 74, 88</td>
</tr>
<tr>
<td>clvarseloutput</td>
<td>(clvarseloutput-class), 10</td>
</tr>
<tr>
<td>clvarseloutput</td>
<td>class, 10</td>
</tr>
<tr>
<td>CMA</td>
<td>(CMA-package), 3</td>
</tr>
<tr>
<td>compare</td>
<td>4, 11, 14, 23, 40, 81</td>
</tr>
<tr>
<td>compare</td>
<td>list-method (compare-methods), 13</td>
</tr>
<tr>
<td>compare</td>
<td>methods, 13</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>4, 8, 10, 11, 14, 16, 17, 19, 20, 25, 29, 32, 42, 44, 46, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>data.frame, missing, formula-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
<tr>
<td>compBoostCMA</td>
<td>ExpressionSet, character, missing-method</td>
</tr>
<tr>
<td></td>
<td>(compBoostCMA-methods), 16</td>
</tr>
</tbody>
</table>
INDEX

filter, 27, 38, 88
flexdaCMA, 4, 8, 10, 15, 17, 20, 25, 27, 29, 32, 42, 44, 46, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82
flexdaCMA, data.frame, missing, formula-method (flexdaCMA-methods), 29
flexdaCMA, ExpressionSet, character, missing-method (flexdaCMA-methods), 29
flexdaCMA, matrix, factor, missing-method (flexdaCMA-methods), 29
flexdaCMA, matrix, numeric, missing-method (flexdaCMA-methods), 29
flexdaCMA-methods, 29
ftable, 30
ftable, cloutput-method (ftable), 30
ftest (filter), 27
gbmCMA, 4, 8, 10, 15, 17, 20, 25, 29, 30, 32, 42, 44, 46, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
gbmCMA, data.frame, missing, formula-method (gbmCMA-methods), 32
gbmCMA, ExpressionSet, character, missing-method (gbmCMA-methods), 32
gbmCMA, matrix, factor, missing-method (gbmCMA-methods), 32
gbmCMA, matrix, numeric, missing-method (gbmCMA-methods), 32
gbmCMA-methods, 32
GenerateLearningsets, 4, 33, 38, 47, 48
genesel, 4, 5, 7, 37, 68, 83, 84
genesel (genesel-class), 34
genesel-class, 34
GeneSelection, 4, 5, 7, 8, 14, 16, 19, 27, 34, 35, 35, 38, 43, 45, 47, 48, 54, 55, 69, 83–86, 89, 91
GeneSelection, data.frame, missing, formula-method (GeneSelection-methods), 38
GeneSelection, ExpressionSet, character, missing-method (GeneSelection-methods), 38
GeneSelection, matrix, factor, missing-method (GeneSelection-methods), 38
GeneSelection, matrix, numeric, missing-method (GeneSelection-methods), 38
GeneSelection-methods, 38
golub, 36, 38
golubcrit (filter), 27
internals, 39
join, 8, 39, 40, 56
join, list-method (join-methods), 40
join-methods, 40
khan, 40
knnCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 41, 43, 44, 46, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
knnCMA, data.frame, missing, formula-method (knnCMA-methods), 43
knnCMA, ExpressionSet, character, missing-method (knnCMA-methods), 43
knnCMA, matrix, factor, missing-method (knnCMA-methods), 43
knnCMA, matrix, numeric, missing-method (knnCMA-methods), 43
knnCMA-methods, 41, 43
kruskal.test (filter), 27
LassoCMA, 4, 8, 10, 11, 15, 17, 20, 25, 29, 32, 36, 42, 43, 45, 46, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
LassoCMA, data.frame, missing, formula-method (LassoCMA-methods), 45
LassoCMA, ExpressionSet, character, missing-method (LassoCMA-methods), 45
LassoCMA, matrix, factor, missing-method (LassoCMA-methods), 45
LassoCMA, matrix, numeric, missing-method (LassoCMA-methods), 45
LassoCMA-methods, 45
ldaCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 45, 47, 49, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 84
ldaCMA, data.frame, missing, formula-method (ldaCMA-methods), 47
ldaCMA, ExpressionSet, character, missing-method (ldaCMA-methods), 47
ldaCMA, matrix, factor, missing-method (ldaCMA-methods), 47
ldaCMA, matrix, numeric, missing-method (ldaCMA-methods), 47
ldaCMA-methods, 47
learningsets, 7, 33, 34, 36, 56, 84
learningsets (learningsets-class), 47
learningsets-class, 47
limmatest (filter), 27
mklt (internals), 39
mkreg (internals), 39
my.care.exp (internals), 39
nnetCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 46, 49, 50, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
nnetCMA, data.frame, missing, formula-method  
(nnetCMA-methods), 50
nnetCMA, ExpressionSet, character, missing-method  
(nnetCMA-methods), 50
nnetCMA, matrix, factor, missing-method  
(nnetCMA-methods), 50
nnetCMA, matrix, numeric, missing-method  
(nnetCMA-methods), 50
nnetCMA-methods, 48, 50

obsinfo, 21, 51, 81
obsinfo, evaloutput-method  
(evaloutput-class), 21

pknnCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 46, 49, 52, 53, 55, 59, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
pknnCMA, data.frame, missing, formula-method  
(pknnCMA-methods), 53
pknnCMA, ExpressionSet, character, missing-method  
(pknnCMA-methods), 53
pknnCMA, matrix, factor, missing-method  
(pknnCMA-methods), 53
pknnCMA, matrix, numeric, missing-method  
(pknnCMA-methods), 53
pknnCMA-methods, 52, 53
Planarplot, 54, 55
Planarplot, data.frame, missing, formula-method  
(Planarplot-methods), 55
Planarplot, ExpressionSet, character, missing-method  
(Planarplot-methods), 55
Planarplot, matrix, factor, missing-method  
(Planarplot-methods), 55
Planarplot, matrix, numeric, missing-method  
(Planarplot-methods), 55
Planarplot-methods, 55
plot, 56
plot, tuningresult, 57
plot, cloutput, missing-method(plot), 56
plot, cloutput-method(plot), 56
plot, genesel, missing-method(Barpplot), 4
plot, genesel-method(Barpplot), 4
plot, tuningresult, missing-method(plot tuningresult), 57
plot, tuningresult-method(plot tuningresult), 57
plotprob (internals), 39
plrCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 46, 49, 53, 55, 58, 60, 61, 63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
plrCMA, data.frame, missing, formula-method  
(plrCMA-methods), 59
plrCMA, ExpressionSet, character, missing-method  
(plrCMA-methods), 59
plrCMA, matrix, factor, missing-method  
(plrCMA-methods), 59
plrCMA, matrix, numeric, missing-method  
(plrCMA-methods), 59
plrCMA-methods, 58, 59
pls ldaCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 46, 49, 53, 55, 59, 60, 61–63, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
pls ldaCMA, data.frame, missing, formula-method  
(pls ldaCMA-methods), 61
pls ldaCMA, ExpressionSet, character, missing-method  
(pls ldaCMA-methods), 61
pls ldaCMA, matrix, factor, missing-method  
(pls ldaCMA-methods), 61
pls ldaCMA, matrix, numeric, missing-method  
(pls ldaCMA-methods), 61
pls ldaCMA-methods, 61
pls lrCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 46, 49, 53, 55, 59, 61, 62, 64, 65, 67, 69, 71, 72, 74, 77, 79, 82, 85
pls lrCMA, data.frame, missing, formula-method  
(pls lrCMA-methods), 63
pls lrCMA, ExpressionSet, character, missing-method  
(pls lrCMA-methods), 63
pls lrCMA, matrix, factor, missing-method  
(pls lrCMA-methods), 63
pls lrCMA, matrix, numeric, missing-method  
(pls lrCMA-methods), 63
pls lrCMA-methods, 63
pls rfCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 46, 49, 53, 55, 59, 61, 63, 64, 66, 67, 69, 71, 72, 74, 77, 79, 82, 85
pls rfCMA, data.frame, missing, formula-method  
(pls rfCMA-methods), 65
pls rfCMA, ExpressionSet, character, missing-method  
(pls rfCMA-methods), 65
pls rfCMA, matrix, factor, missing-method  
(pls rfCMA-methods), 65
pls rfCMA, matrix, numeric, missing-method  
(pls rfCMA-methods), 65
pls rfCMA-methods, 65
pnnCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44, 46, 49, 53, 55, 59, 61, 63, 65, 66, 68, 69, 71, 72, 74, 77, 79, 82, 85
pnnCMA, data.frame, missing, formula-method  
(pnnCMA-methods), 67
pnnCMA, ExpressionSet, character, missing-method  
(pnnCMA-methods), 67
pnnCMA, matrix, factor, missing-method  
(pnnCMA-methods), 67
pnnCMA-methods, 67

INDEX
INDEX

index

97

prediction, 68

prediction, data.frame, missing, data.frame, formula-method
(prediction-methods), 70

prediction, expressionSet, character, expressionSet, missing-method
(prediction-methods), 70

prediction, matrix, ANY, matrix, missing-method
(prediction-methods), 70

prediction-methods, 70

predoutput (predoutput-class), 70

predoutput-class, 70

qdaCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44,
46, 49, 53, 55, 59, 61, 63, 65, 67, 69,
71, 72, 73, 74, 77, 79, 82

qdaCMA, data.frame, missing, formula-method
(qdaCMA-methods), 73

qdaCMA, expressionSet, character, missing-methods
(shrinkldaCMA-methods), 73

qdaCMA, matrix, factor, missing-method
(qdaCMA-methods), 73

qdaCMA, matrix, numeric, missing-method
(qdaCMA-methods), 73

qdaCMA-methods, 71, 73

rfCMA, 4, 8, 10, 11, 15, 17, 20, 25, 29, 32, 42,
44, 46, 49, 53, 55, 59, 61, 63–65, 67,
69, 71, 72, 73, 75, 77, 79, 82, 85

rfCMA, data.frame, missing, formula-method
(rfCMA-methods), 75

rfCMA, expressionSet, character, missing-methods
(rfCMA-methods), 75

rfCMA, matrix, factor, missing-method
(rfCMA-methods), 75

rfCMA, matrix, numeric, missing-method
(rfCMA-methods), 75

rfCMA-methods, 73, 75

rfe (filter), 27

roc, 75

roc, cloutput-method (roc), 75

rocinternals (internals), 39

rocinternal (internals), 39

rowswaps (internals), 39

safeexp (internals), 39

scdaCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42,
44, 46, 49, 53, 55, 59, 61, 63, 65, 67,
69, 71, 72, 74, 76, 78, 79, 82, 85

scdaCMA, data.frame, missing, formula-method
(scdaCMA-methods), 78

scdaCMA, expressionSet, character, missing-method
(scdaCMA-methods), 78

scdaCMA, matrix, factor, missing-method
(scdaCMA-methods), 78

scdaCMA, matrix, numeric, missing-method
(scdaCMA-methods), 78

scdaCMA-methods, 78

show, cloutput-method (cloutput-class), 9

show, evaloutput-method
(evaloutput-class), 21

show, genesel-method (genesel-class), 34

show, learningsets-method
(learningsets-class), 47

show, predoutput-method
(predoutput-class), 70

show, tuningresult-method
(tuningresult-class), 87

show, wmcr.result-method
(wmcr.result-class), 92

shrinkldaCMA (filter), 27

shrinkldaCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32,
42, 44, 46, 49, 53, 55, 59, 61, 63, 65,
67, 69, 71, 72, 74, 77, 78, 80, 82

shrinkldaCMA, data.frame, missing, formula-method
(shrinkldaCMA-methods), 80

shrinkldaCMA, expressionSet, character, missing-method
(shrinkldaCMA-methods), 80

shrinkldaCMA, matrix, factor, missing-method
(shrinkldaCMA-methods), 80

shrinkldaCMA, matrix, numeric, missing-method
(shrinkldaCMA-methods), 80

shrinkldaCMA-methods, 78, 80, 80

summary, 80

summary, evaloutput-method (summary), 80

svmCMA, 4, 8, 10, 15, 17, 20, 25, 29, 32, 42, 44,
46, 49, 53, 55, 59, 61, 63, 65, 67, 69,
71, 72, 74, 77, 79, 81, 83, 85

svmCMA, data.frame, missing, formula-method
(svmCMA-methods), 83

svmCMA, expressionSet, character, missing-method
(svmCMA-methods), 83

svmCMA, matrix, factor, missing-method
(svmCMA-methods), 83

svmCMA, matrix, numeric, missing-method
(svmCMA-methods), 83

svmCMA-methods, 81, 83

toplist, 4, 5, 35, 83

toplist, genesel-method (toplist), 83

ttest (filter), 27

tune, 4, 8, 14, 31, 34, 38, 47, 48, 57, 58, 60,
62, 66, 69, 82, 84, 87–89, 91
tune, data.frame, missing, formula-method
   (tune-methods), 87
  tune, ExpressionSet, character, missing-method
    (tune-methods), 87
    tune, matrix, factor, missing-method
      (tune-methods), 87
    tune, matrix, numeric, missing-method
      (tune-methods), 87
    tune-methods, 87
    tuningresult, 5, 8, 57, 69, 86
    tuningresult (tuningresult-class), 87
    tuningresult-class, 87

  varseloutput, 11, 27
    varseloutput (varseloutput-class), 88
    varseloutput-class, 88

  weighted.mcr, 88, 90, 91
    weighted.mcr, character, character, missing, character, matrix, factor-method
      (weighted.mcr-methods), 90
    weighted.mcr, character, character, numeric, character, matrix, factor-method
      (weighted.mcr-methods), 90
    weighted.mcr, character, character, numeric, character, matrix, numeric-method
      (weighted.mcr-methods), 90
    weighted.mcr-methods, 90
  welchtest (filter), 27
  wilcoxtest (filter), 27
  wmc, 89, 90, 91
    wmc, matrix, numeric, numeric-method
      (wmc-methods), 91
    wmc-methods, 91
    wmcr.result (wmcr.result-class), 92
    wmcr.result-class, 92