Package 'rda'

July 2, 2014

Version 1.0.2-2

Date 2012-06-30

Title Shrunken Centroids Regularized Discriminant Analysis

Author Yaqian Guo <yaqiang@stat.stanford.edu> Trevor Hastie <hastie@stat.stanford.edu> Robert Tibshirani <tibs@stanford.edu>

Maintainer Rob Tibshirani <tibs@stanford.edu>

Depends R (>= 2.10)

Description Shrunken Centroids Regularized Discriminant Analysis for the classification purpose in high dimensional data.

License GPL (>= 2)

URL http://www.r-project.org

Repository CRAN

Date/Publication 2012-07-02 10:03:47

NeedsCompilation no

R topics documented:

brain *Brain Cancer Microarray Data*

Description

The brain data contains two objects: brain.x, microarray expression data for 42 brain cancer samples and brain.y, the class labels for these samples.

Usage

```
data(brain)
```

Format

An expression data matrix (42x5597) brain.x and a class label vector (42) brain.y for 42 samples.

Details

brain.y is the class labels of the 42 samples. brain.x is the microarray expression data matrix with each row representing a sample.

Source

Pomeroy, S. et al. (2002) *Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression*. Nature, Vol 415, p436-442. The data set is available at http://www.broad.mit.edu/mpr/CNS/.

References

Guo, Y. et al. (2004) *Regularized Discriminant Analysis and Its Application in Microarrays*, Technical Report, Department of Statistics, Stanford University.

brain.x *Brain Cancer Microarray Expression Data*

Description

brain.x is the microarray expression data for 42 brain cancer samples.

brain.y *Brain Cancer Microarray Data Class Labels*

Description

brain.y is the class labels for the brain cancer samples.

colon *Colon Cancer Microarray Data*

Description

The colon data contains two objects: colon.x, microarray expression data for 62 colon cancer samples and colon.y, the class labels for these samples.

Usage

```
data(colon)
```

Format

An expression data matrix (62x2000) colon.x and a class label vector (62) colon.y for 62 samples.

Details

colon.y is the class labels of the 62 samples. colon.x is the microarray expression data matrix with each row representing a sample.

Source

Alon, U. et al. (1999) *Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays*. PNAS, Vol 96, p6745-6750. The data set is available at http://microarray.princeton.edu/oncology/.

References

Guo, Y. et al. (2004) *Regularized Discriminant Analysis and Its Application in Microarrays*, Technical Report, Department of Statistics, Stanford University.

colon.x *Colon Cancer Microarray Expression Data*

Description

colon.x is the microarray expression data for 62 colon cancer samples.

colon.y *Colon Cancer Microarray Data Class Labels*

Description

colon.y is the class labels for the colon cancer samples.

genelist.rda *RDA Shrunken Gene List Function*

Description

A function that returns the shrunken gene (variable) names by RDA for a particular (alpha, delta) combination.

Usage

```
genelist.rda(x, y, alpha, delta, prior=table(y)/length(y),
             gnames=NULL, regularization="S")
```

Arguments

x	The training data set for which you want to obtain the shrunken gene list. It must be a numerical matrix. The columns are sample observations and the rows are variables.
y	The class labels for the columns of 'x'.
prior	A numerical vector that gives the prior proportion of each class. By default, it is set to be the sample frequencies unless users want to specify a different one.
alpha	A single regularization value for alpha. Users must supply this option.
delta	A threshold value for delta. Users must supply this option.
gnames	A character vector that specifies the names of the variables of the training data set 'x'. By default, it is set to be NULL and the function uses either the row names of 'x' (if it exists) or the row index 1:nrow(x). Users can provide their customized gene name list. However, the length of the name vector must be the same as the number of rows of 'x'.
regularization	The type of regularization. It is either 'S' or 'R'. The default value is 'S'.

Details

genelist.rda will return a vector of names for those shrunken genes by RDA for a particular (alpha, delta).

Value

A character vector of the names of the shrunken genes.

Author(s)

Yaqian Guo, Trevor Hastie and Robert Tibshirani

References

Guo, Y. et al. (2004) *Regularized Discriminant Analysis and Its Application in Microarrays*, Technical Report, Department of Statistics, Stanford University.

Examples

```
data(colon)
colon.x <- t(colon.x)
genenames <- genelist.rda(colon.x, colon.y, alpha=0.1, delta=0.3)
```

`plot.rdacv` *A function that plots the result from rda.cv*

Description

Plot the cross validation error matrix and the number of shrunken gene matrix obtained from RDA cross-validation analysis.

Usage

```
## S3 method for class 'rdacv'
plot(x, type=c("both", "error", "gene"), nice=FALSE, ...)
```

Arguments

`x`	The fit from `rda.cv`.
`type`	A character string specifying what to plot. If 'both', then heatmaps for both cv error and shrunken genes are plotted; if 'error', only the error map is produced; if 'gene', only the gene map is produced. This option is useful if users want to generate a specific plot. Default is 'both'.
`nice`	A logical flag. If 'TRUE', then 1-dim curves are plotted instead of heatmaps. This is useful when the length of alpha or delta is small. Heatmap in this case can be awful-looking. For example, if alpha=0.5 is a single value, while delta=seq(10), then both cv error and shrunken genes will be plotted as a 1-dim function of delta or vice versa when the length of delta is small.
`...`	Additional arguments for generic plot.

Details

`plot.rdacv` produces two heatmaps for the cross validation error matrix and the number of shrunken genes matrix obtained from `rda.cv`.

Value

A list of returning values:

`one.se.pos`	A 2-column matrix of the positions of the one standard error boundary points on the CV error heatmap. The first column indicates the alpha positions and the second column for the delta positions.
`min.cv.pos`	A 2-column matrix of the positions of the minimal CV error points on the CV error heatmap. The first column indicates the alpha positions and the second column for the delta positions.

Author(s)

Yaqian Guo, Trevor Hastie and Robert Tibshirani

References

Guo, Y. et al. (2004) *Regularized Discriminant Analysis and Its Application in Microarrays*, Technical Report, Department of Statistics, Stanford University.

Examples

```
data(colon)
fit <- rda(t(colon.x), colon.y)
fit.cv <- rda.cv(fit, x=t(colon.x), y=colon.y)
plot.rdacv(fit.cv)
```

---

predict.rda *RDA Prediction Function*

Description

A function that predicts the class labels for new samples using RDA.

Usage

```
## S3 method for class 'rda'
predict(object, x, y, xnew, prior, alpha, delta,
            type=c("class", "posterior", "nonzero"),
            trace=FALSE, ...)
```

Arguments

object	An rda fit object obtained from the function rda.
x	The training data matrix as used in the 'fit' object.
y	The class labels for the columns of 'x' as used in the 'fit' object.
xnew	The new data matrix used to predict the class labels of the new samples. Must be a numerical matrix with rows corresponding to variables and columns corresponding to the samples. The number of rows must be the same as 'x'.
prior	A numerical vector that gives the prior proportion of each class. By default, it is set to the fit component from the training step unless users want to specify a new one for prediction.
alpha	A particular regularization value for alpha. Often, this is the optimal alpha value obtained from the cross-validation step. But it could be any other value that users set. A vector of values is also acceptable. If missing, the function will use the default values from the fit component.

`delta`	A particular threshold value for delta. Often, this is the optimal delta value obtained from the cross-validation step. But it could be any other value that users set. A vector of values is also acceptable. If missing, the function will use the default values from the `fit` component.
`type`	A character string specifying which type of prediction is desired. If 'class', then the predicted class labels are returned; if 'posterior', then the predicted posterior probabilities for each sample belonging to a class are returned; if 'nonzero', then the indicators of shrunken genes are returned. 'class' is the default value.
`trace`	A logical flag indicating whether the intermediate steps should be printed.
`...`	Additional arguments for generic predict.

Details

`predict.rda` does various predictions on the new test samples based on fit from the training samples.

Value

If option "type='class'", the function will return the predicted class labels for the new test samples. The format is a 3-dim array. The first index corresponds to the alpha value(s) while the second index corresponds to the delta value(s). The last index is the predicted labels for the new samples. A reduced-dimensional array is possible if the length of alpha or delta is 1.

If option "type='posterior'", the function will return the predicted posterior probabilities of the new test samples belonging to different classes. The format is a 4-dim array. The first index corresponds to the alpha value(s) while the second index corresponds to the delta value(s). The third index represents the samples in 'xnew'. The last index is the class labels. A reduced-dimensional array is possible if the length of alpha or delta is 1.

If option "type='nonzero'", the function will return a 3-dim indicator array of the shrunken genes by RDA with 3 indices corresponding to alpha, delta and the indices of the genes respectively. A reduced-dimensional array is possible if the length of alpha or delta is 1.

Author(s)

Yaqian Guo, Trevor Hastie and Robert Tibshirani

References

Guo, Y. et al. (2004) *Regularized Discriminant Analysis and Its Application in Microarrays*, Technical Report, Department of Statistics, Stanford University.

See Also

Also see `rda` and `rda.cv`.

Examples

```
data(colon)
colon.x <- t(colon.x)

## divide the data set into a training set and a test
## set using a ratio of 2:1.
tr.index <- sample(1:62, 40)
fit <- rda(colon.x[, tr.index], colon.y[tr.index])

## predict the class labels of the test set at alpha=0.1
## and delta=0.5
ynew <- predict(fit, x=colon.x[, tr.index], y=colon.y[tr.index],
                xnew=colon.x[, -tr.index], alpha=0.1, delta=0.5)

## calculate the prediction error
sum(ynew != colon.y[-tr.index])
```

---

rda *Main RDA Function*

Description

The function that does RDA analysis on high dimensional data, e.g., microarray expression data.

Usage

```
rda(x, y, xnew=NULL, ynew=NULL, prior=table(y)/length(y),
    alpha=seq(0, 0.99, len=10), delta=seq(0, 3, len=10),
    regularization="S", genelist=FALSE, trace=FALSE)
```

Arguments

x	The training data set. It must be a numerical matrix. The columns are sample observations and the rows are variables. For example, in the microarray settings, "x" is the gene expression matrix with the columns corresponding to the arrays while the rows corresponding to the genes.
y	The class labels of the training samples (columns) in 'x', which must be consecutive integers starting from 1.
xnew	The test data matrix. It has the same structure as 'x'. The columns are samples and the rows are variables.
ynew	The class labels of the test samples. Same requirement as for 'y'.
prior	A numerical vector that gives the prior proportion of each class. Its length is equal to the number of classes. If not supplied, it is set to the sample proportions by default.
alpha	A numerical vector of the regularization values for alpha. A single value is allowed. If not supplied, the default one will be used.

delta	A numerical vector of the threshold values for delta. A single value is allowed. If not supplied, the default one will be used.
regularization	Define which regularization method to use. 'S' stands for regularization on covariance; 'R' stands for regularization on correlation. 'S' is the default option.
genelist	A logical flag. If 'TRUE', then the function will return an array of indices indicating the genes remained for each (alpha, delta) combination. By default, this is set to 'FALSE'.
trace	A logical flag. If 'TRUE', then the intermediate computation steps will be displayed. Caution: this would lead to a very long output display. By default, this is set to 'FALSE'.

Details

rda does RDA analysis on high dimensional data. This is the main function of the package.

Value

The function will return an 'rda' object with the following list of components:

alpha	The vector of the regularization values for alpha used in the function.
delta	The vector of the threshold values for delta used in the function.
prior	The vector of the prior proportion of each class used in the function.
error	The training error matrix. The rows correspond to the alpha values while the columns correspond to the delta values.
yhat	A 3-dim array giving the predicted class labels of 'y'. The first index corresponds to the alpha values while the second index corresponds to the delta values. The third index is the predicted class labels for the corresponding samples. However, when the length of alpha or delta is 1, this could be a 2-dim matrix or even a 1-dim vector.
ngene	The matrix of the number of shrunken genes. The rows correspond to the alpha values while the columns correspond to the delta values.
centroids	The group centroids matrix. It has the same number of rows as 'x' and the number of columns is the total number of classes. Each column is the centroids vector of the samples within that class.
centroid.overall	A single vector giving the grand mean vector of all the samples in the 'x' matrix.
yhat.new	A 3-dim array of the predicted class labels for the columns of 'xnew' if 'xnew' is provided. The first index corresponds to the alpha values while the second index corresponds to the delta values. The third index is the predicted class labels for the corresponding samples. However, when the length of alpha or delta is 1, this can be a 2-dim matrix or even a 1-dim vector.
posterior	A 4-dim array giving the posterior probabilities of each column of 'xnew' belonging to a class if 'xnew' is provided. The first index corresponds to the alpha values while the second index corresponds to the delta values. The third index is the corresponding columns in 'xnew'. The last index corresponds to different classes. However, an array of reduced dimensions may be produced if any of these four indices has length of 1.

testerror	The test error matrix if both xnew and ynew are supplied. The rows correspond to the alpha values while the columns correspond to the delta values.
gene.list	A 3-dim array giving the indicator whether a gene is shrunken or not for a particular (alpha, delta) if "genelist" option is 'TRUE'. '0' means that gene is shrunken while '1' otherwise. The first two indices correspond to alpha and delta. A reduced-dimensional array is possible if either alpha or delta is of length 1.
reg	The type of regularization used in calculation.

Author(s)

Yaqian Guo, Trevor Hastie and Robert Tibshirani

References

Guo, Y. et al. (2004) *Regularized Discriminant Analysis and Its Application in Microarrays*, Technical Report, Department of Statistics, Stanford University.

See Also

rda.cv and predict.rda.

Examples

```
data(colon)
colon.x <- t(colon.x)
fit <- rda(colon.x, colon.y)
```

---

rda.cv *RDA Cross Validation Function*

Description

A function that does RDA cross-validation analysis on the training data set.

Usage

```
rda.cv(fit, x, y, prior, alpha, delta, nfold=min(table(y), 10),
       folds=balanced.folds(y), trace=FALSE)
```

Arguments

fit	An rda fit object obtained from the rda function.
x	The training data set as used in the rda function.
y	The class labels of the training samples (columns) in "x" as used in rda function.
prior	A numerical vector that gives the prior proportion of each class. Its length should be equal to the number of classes. By default, the function uses the one coming along with the fit object unless users want to specify some other prior vector.

alpha	A numerical vector of the regularization values for alpha. By default, the function uses the one coming along with the fit object unless users want to do cross-validation based on some other values of alpha.
delta	A numerical vector of the threshold values for delta. By default, the function uses the one coming along with the fit object unless users want to do cross-validation based on some other values of delta.
nfold	An integer number to specify the number of folds in the cross-validation analysis. This option is overwritten when the folds option is specified at the same time.
folds	A list that provides the folds used in the cross-validation analysis. Each component of the list is an integer vector of the sample indices. See examples below for more details.
trace	A logical flag indicating whether the intermediate steps should be printed.

Details

rda.cv does the RDA-based cross-validation on the training data set.

Value

The rda.cv function will return an object of class rdacv with the following list of components:

alpha	The vector of the regularization values for alpha used in the cross-validation.
delta	The vector of the threshold values for delta used in the cross-validation.
prior	The vector of the prior proportion of each class used in the cross-validation.
nfold	The number of folds used in the cross-validation.
folds	The folds used in the cross-validation.
yhat.new	The 3-dim array of the predicted class labels of the training samples for each combination (alpha, delta). The first index corresponds to the alpha values while the second index corresponds to the delta values. The third index is the predicted class labels for the corresponding samples.
err	The training error matrix from cross-validation. The rows correspond to the alpha values while the columns correspond to the delta values. It is automatically generated by the function.
cv.err	The test error (or cross-validation error) matrix. The rows correspond to the alpha values while the columns correspond to the delta values.
ngene	The matrix of the number of shrunken genes. The rows correspond to the alpha values while the columns correspond to the delta values. Note: the number of shrunken genes is based on the average result from cross-validation.
reg	The type of regularization used in cross-validation.
n	The sample size of the training data set.

Author(s)

Yaqian Guo, Trevor Hastie and Robert Tibshirani

References

Guo, Y. et al. (2004) *Regularized Discriminant Analysis and Its Application in Microarrays*, Technical Report, Department of Statistics, Stanford University.

See Also

Also see rda and predict.rda.

Examples

```
data(colon)
colon.x <- t(colon.x)
fit <- rda(colon.x, colon.y)
fit.cv <- rda.cv(fit, x=colon.x, y=colon.y)

## to use the customized folds in cross-validation,
## for example, 6-fold with 11, 11, 10, 10, 10, 10 samples
## in the respective folds, you can do the follows:
index <- sample(1:62, 62)
folds <- list()
folds[[1]] <- index[1:11]
folds[[2]] <- index[12:22]
folds[[3]] <- index[23:32]
folds[[4]] <- index[33:42]
folds[[5]] <- index[43:52]
folds[[6]] <- index[53:62]
fit.cv <- rda.cv(fit, colon.x, colon.y, folds=folds)
```

Index