Package 'kmr4toxicogenetics'

August 17, 2017

Title Kernel multitask regression for toxicogenetics

Version 0.1.0

Description The package delivers a vignette containing code to reproduce the experiments in the paper ``Kernel multitask regression for toxicogenetics'' (bioRxiv-171298).

Depends R (>= 3.2)

Imports kmr (>= 0.1), glmnet, randomForest, parallel, Hmisc, stats, methods

Suggests knitr, rmarkdown, WGCNA, reshape2, gplots

License GPL-3 + file LICENSE

Encoding UTF-8

LazyData true

URL https://github.com/jpvert/kmr4toxicogenetics

BugReports https://github.com/jpvert/kmr4toxicogenetics/issues

RoxygenNote 6.0.1

VignetteBuilder knitr

R topics documented:

design 2
evaluateCV 2
id.rna 4
kcell 4
kchem 5
nontoxic 6
predictorElasticNet 6
predictorKMR 7
predictorLasso 8
predictorRF 9
tox 9

Index **11**

design *The design matrix*

Description

The design matrix processed from the Dream 8 toxicogenetics challenge dataset.

Usage

```
design
```

Format

A data frame with 884 cell lines in rows and 1,237 variables in columns of three categories:

covariate 16 variables binarized from 3 attributes (sex, population, batch) characterizing the nature of cell lines provided by the challenge.

RNAgram 337 variables corresponding to RNA Gram matrix (linear kernel) of normalized and NA-fixed RNA-seq counts of cell lines.

SNPgram 884 variables corresponding to genotypic SNP distance matrix (squared Euclidean distance) of SNP data of cell lines.

Note

RNA-seq data are available only for 337 cells from the challenge.

Source

https://doi.org/10.7303/syn1761567

evaluateCV *Performance evaluation by cross-validation*

Description

Evaluate regression performance of a predictor by cross-validation.

Usage

```
evaluateCV(mypredictor = c("predictorKMR", "predictorElasticNet",
  "predictorLasso", "predictorRF"), celllines, celllinesKernel, chemicals,
  chemicalsKernel, toxicity, nfolds = 5, nrepeats = 10, seed = 47,
  mc.cores = 1, ...)
```

Arguments

mypredictor	Character indicating a predictor function. Possible options are KMR (default), ElasticNet, Lasso and RF.
celllines	Matrix of descriptors for ncell cell lines, of dimension ncell x pcell. Only used for ElasticNet, Lasso and RF.
celllinesKernel	Kernel Gram matrix for ncell cell lines, of dimension ncell x ncell. Only used for KMR.
chemicals	Matrix of descriptors for nchem chemicals, of dimension nchem x pchem. Only used for ElasticNet, Lasso and RF.
chemicalsKernel	Kernel Gram matrix for nchem chemicals, of dimension nchem x nchem. Only used for KMR.
toxicity	Matrix of toxicity values for ncell cell lines responding to nchem chemicals, of dimension ncell x nchem.
nfolds	Number of folds for cross-validation. Default is 5.
nrepeats	Number of times the k-fold cross-validation is performed. Default is 1.
seed	A seed number for the random number generator (useful to have the same CV splits).
mc.cores	Number of parallelable CPU cores to use.
...	Other arguments to pass to predictor function.

Value

A list with matrices of cross-validation performance scores. Each score matrix is of dimension nexp x nchem (per CV experiment, per chemical) where nexp=nfolds*nrepeats and corresponds to one of the evaluation criteria:

matrix.ci	Concordance index.
matrix.rho	Pearson correlation.

References

Bernard, E., Jiao, Y., Scornet, E., Stoven, V., Walter, T., and Vert, J.-P. (2017). "Kernel multitask regression for toxicogenetics." bioRxiv-171298.

See Also

predictorKMR, predictorElasticNet, predictorLasso, predictorRF

id.rna *The identifiers of RNA-seq only cell lines*

Description

The identifiers of RNA-seq only cell lines of the Dream 8 toxicogenetics challenge dataset.

Usage

```
id.rna
```

Format

A vector of character strings of identifiers of the 337 RNA-seq only cell lines.

Source

https://doi.org/10.7303/syn1761567

References

Eduati, F., et al. "Prediction of human population responses to toxic compounds by a collaborative competition." Nature biotechnology 33.9 (2015): 933-940. doi:10.1038/nbt.3299.

kcell *Kernel matrices for cell lines*

Description

A list of example kernel matrices of cell line features.

Usage

```
kcell
```

Format

A list of 24 kernel matrices for cell lines of three categories of cell line features:

Kcovariates 4 kernel matrices of cell line covariates provided by the challenge, three of which correspond to a linear kernel of each attribute (sex, population, batch) plus one more combining all three attributes.

KrnaseqRbf 10 kernel matrices of normalized and NA-fixed RNA-seq counts of cell lines, corresponding to Gaussian RBF kernel with various bandwidth.

KsnpRbf 10 kernel matrices of genotypic SNP data of cell lines, corresponding to Gaussian RBF kernel with various bandwidth.

Source

https://doi.org/10.7303/syn1761567

References

Bernard, E., Jiao, Y., Scornet, E., Stoven, V., Walter, T., and Vert, J.-P. (2017). "Kernel multitask regression for toxicogenetics." bioRxiv-171298.

`kchem` *Kernel matrices for chemicals*

Description

A list of example kernel matrices of chemical features.

Usage

```
kchem
```

Format

A list of 32 kernel matrices for chemicals of five categories of cell line features:

KcdkRbf 10 kernel matrices of chemical descriptors calculated using the Chemistry Development Kit (CDK) provided by the challenge, corresponding to Gaussian RBF kernel with various bandwidth.

Kchemcpp 1 kernel matrix of a marginalized graph kernel, as implemented in the `ChemCPP` package, in the 2D structure of the chemicals.

KpredtargetRbf 10 kernel matrices based on the chemical descriptors by their predicted targets, corresponding to Gaussian RBF kernel with various bandwidth.

KsirmsRbf 10 kernel matrices of chemical descriptors generated by the Simplex representation of molecular structure (SIRMS) provided by the challenge, corresponding to Gaussian RBF kernel with various bandwidth.

Ksubstructure 1 kernel matrix (linear kernel) of the presence or absence of a list of predefined substructures from the PubChem fingerprint in the 2D structure of the chemicals.

Source

`https://doi.org/10.7303/syn1761567`

References

Bernard, E., Jiao, Y., Scornet, E., Stoven, V., Walter, T., and Vert, J.-P. (2017). "Kernel multitask regression for toxicogenetics." bioRxiv-171298.

nontoxic *Non-toxic chemical compounds*

Description

15 out of 106 chemical compounds in toxicity data that were shown to have no toxicity across the human cell population.

Usage

```
nontoxic()
```

Value

A sequence of identifiers for those 15 non-toxic chemical compounds in toxicity data.

References

Eduati, F., et al. "Prediction of human population responses to toxic compounds by a collaborative competition." Nature biotechnology 33.9 (2015): 933-940. doi:10.1038/nbt.3299.

predictorElasticNet *Wrapper function for elastic net regression*

Description

Wrapper function to perform elastic net regression with `cv.glmnet` that trains a model on training set and then predicts on test set for multiple tasks.

Usage

```
predictorElasticNet(patientsTrain, patientsTest, response, alpha = 0.5)
```

Arguments

`patientsTrain`	Matrix of training descriptors, of dimension `n x p`, for `n` training patients with `p` descriptors.
`patientsTest`	Matrix of test descriptors, of dimension `m x p`, for `m` test patients with the same set of `p` descriptors.
`response`	Matrix of observed toxicity values, of dimension `n x t`, for the `n` training patients responding to `t` drugs.
`alpha`	The elasticnet mixing parameter. `alpha=1` is the lasso penalty, and `alpha=0` the ridge penalty. Default is 0.5. All other arguments are taken by default implementation of `randomForest`.

Value

A matrix of predicted toxicity values, of dimension `m x t`, for the `m` test patients responding to the `t` drugs.

Note

Prediction is made per task with no special treatment for multitask learning, nor are task features needed.

See Also

cv.glmnet

---

predictorKMR *Wrapper function for kernel multitask regression*

Description

Wrapper function to perform kernel multitask regression with cv.kmr that trains a model on training set and then predicts on test set for multiple tasks.

Usage

```
predictorKMR(patientsKernelTrain, patientsKernelTest, response, drugsKernel,
  lambdas = exp(-15:25), nfolds = 5, nrepeats = 1)
```

Arguments

patientsKernelTrain	Precomputed kernel Gram matrix of n training patients, of dimension n x n.
patientsKernelTest	Precomputed kernel Gram matrix of m test patients crossing n training patients, of dimension m x n.
response	Matrix of observed toxicity values, of dimension n x t, for the n training patients responding to t drugs.
drugsKernel	Kernel Gram matrix of the t drugs, of dimension t x t.
lambdas	Sequence of lambdas that must be tested to fit a cross-validated KMR model. Default is exp(-15:25).
nfolds	Number of folds for cross-validation. Default is 5.
nrepeats	Number of times the k-fold cross-validation is performed. Default is 1.

Value

A matrix of predicted toxicity values, of dimension m x t, for the m test patients responding to the t drugs.

Note

Multitask prediction is made, for which task relationships are encoded in drugsKernel.

References

Bernard, E., Jiao, Y., Scornet, E., Stoven, V., Walter, T., and Vert, J.-P. (2017). "Kernel multitask regression for toxicogenetics." bioRxiv-171298.

See Also

cv.kmr

---

predictorLasso *Wrapper function for lasso regression*

Description

Wrapper function to perform lasso regression with cv.glmnet that trains a model on training set and then predicts on test set for multiple tasks.

Usage

```
predictorLasso(patientsTrain, patientsTest, response)
```

Arguments

patientsTrain	Matrix of training descriptors, of dimension n x p, for n training patients with p descriptors.
patientsTest	Matrix of test descriptors, of dimension m x p, for m test patients with the same set of p descriptors.
response	Matrix of observed toxicity values, of dimension n x t, for the n training patients responding to t drugs.

Value

A matrix of predicted toxicity values, of dimension m x t, for the m test patients responding to the t drugs.

Note

Prediction is made per task with no special treatment for multitask learning, nor are task features needed.

See Also

cv.glmnet, lasso implements a special case of elastic net predictorElasticNet

predictorRF *Wrapper function for random forest regression*

Description

Wrapper function to perform random forest regression with randomForest that trains a model on training set and then predicts on test set for multiple tasks.

Usage

```
predictorRF(patientsTrain, patientsTest, response, ntree = 500)
```

Arguments

patientsTrain	Matrix of training descriptors, of dimension n x p, for n training patients with p descriptors.
patientsTest	Matrix of test descriptors, of dimension m x p, for m test patients with the same set of p descriptors.
response	Matrix of observed toxicity values, of dimension n x t, for the n training patients responding to t drugs.
ntree	Number of trees to grow a random forest. Default is 500. All other arguments are taken by default implementation of randomForest.

Value

A matrix of predicted toxicity values, of dimension m x t, for the m test patients responding to the t drugs.

Note

Prediction is made per task with no special treatment for multitask learning, nor are task features needed.

See Also

randomForest

tox *The toxicity matrix*

Description

The toxicity values of cell lines, represented by the one-tenth maximal effective concentration-response exposure data (EC10) responding to various chemicals, from the Dream 8 toxicogenetics challenge dataset.

Usage

```
tox
```

Format

A data frame of EC10 values of 884 cell lines in rows and 91 toxic chemical compounds in columns.

Source

https://doi.org/10.7303/syn1761567

Index

∗Topic **datasets**
design, 2
id.rna, 4
kcell, 4
kchem, 5
tox, 9

cv.glmnet, *7*, *8*
cv.kmr, *8*

design, 2

evaluateCV, 2

id.rna, 4

kcell, 4
kchem, 5

nontoxic, 6

predictorElasticNet, *3*, 6, *8*
predictorKMR, *3*, 7
predictorLasso, *3*, 8
predictorRF, *3*, 9

randomForest, *9*

tox, 9