Package 'CALF'

May 29, 2020

Type Package

Title Coarse Approximation Linear Function

Version 1.0.15

Date 2020-05-5

Author Stephanie Lane [aut, cre], John Ford [aut], Clark Jeffries [aut], Diana Perkins [aut]

Maintainer John Ford <JoRuFo@gmail.com>

Description Forward selection linear regression greedy algorithm where selection is driven by optimization of Welch t-test p-value or AUC (binary dependent vector), or Pearson correlation (non-binary). Functions now enabled include data outputs for permutation tests of several types plus cross-validation features. Please see preprint at <https://www.biorxiv.org/content/10.1101/2020.03.27.011700v1> or <doi:10.1101/2020.03.27.

License GPL-2

Imports data.table, ggplot2

LazyData TRUE

RoxygenNote 7.1.0

Encoding UTF-8

NeedsCompilation no

Repository CRAN

Date/Publication 2020-05-28 23:20:22 UTC

R topics documented:

`CALF-package` *Coarse Approximation Linear Function*

Description

Forward selection linear regression greedy algorithm.

Details

The Coarse Approximation Linear Function (CALF) algorithm is a type of forward selection linear regression greedy algorithm. Nonzero weights are restricted to the values +1 and -1 and their number limited by an input parameter. CALF operates similarly on two different types of samples, binary and nonbinary, with some notable distinctions between the two. All sample data is provided to CALF as a data matrix. A binary sample must contain a distinguished first column with at least one 0 entries (e.g. controls) and at least one 1 entry (e.g. cases); at least one other column contains predictor values of some type. A nonbinary sample is similar but must contain a first column with real dependent (target) values. Columns containing values other that 0 or 1 must be normalized, e.g. as z-scores. As its score of differentiation, CALF uses either the Welch t-statistic p-value or AUC for binary samples and the Pearson correlation for non-binary samples, selected by input parameter. When initiated CALF selects from all predictors (markers) (first in the case of a tie) the one that yields the best score. CALF then checks if the number of selected markers is equal to the limit provided and terminates if so. Otherwise, CALF seeks a second marker, if any, that best improves the score of the sum function generated by adding the newly selected marker to the previous markers with weight +1 or weight -1. The process continues until the limit is reached or until no additional marker can be included in the sum to improve the score. By default, for binary samples, CALF assumes control data is designated with a 0 and case data with a 1. It is allowable to use the opposite convention, however the weights in the final sum may need to be reversed.

Author(s)

Stephanie Lane [aut, cre],
John Ford [aut],
Clark Jeffries [aut],
Diana Perkins [aut]

Maintainer: John Ford <JoRuFo@gmail.com>

`calf` *calf*

Description

Coarse Approximation Linear Function

Usage

```
calf(data, nMarkers, targetVector, optimize = "pval", verbose = FALSE)
```

Arguments

data	Matrix or data frame. First column must contain case/control dummy coded variable (if targetVector = "binary"). Otherwise, first column must contain real number vector corresponding to selection variable (if targetVector = "nonbinary"). All other columns contain relevant markers.
nMarkers	Maximum number of markers to include in creation of sum.
targetVector	Indicate "binary" for target vector with two options (e.g., case/control). Indicate "nonbinary" for target vector with real numbers.
optimize	Criteria to optimize, "pval" or "auc", (if targetVector = "binary") or "corr" (if targetVector = "nonbinary"). Defaults to "pval".
verbose	Logical. Indicate TRUE to print activity at each iteration to console. Defaults to FALSE.

Value

A data frame containing the chosen markers and their assigned weight (-1 or 1)

The optimal AUC, pval, or correlation for the classification.

If targetVector is binary, rocPlot. A plot object from ggplot2 for the receiver operating curve.

Examples

```
calf(data = CaseControl, nMarkers = 6, targetVector = "binary", optimize = "pval")
```

---

calf_exact_binary_subset

calf_exact_binary_subset

Description

Runs Coarse Approximation Linear Function on a random subset of binary data provided, with the ability to precisely control the number of case and control data used.

Usage

```
calf_exact_binary_subset(
  data,
  nMarkers,
  nCase,
  nControl,
  times = 1,
  optimize = "pval",
  verbose = FALSE
)
```

Arguments

data	Matrix or data frame. First column must contain case/control dummy coded variable.
nMarkers	Maximum number of markers to include in creation of sum.
nCase	Numeric. A value indicating the number of case data to use.
nControl	Numeric. A value indicating the number of control data to use.
times	Numeric. Indicates the number of replications to run with randomization.
optimize	Criteria to optimize. Indicate "pval" to optimize the p-value corresponding to the t-test distinguishing case and control. Indicate "auc" to optimize the AUC.
verbose	Logical. Indicate TRUE to print activity at each iteration to console. Defaults to FALSE.

Value

A data frame containing the chosen markers and their assigned weight (-1 or 1)

The optimal AUC or pval for the classification. If multiple replications are requested, a data.frame containing all optimized values across all replications is returned.

aucHist A histogram of the AUCs across replications, if applicable.

Examples

```
calf_exact_binary_subset(data = CaseControl, nMarkers = 6, nCase = 5, nControl = 8, times = 5)
```

---

calf_fractional *calf_fractional*

Description

Randomly selects from binary input provided to data parameter while ensuring the requested proportions of case and control variables are used and runs Coarse Approximation Linear Function.

Usage

```
calf_fractional(
  data,
  nMarkers,
  controlProportion = 0.8,
  caseProportion = 0.8,
  optimize = "pval",
  verbose = FALSE
)
```

Arguments

data	Matrix or data frame. Must be binary data such that the first column must contain case/control dummy coded variable, as function is only approprite for binary data.
nMarkers	Maximum number of markers to include in creation of sum.
controlProportion	Proportion of control samples to use, default is .8.
caseProportion	Proportion of case samples to use, default is .8.
optimize	Criteria to optimize, "pval" or "auc". Defaults to "pval".
verbose	Logical. Indicate TRUE to print activity at each iteration to console. Defaults to FALSE.

Value

A data frame containing the chosen markers and their assigned weight (-1 or 1)

The optimal AUC or pval for the classification.

rocPlot. A plot object from ggplot2 for the receiver operating curve.

Examples

```
calf_fractional(data = CaseControl, nMarkers = 6, controlProportion = .8, caseProportion = .4)
```

---

calf_randomize *calf_randomize*

Description

Randomly selects from binary input provided to data parameter and runs Coarse Approximation Linear Function.

Usage

```
calf_randomize(
  data,
  nMarkers,
  targetVector,
  times = 1,
  optimize = "pval",
  verbose = FALSE
)
```

Arguments

data	Matrix or data frame. Must be binary data such that the first column must contain case/control dummy coded variable, as function is only approprite for binary data.
nMarkers	Maximum number of markers to include in creation of sum.
targetVector	Indicate "binary" for target vector with two options (e.g., case/control). Indicate "nonbinary" for target vector with real numbers.
times	Numeric. Indicates the number of replications to run with randomization.
optimize	Criteria to optimize if targetVector = "binary." Indicate "pval" to optimize the p-value corresponding to the t-test distinguishing case and control. Indicate "auc" to optimize the AUC.
verbose	Logical. Indicate TRUE to print activity at each iteration to console. Defaults to FALSE.

Value

A data frame containing the chosen markers and their assigned weight (-1 or 1)

The optimal AUC, pval, or correlation for the classification.

aucHist A histogram of the AUCs across replications, if applicable.

Examples

```
calf_randomize(data = CaseControl, nMarkers = 6, targetVector = "binary", times = 5)
```

---

calf_subset *calf_subset*

Description

Runs Coarse Approximation Linear Function on a random subset of the data provided, resulting in the same proportion applied to case and control, when applicable.

Usage

```
calf_subset(
  data,
  nMarkers,
  proportion = 0.8,
  targetVector,
  times = 1,
  optimize = "pval",
  verbose = FALSE
)
```

Arguments

data	Matrix or data frame. First column must contain case/control dummy coded variable (if targetVector = "binary"). Otherwise, first column must contain real number vector corresponding to selection variable (if targetVector = "nonbinary"). All other columns contain relevant markers.
nMarkers	Maximum number of markers to include in creation of sum.
proportion	Numeric. A value between 0 and 1 indicating the proportion of cases and controls to use in analysis (if targetVector = "binary"). If targetVector = "nonbinary", this is just a proportion of the full sample. Used to evaluate robustness of solution. Defaults to 0.8.
targetVector	Indicate "binary" for target vector with two options (e.g., case/control). Indicate "nonbinary" for target vector with real numbers.
times	Numeric. Indicates the number of replications to run with randomization.
optimize	Criteria to optimize if targetVector = "binary." Indicate "pval" to optimize the p-value corresponding to the t-test distinguishing case and control. Indicate "auc" to optimize the AUC.
verbose	Logical. Indicate TRUE to print activity at each iteration to console. Defaults to FALSE.

Value

A data frame containing the chosen markers and their assigned weight (-1 or 1)

The optimal AUC, pval, or correlation for the classification. If multiple replications are requested, a data.frame containing all optimized values across all replications is returned.

aucHist A histogram of the AUCs across replications, if applicable.

Examples

```
calf_subset(data = CaseControl, nMarkers = 6, targetVector = "binary", times = 5)
```

CaseControl *Example data containing case and control data*

Description

This data contains 136 marker variables for 68 individuals who are distinguished as case/control.

Usage

```
data(CaseControl)
```

Format

A data frame with 136 marker variables and 68 individuals.

cv.calf *cv.calf*

Description

Performs cross-validation using CALF data input

Usage

```
cv.calf(
  data,
  limit,
  proportion = 0.8,
  times,
  targetVector,
  optimize = "pval",
  outputPath = NULL
)
```

Arguments

data	Matrix or data frame. First column must contain case/control dummy coded variable (if targetVector = "binary"). Otherwise, first column must contain real number vector corresponding to selection variable (if targetVector = "nonbinary"). All other columns contain relevant markers.
limit	Maximum number of markers to include in creation of sum.
proportion	Numeric. A value between 0 and 1 indicating the proportion of cases and controls to use in analysis (if targetVector = "binary") or proportion of the full sample (if targetVector = "nonbinary"). Defaults to 0.8.
times	Numeric. Indicates the number of replications to run with randomization.
targetVector	Indicate "binary" for target vector with two options (e.g., case/control). Indicate "nonbinary" for target vector with real numbers.
optimize	Criteria to optimize if targetVector = "binary." Indicate "pval" to optimize the p-value corresponding to the t-test distinguishing case and control. Indicate "auc" to optimize the AUC. Defaults to pval.
outputPath	The path where files are to be written as output, default is NULL meaning no files will be written. When targetVector is "binary" file binary.csv will be output in the provided path, showing the reults. When targetVector is "nonbinary" file nonbinary.csv will be output in the provided path, showing the results. In the same path, the kept and unkept variables from the last iteration, will be output, prefixed with the targetVector type "binary" or "nonbinary" followed by Kept and Unkept and suffixed with .csv. Two files containing the results from each run have List in the filenames and suffixed with .txt.

Value

A data frame containing "times" rows of CALF runs where each row represents a run of CALF on a randomized "proportion" of "data". Colunns start with the numer selected for the run, followed by AUC or pval and then all markers from "data". An entry in a marker column signifys a chosen marker for a particular run (a row) and their assigned coarse weight (-1, 0, or 1).

Examples

```
## Not run:
cv.calf(data = CaseControl, limit = 5, times = 100, targetVector = 'binary', optimize = 'pval')

## End(Not run)
```

`write.calf` *write.calf*

Description

Writes output of the CALF dataframe

Usage

```
write.calf(x, filename)
```

Arguments

`x`	A CALF data frame.
`filename`	The output filename

`write.calf_randomize` *write.calf_randomize*

Description

Writes output of the CALF randomize dataframe

Usage

```
write.calf_randomize(x, filename)
```

Arguments

`x`	A CALF randomize data frame.
`filename`	The output filename

write.calf_subset *write.calf_subset*

Description

Writes output of the CALF subset dataframe

Usage

```
write.calf_subset(x, filename)
```

Arguments

x	A CALF subset data frame.
filename	The output filename

Index