Package 'CALF'

May 19, 2017

Type Package

Title Coarse Approximation Linear Function

Version 0.2.0

Date 2017-05-15

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

Maintainer Stephanie Lane <slane@unc.edu>

Description Contains greedy algorithms for coarse approximation linear functions.

License GPL-2

Imports ggplot2

LazyData TRUE

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2017-05-19 15:08:51 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. The number of nonzero weights used is limited by a parameter. Samples are controls (at least 2) and cases (at least 2). A data matrix consists of a distinguished column that labels every row as either a control (0) or a case (1). Other columns (at least one) contain real number marker measurement data. Another input is a limit (positive integer) on the number of markers that can be selected for use in a linear sum. The present version uses as a score of differentiation the two-tailed, two sample unequal variance Student t-test p-value. Thus, any real-valued function applied to all samples generates values for controls and cases that are used to calculate the score. CALF selects the one marker (first in case of tie) that best distinguishes controls from cases (score is smallest p-value). CALF then checks the limit. If the number of selected markers is the limit, CALF ends. Else, 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.

Author(s)

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

Maintainer: Stephanie Lane <slane@unc.edu>

calf *calf*

Description

Coarse approximation linear function

Usage

```
calf(data, nMarkers, targetVector, margin, 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 = "real"). 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 "real" for target vector with real numbers.
margin	Real number from 0 to 1. Indicates the amount a potential marker must improve the target criterion (Pearson correlation or p-value) in order to add the marker.
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 AUC value for the classification

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

Examples

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

---

calf_randomize *calf_randomize*

Description

Coarse approximation linear function, randomized

Usage

```
calf_randomize(data, nMarkers, randomize = TRUE, targetVector, times = 1,
  margin = NULL, 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 = "real"). All other columns contain relevant markers.
nMarkers	Maximum number of markers to include in creation of sum.

randomize	Logical. Indicate TRUE to randomize the case/control status (or real number vector) for each individual. Used to compare results from true data with results from randomized data.
targetVector	Indicate "binary" for target vector with two options (e.g., case/control). Indicate "real" for target vector with real numbers.
times	Numeric. Indicates the number of replications to run with randomization.
margin	Real number from 0 to 1. Indicates the amount a potential marker must improve the target criterion (Pearson correlation or p-value) in order to add the marker.
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 AUC value for the classification

aucHist A histogram of the AUCs across replications.

Examples

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

`calf_subset` *calf_subset*

Description

Coarse approximation linear function, randomized

Usage

```
calf_subset(data, nMarkers, proportion = 0.8, targetVector, times = 1,
  margin = NULL, 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 = "real"). All other columns contain relevant markers.
nMarkers	Maximum number of markers to include in creation of sum.

proportion	Numeric. A value (where 0 < proportion <= 1) indicating the proportion of cases and controls to use in analysis (if targetVector = "binary"). If targetVector = "real", 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 "real" for target vector with real numbers.
times	Numeric. Indicates the number of replications to run with randomization.
margin	Real number from 0 to 1. Indicates the amount a potential marker must improve the target criterion (Pearson correlation or p-value) in order to add the marker.
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 AUC value for the classification. If multiple replications are requested, this will be a data.frame containing all AUCs across replications.

aucHist A histogram of the AUCs across replications.

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

```
CaseControl
```

Format

A data frame with 136 marker variables and 68 individuals.

Index