Package 'bolasso'

October 12, 2022

Title Model Consistent Lasso Estimation Through the Bootstrap

Version 0.2.0

Description Implements the bolasso algorithm for consistent variable selection and estimation accuracy. Includes support for many parallel backends via the future package. For details see: Bach (2008), 'Bolasso: model consistent Lasso estimation through the bootstrap', <arXiv:0804.1302>.

Depends Matrix (>= 1.0-6), R (>= 3.6.0)

Imports future.apply (>= 1.1.0), gamlr (>= 1.0), ggplot2, glmnet (>= 3.0), progressr, Rdpack, stats, tibble

Suggests testthat (>= 3.0.0), mlbench, covr

RdMacros Rdpack

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.2

URL https://www.dmolitor.com/bolasso/

Config/testthat/edition 3

NeedsCompilation no

Author Daniel Molitor [aut, cre]

Maintainer Daniel Molitor <molitdj97@gmail.com>

Repository CRAN

Date/Publication 2022-05-09 08:10:10 UTC

R topics documented:

`bolasso` *Bootsrap-enhanced Lasso*

Description

This function implements model-consistent Lasso estimation through the bootstrap. It supports parallel processing by way of the future package, allowing the user to flexibly specify many parallelization methods. This method was developed as a variable-selection algorithm, but this package also supports making ensemble predictions on new data using the bagged Lasso models.

Usage

```
bolasso(
  formula,
  data,
  n.boot = 100,
  progress = TRUE,
  implement = "glmnet",
  x = NULL,
  y = NULL,
  ...
)
```

Arguments

`formula`	An optional object of class formula (or one that can be coerced to that class): a symbolic description of the model to be fitted. Can be omitted when `x` and `y` are non-missing.
`data`	An optional object of class data.frame that contains the modeling variables referenced in `form`. Can be omitted when `x` and `y` are non-missing.
`n.boot`	An integer specifying the number of bootstrap replicates.
`progress`	A boolean indicating whether to display progress across bootstrap folds.
`implement`	A character; either 'glmnet' or 'gamlr', specifying which Lasso implementation to utilize. For specific modeling details, see `glmnet::cv.glmnet` or `gamlr::cv.gamlr`.
`x`	An optional predictor matrix in lieu of `form` and `data`.
`y`	An optional response vector in lieu of `form` and `data`.
`...`	Additional parameters to pass to either `glmnet::cv.glmnet` or `gamlr::cv.gamlr`.

Value

An object of class `bolasso`. This object is a list of length `n.boot` of `cv.glmnet` or `cv.gamlr` objects.

References

Bach FR (2008). "Bolasso: model consistent Lasso estimation through the bootstrap." *CoRR*, **abs/0804.1302**. 0804.1302, https://arxiv.org/abs/0804.1302.

See Also

glmnet::cv.glmnet and gamlr::cv.gamlr for full details on the respective implementations and arguments that can be passed to `...`.

Examples

```
mtcars[, c(2, 10:11)] <- lapply(mtcars[, c(2, 10:11)], as.factor)
idx <- sample(nrow(mtcars), 22)
mtcars_train <- mtcars[idx, ]
mtcars_test <- mtcars[-idx, ]

## Formula Interface

# Train model
set.seed(123)
bolasso_form <- bolasso(
  form = mpg ~ .,
  data = mtcars_train,
  n.boot = 20,
  nfolds = 5,
  implement = "glmnet"
)

# Extract selected variables
selected_vars(bolasso_form, threshold = 0.9, select = "lambda.min")

# Bagged ensemble prediction on test data
predict(bolasso_form,
        new.data = mtcars_test,
        select = "lambda.min")

## Alternal Matrix Interface

# Train model
set.seed(123)
bolasso_mat <- bolasso(
  x = model.matrix(mpg ~ . - 1, mtcars_train),
  y = mtcars_train[, 1],
  data = mtcars_train,
  n.boot = 20,
  nfolds = 5,
  implement = "glmnet"
)

# Extract selected variables
selected_vars(bolasso_mat, threshold = 0.9, select = "lambda.min")
```

```
# Bagged ensemble prediction on test data
predict(bolasso_mat,
        new.data = model.matrix(mpg ~ . - 1, mtcars_test),
        select = "lambda.min")
```

selected_vars *Bolasso-selected Variables*

Description

Identifies independent variables that are selected by the Bolasso algorithm at least the fraction of the time specified by the user-defined threshold. The typical value for this threshold is 0.9 and typically shouldn't be lower than that.

Usage

```
selected_vars(object, threshold = 0.9, summarise = TRUE, ...)
```

Arguments

`object`	An object of class bolasso.
`threshold`	A numeric between 0 and 1, specifying the fraction of bootstrap replicates for which Lasso must select a variable for it to be considered a selected variable.
`summarise`	A Boolean indicator where `FALSE` indicates returning the full set of coefficients at the selected variable/bootstrap replicate level and `TRUE` indicates taking the average of each variable's coefficient across bootstrap replicates. The default value is `TRUE` as it's more efficient and interpretable.
`...`	Additional arguments to pass to `predict` on objects with class cv.glmnet or cv.gamlr.

Value

A tibble with each selected variable and its respective coefficient for each bootstrap replicate.

See Also

`glmnet::predict.glmnet()` and `gamlr:::predict.gamlr` for details on additional arguments to pass to `...`.

Index