Package 'bigdata'

July 2, 2014

Type Package

Title Big Data Analytics

Version 0.1

Date 2011-02-12

Author Han Liu, Tuo Zhao

Maintainer Han Liu <hanliu@cs.jhu.edu>

Depends glmnet, Matrix, lattice,

Description The big data package is a collection of scalable methods for large-scale data analysis.

License GPL-2

Repository CRAN

Date/Publication 2012-04-06 15:55:53

NeedsCompilation no

R topics documented:

`bigdata-package` *Big Data Analytics*

Description

a collection of scalable methods for large-scale data analysis.

Details

Package:	bigdata
Type:	Package
Version:	0.1
Date:	2012-04-06
License:	GPL-2
LazyLoad:	yes

Author(s)

Han Liu, Tuo Zhao Maintainers: Han Liu <hanliu@cs.jhu.edu>

References

1. Han Liu, Kathryn Roeder and Larry Wasserman. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Advances in Neural Information Processing Systems*(NIPS), 2010.

See Also

lasso.stars

`lasso.stars` *Stability Approach to Regularization Selection for Lasso*

Description

Implements the Stability Approach to Regularization Selection (StARS) for Lasso

Usage

```
lasso.stars(x, y, rep.num = 20, lambda = NULL, nlambda = 100,
lambda.min.ratio = 0.001, stars.thresh = 0.1, sample.ratio = NULL,
alpha = 1, verbose = TRUE)
```

Arguments

`x`	The `n` by `d` data matrix representing `n` observations in `d` dimensions
`y`	The `n`-dimensional response vector
`rep.num`	The number of subsampling for StARS. The default value is `20`.

`lambda` A sequence of decresing positive numbers to control regularization. Typical usage is to leave the input `lambda = NULL` and have the program compute its own `lambda` sequence based on `nlambda` and `lambda.min.ratio`. Users can also specify a sequence to override this. Use with care - it is better to supply a decreasing sequence values than a single (small) value.

`nlambda` The number of regularization paramters. The default value is `100`.

`lambda.min.ratio`
The smallest value for `lambda`, as a fraction of the uppperbound (`MAX`) of the regularization parameter which makes all estimates equal to `0`. The program can automatically generate `lambda` as a sequence of length = `nlambda` starting from `MAX` to `lambda.min.ratio*MAX` in log scale. The default value is `0.001`.

`stars.thresh` The threshold of the variability in StARS. The default value is `0.1`. The alternative value is `0.05`. Only applicable when `criterion = "stars"`

`sample.ratio` The subsampling ratio. The default value is `10*sqrt(n)/n` when `n>144` and `0.8` when `n<=144`, where `n` is the sample size.

`alpha` The tuning parameter for the elastic-net regression. The default value is `1` (lasso).

`verbose` If `verbose = FALSE`, tracing information printing is disabled. The default value is `TRUE`.

Details

StARS selects the optimal regularization parameter based on the variability of the solution path. It chooses the least sparse graph among all solutions with the same variability. An alternative threshold `0.05` is chosen under the assumption that the model is correctly specified. In applications, the model is usually an approximation of the true model, `0.1` is a safer choice. The implementation is based on the popular package "glmnet".

Value

An object with S3 class "stars" is returned:

`path` The solution path of regression coefficients (in an `d` by `nlambda` matrix)

`lambda` The regularization parameters used in Lasso

`opt.index` The index of the optimal regularization parameter.

`opt.beta` The optimal regression coefficients.

`opt.lambda` The optimal regularization parameter.

`Variability` The variability along the solution path.

Note

This function can only work under the setting when `d>1`

Author(s)

Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman
Maintainers: Tuo Zhao<tourzhao@andrew.cmu.edu>; Han Liu <hanliu@cs.jhu.edu>

References

1.Han Liu, Kathryn Roeder and Larry Wasserman. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Advances in Neural Information Processing Systems*, 2010.
2.Jerome Friedman, Trevor Hastie and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, Vol.33, No.1, 2008.

See Also

bigdata-package

Examples

```
#generate data
x = matrix(rnorm(50*80),50,80)
beta = c(3,2,1.5,rep(0,77))
y = rnorm(50) + x%*%beta

#StARS for Lasso
z1 = lasso.stars(x,y)
summary(z1)
plot(z1)

#StARS for Lasso
z2 = lasso.stars(x,y, stars.thresh = 0.05)
summary(z2)
plot(z2)

#StARS for Lasso
z3 = lasso.stars(x,y,rep.num = 50)
summary(z3)
plot(z3)
```

---

plot.stars *Plot function for S3 class "stars"*

Description

Visualize the solution path and plot the optimal solution by model selection

Usage

```
## S3 method for class 'stars'
plot(x, ...)
```

Arguments

x	An object with S3 class "stars"
...	System reserved (No specific usage)

Author(s)

Han Liu and Tuo Zhao
Maintainers: Han Liu <hanliu@cs.jhu.edu>

See Also

lasso.stars

`print.stars` *Print function for S3 class "stars"*

Description

Print the information about the solution path length and the degree of freedom's along the solution path.

Usage

```
## S3 method for class 'stars'
print(x, ...)
```

Arguments

`x`	An object with S3 class `"stars"`
`...`	System reserved (No specific usage)

Author(s)

Han Liu and Tuo Zhao
Maintainers: Han Liu <hanliu@cs.jhu.edu>

See Also

lasso.stars

Index