Package 'abcrlda'

October 19, 2019

Type Package

Title Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis

Version 0.1.1

Maintainer Dmitriy Fedorov <dmitriy.fedorov@nu.edu.kz>

Description This package offers methods to perform asymptotically bias-corrected regularized linear discriminant analysis for cost-sensitive binary classification.

Imports stats

License GPL-3

URL https://ieeexplore.ieee.org/document/8720003/

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

R topics documented:

`abcrlda` *Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis for Cost-Sensitive Binary Classification*

Description

Performs Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis

Usage

```
abcrlda(x, y, gamma = 1, cost = c(0.5, 0.5))
```

Arguments

x	Input matrix or data.frame of dimension nobs x nvars; each row is an observation vector.
y	a numeric or factor vector of class labels. Factor should have two levels or be a vector with two distinct values. If y is presented as a vector, it will be coerced into a factor. Length of y has to correspond to number of samples in x.
gamma	Regularization parameter γ{gamma} in the following equation $$W_{ABC}^{RLDA} = \gamma(x - \frac{\bar{x}_0 + \bar{x}_1}{2})^T H(\bar{x}_0 - \bar{x}_1) - log(\frac{C_{01}}{C_{10}}) + \breve{\omega}_{opt}$$ Formulas and derivations for parameters used in above equation can be found in the journal paper under reference section.
cost	parameter that controls prioretization of classes. This is a vector of length 1 or 2 where first value is C_{10} (represents prioretization of class 0) and second value if provided is C_{01} (represents prioretization of class 1). Default value is c(0.5, 0.5), so both classes have equal priority and risk essentially becomes equivalent to error rate. If single value is provided it should be normalized to be between 0 and 1 (but not including 0 or 1). This value will be assigned to C_{10} and C_{01} will be equal to $(1 - C_{10})$ In a vector of length 1, values bigger than 0.5 prioretizes correct classification of 0 class while values less than 0.5 prioretizes 1 class.

Value

An object of class "abcrlda" is returned which can be used for class prediction (see predict())

a	Slope of a discriminant hyperplane. W(**x**) = **a**' **x** + m.
m	Bias term. W(**x**) = **a**'**x** + m.
cost	Vector of cost values that were used to fit this model
ncost	Normilized cost such that C_{10} + C_{01} == 1.
gamma	Regularization parameter value provided during fitting.
lev	Levels. Corresponds to the labels in y.

Reference

A. Zollanvari, M. Abdirash, A. Dadlani and B. Abibullaev, "Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis for Cost-Sensitive Binary Classification," in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1300-1304, Sept. 2019. doi: 10.1109/LSP.2019.2918485 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8720003&isnumber=8770167

See Also

Other functions in the package: cross_validation, da_risk_estimator, grid_search, predict.abcrlda

Examples

```
data(iris)
train_data <- iris[which(iris[, ncol(iris)] == "virginica" |
                           iris[, ncol(iris)] == "versicolor"), 1:4]
train_label <- factor(iris[which(iris[, ncol(iris)] == "virginica" |
                                   iris[, ncol(iris)] == "versicolor"), 5])
```

```
model <- abcrlda(train_data, train_label, gamma = 0.5, cost = 0.75)
a <- predict(model, train_data)
# same params but more explicit
model <- abcrlda(train_data, train_label, gamma = 0.5, cost = c(0.75, 0.25))
b <- predict(model, train_data)
# same class costs ratio
model <- abcrlda(train_data, train_label, gamma = 0.5, cost = c(3, 1))
c <- predict(model, train_data)
# all this model will give the same predictions
all(a == b & a == c & b == c)
#' [1] TRUE
```

`cross_validation` *Cross Validation for separate sampling adjasted for cost*

Description

Cross Validation for separate sampling adjasted for cost

Usage

```
cross_validation(x, y, gamma = 1, cost = c(0.5, 0.5), nfolds = 10)
```

Arguments

`x`	Input matrix or data.frame of dimension `nobs` x `nvars`; each row is an observation vector.
`y`	a numeric or factor vector of class labels. Factor should have two levels or be a vector with two distinct values. If `y` is presented as a vector, it will be coerced into a factor. Length of `y` has to correspond to number of samples in `x`.
`gamma`	Regularization parameter γ{`gamma`} in the following equation $$W_{ABC}^{RLDA} = \gamma(x - \frac{\bar{x}_0 + \bar{x}_1}{2})^T H(\bar{x}_0 - \bar{x}_1) - log(\frac{C_{01}}{C_{10}}) + \breve{\omega}_{opt}$$ Formulas and derivations for parameters used in above equation can be found in the journal paper under reference section.
`cost`	parameter that controls prioretization of classes. This is a vector of length 1 or 2 where first value is C_{10} (represents prioretization of class 0) and second value if provided is C_{01} (represents prioretization of class 1). Default value is c(0.5, 0.5), so both classes have equal priority and risk essentially becomes equivalent to error rate. If single value is provided it should be normalized to be between 0 and 1 (but not including 0 or 1). This value will be assigned to C_{10} and C_{01} will be equal to $(1 - C_{10})$ In a vector of length 1, values bigger than 0.5 prioretizes correct classification of 0 class while values less than 0.5 prioretizes 1 class.
`nfolds`	number of fold to use with cross-validation. Default is 10.

Value

Returns average risk after cross validation

Reference

Braga-Neto, Ulisses & Zollanvari, Amin & Dougherty, Edward. (2014). Cross-Validation Under Separate Sampling: Strong Bias and How to Correct It. Bioinformatics (Oxford, England). 30. 10.1093/bioinformatics/btu527. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4296143/pdf/btu527.pdf

See Also

Other functions in the package: abcrlda, da_risk_estimator, grid_search, predict.abcrlda

Examples

```
data(iris)
train_data <- iris[which(iris[, ncol(iris)] == "virginica" |
                         iris[, ncol(iris)] == "versicolor"), 1:4]
train_label <- factor(iris[which(iris[, ncol(iris)] == "virginica" |
                                 iris[, ncol(iris)] == "versicolor"), 5])
cross_validation(train_data, train_label, gamma = 10)
```

---

da_risk_estimator *Double Asymptotic Risk Estimator*

Description

Generalized consistent estimator of risk

Usage

```
da_risk_estimator(object)
```

Arguments

object	An object of class "abcrlda".

Value

Calculates risk based on estimated class error rates and misclassification costs

$$\Re = \varepsilon_0 * cost_{10} + \varepsilon_1 * cost_{01}$$

Reference

A. Zollanvari, M. Abdirash, A. Dadlani and B. Abibullaev, "Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis for Cost-Sensitive Binary Classification," in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1300-1304, Sept. 2019. doi: 10.1109/LSP.2019.2918485 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8720003&isnumber=8770167

See Also

Other functions in the package: abcrlda, cross_validation, grid_search, predict.abcrlda

Examples

```
data(iris)
train_data <- iris[which(iris[, ncol(iris)] == "virginica" |
                         iris[, ncol(iris)] == "versicolor"), 1:4]
train_label <- factor(iris[which(iris[, ncol(iris)] == "virginica" |
                                 iris[, ncol(iris)] == "versicolor"), 5])
model <- abcrlda(train_data, train_label, gamma = 0.5, cost = 0.75)
da_risk_estimator(model)
```

`grid_search` *Grid Search*

Description

Performs grid search for optimal hyperparameters (codegamma and codecost) within specified space based on double asymptotic risk estimation or cross validation. Double asymptotic risk estimation is faster option because it uses closed form formula for risk estimation. For further details refer to paper in the refernce section.

$$\Re = \varepsilon_0 * cost_{10} + \varepsilon_1 * cost_{01}$$

$$\varepsilon_i = \Phi(\frac{(-1)^{i+1}(\hat{G}_i + \hat{\omega}_{opt}/\gamma)}{\sqrt{\hat{D}}})$$

Cross validation was adapted to work with cost based risk estimation and works optimally with separate sampling

Usage

```
grid_search(x, y, range_gamma, range_cost, method = "estimator",
  nfolds = 10)
```

Arguments

`x`	Input matrix or data.frame of dimension `nobs` x `nvars`; each row is an observation vector.
`y`	a numeric or factor vector of class labels. Factor should have two levels or be a vector with two distinct values. If `y` is presented as a vector, it will be coerced into a factor. Length of `y` has to correspond to number of samples in `x`.
`range_gamma`	vector of gamma values to check
`range_cost`	nobs x 1 vector (values should be between 0 and 1) or nobs x 2 matrix (each row is cost pair value c(C_{10}, C_{01})) of cost values to check
`method`	selects method to evaluete risk. "estimator" and "cross"
`nfolds`	number of fold to use with cross-validation. Default is 10.

Value

List of best founded parameters

`cost`	cost value for which risk estimates are lowest during the search.
`gamma`	gamma regularization parameter for which risk estimates are lowest during the search
`risk`	Smalest risk value estimated during grid search.

Reference

A. Zollanvari, M. Abdirash, A. Dadlani and B. Abibullaev, "Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis for Cost-Sensitive Binary Classification," in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1300-1304, Sept. 2019. doi: 10.1109/LSP.2019.2918485 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8720003&isnumber=8770167

Braga-Neto, Ulisses & Zollanvari, Amin & Dougherty, Edward. (2014). Cross-Validation Under Separate Sampling: Strong Bias and How to Correct It. Bioinformatics (Oxford, England). 30. 10.1093/bioinformatics/btu527. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4296143/pdf/btu527.pdf

See Also

Other functions in the package: abcrlda, cross_validation, da_risk_estimator, predict.abcrlda

Examples

```
data(iris)
train_data <- iris[which(iris[, ncol(iris)] == "virginica" |
                         iris[, ncol(iris)] == "versicolor"), 1:4]
train_label <- factor(iris[which(iris[, ncol(iris)] == "virginica" |
                                 iris[, ncol(iris)] == "versicolor"), 5])
cost_range <- seq(0.1, 0.9, by = 0.2)
gamma_range <- c(0.1, 1, 10, 100, 1000)

gs <- grid_search(train_data, train_label,
                  range_gamma = gamma_range,
                  range_cost = cost_range,
                  method = "estimator")
model <- abcrlda(train_data, train_label,
                 gamma = gs$gamma, cost = gs$cost)
predict(model, train_data)

cost_range <- matrix(1:10, ncol = 2)
gamma_range <- c(0.1, 1, 10, 100, 1000)

gs <- grid_search(train_data, train_label,
                  range_gamma = gamma_range,
                  range_cost = cost_range,
                  method = "cross")
model <- abcrlda(train_data, train_label,
                 gamma = gs$gamma, cost = gs$cost)
predict(model, train_data)
```

---

predict.abcrlda *Class Prediction for abcrlda objects*

Description

Computes class predictions for new data based on a given abcrlda object

Usage

```
## S3 method for class 'abcrlda'
predict(object, newx, ...)
```

Arguments

object	An object of class "abcrlda".
newx	Matrix of new values for x at which predictions are to be made.
...	Argument used by generic function predict(object, x, ...).

Value

Returns factor vector with predictions for each observation. Factor levels are inhereted from the object variable.

Reference

A. Zollanvari, M. Abdirash, A. Dadlani and B. Abibullaev, "Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis for Cost-Sensitive Binary Classification," in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1300-1304, Sept. 2019. doi: 10.1109/LSP.2019.2918485 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8720003&isnumber=8770167

See Also

Other functions in the package: abcrlda, cross_validation, da_risk_estimator, grid_search

Examples

```
data(iris)
train_data <- iris[which(iris[, ncol(iris)] == "virginica" |
                           iris[, ncol(iris)] == "versicolor"), 1:4]
train_label <- factor(iris[which(iris[, ncol(iris)] == "virginica" |
                                   iris[, ncol(iris)] == "versicolor"), 5])
model <- abcrlda(train_data, train_label, gamma = 0.5, cost = 0.75)
a <- predict(model, train_data)
# same params but more explicit
model <- abcrlda(train_data, train_label, gamma = 0.5, cost = c(0.75, 0.25))
b <- predict(model, train_data)
# same class costs ratio
model <- abcrlda(train_data, train_label, gamma = 0.5, cost = c(3, 1))
c <- predict(model, train_data)
# all this model will give the same predictions
all(a == b & a == c & b == c)
#' [1] TRUE
```

Index