Package 'abcrlda'

November 12, 2019

Type Package

Title Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis

Version 0.1.1

Author Dmitriy Fedorov [aut, cre],
Amin Zollanvari [aut],
Aresh Dadlani [aut],
Berdakh Abibullaev [aut]

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

Description
This package offers methods to perform asymptotically bias-corrected regularized linear discriminant analysis (ABC_RLDA) for cost-sensitive binary classification. The bias-correction is an estimate of the bias term added to regularized discriminant analysis (RLDA) that minimizes the overall risk. The default magnitude of misclassification costs are equal and set to 0.5; however, the package also offers the options to set them to some predetermined values or, alternatively, take them as hyperparameters to tune.

Imports stats

Suggests knitr, rmarkdown

VignetteBuilder knitr

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

Constructs Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis.

Usage

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

Arguments

x	Input matrix or data.frame of dimension nobs x nvars; each row is an feature vector.
y	A numeric vector or factor of class labels. Factor should have either 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 in the ABC-RLDA discriminant function given by: $$W_{ABC}^{RLDA} = \gamma(x - \frac{\bar{x}_0 + \bar{x}_1}{2})^T H(\bar{x}_0 - \bar{x}_1) - log(\frac{C_{01}}{C_{10}}) + \hat{\omega}_{opt}$$ $$H = (I_p + \gamma\hat{\Sigma})^{-1}$$ Formulas and derivations for parameters used in above equation can be found in the article under reference section.
cost	Parameter that controls the overall misclassification costs. This is a vector of length 1 or 2 where the first value is C_{10} (represents the cost of assigning label 1 when the true label is 0) and the second value, if provided, is C_{01} (represents the cost of assigning label 0 when the true label is 1). The default setting is c(0.5, 0.5), so both classes have equal misclassification costs If a single value is provided, it should be normalized to lie between 0 and 1 (but not including 0 or 1). This value will be assigned to C_{10} while C_{01} will be equal to $(1 - C_{10})$.
bias_correction	Takes in a boolean value. If bias_correction is TRUE, then asymptotic bias correction will be performed. Otherwise, (if bias_correction is FALSE) asymptotic bias correction will not be performed and the ABCRLDA is the classical RLDA. The default is TRUE.

Value

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

a	Coefficient vector of a discriminant hyperplane: W(**x**) = **a**' **x** + m.
m	Intercept of discriminant hyperplane: W(**x**) = **a**'**x** + m.
cost	Vector of cost values that are used to construct ABC-RLDA.
ncost	Normilized cost such that C_{10} + C_{01} = 1.
gamma	Regularization parameter value used in ABC_RLDA discriminant function.
lev	Levels corresponding 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, risk_calculate

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 adjusted for cost.*

Description

This function implements Cross Validation for separate sampling adjusted for cost.

Usage

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

Arguments

x	Input matrix or data.frame of dimension nobs x nvars; each row is an feature vector.
y	A numeric vector or factor of class labels. Factor should have either 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 in the ABC-RLDA discriminant function given by:

$$W_{ABC}^{RLDA} = \gamma(x - \frac{\bar{x}_0 + \bar{x}_1}{2})^T H(\bar{x}_0 - \bar{x}_1) - log(\frac{C_{01}}{C_{10}}) + \hat{\omega}_{opt}$$

$$H = (I_p + \gamma\hat{\Sigma})^{-1}$$

Formulas and derivations for parameters used in above equation can be found in the article under reference section.

`cost` Parameter that controls the overall misclassification costs. This is a vector of length 1 or 2 where the first value is C_{10} (represents the cost of assigning label 1 when the true label is 0) and the second value, if provided, is C_{01} (represents the cost of assigning label 0 when the true label is 1). The default setting is c(0.5, 0.5), so both classes have equal misclassification costs

If a single value is provided, it should be normalized to lie between 0 and 1 (but not including 0 or 1). This value will be assigned to C_{10} while C_{01} will be equal to $(1 - C_{10})$.

`nfolds` Number of folds to use with cross-validation. Default is 10. In case of imbalanced data, `nfolds` should not be greater than the number of observations in smaller class.

`bias_correction`

Takes in a boolean value. If `bias_correction` is TRUE, then asymptotic bias correction will be performed. Otherwise, (if `bias_correction` is FALSE) asymptotic bias correction will not be performed and the ABCRLDA is the classical RLDA. The default is TRUE.

Value

Returns list of parameters.

`risk_cross` Returns risk estimation where $\Re = \varepsilon_0 * C_{10} + \varepsilon_1 * C_{01}$

`e_0` Error estimate for class 0.

`e_1` Error estimate for class 1.

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`, `risk_calculate`

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

This function implements the generalized (double asymptotic) 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

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

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, risk_calculate

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 to estimate the optimal hyperparameters (gamma and cost) within specified space based on double asymptotic risk estimation or cross validation. Double asymptotic risk estimation is more efficient to compute because it uses closed form for risk estimation. For further details, refer to the article in the reference section.

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

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

Separate sampling cross-validation (see cross-validation function) was adapted to work with cost-based risk estimation.

Usage

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

Arguments

x	Input matrix or data.frame of dimension nobs x nvars; each row is an feature vector.
y	A numeric vector or factor of class labels. Factor should have either 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 folds to use with cross-validation. Default is 10. In case of imbalanced data, nfolds should not be greater than the number of observations in smaller class.
bias_correction	Takes in a boolean value. If bias_correction is TRUE, then asymptotic bias correction will be performed. Otherwise, (if bias_correction is FALSE) asymptotic bias correction will not be performed and the ABCRLDA is the classical RLDA. The default is TRUE.

Value

List of estimated 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	Lowest 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, risk_calculate

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

Classifies observations 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 (i.e., assigned labels) 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, risk_calculate

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
```

`risk_calculate` *Risk Calculate*

Description

Estimates risk and error by applying a constructed classifier (an object of class abcrlda) to a given set of observations.

Usage

```
risk_calculate(object, x_true, y_true)
```

Arguments

`object`	An object of class "abcrlda".
`x_true`	Matrix of values for x for which true class labels are known.
`y_true`	A numeric vector or factor of true class labels. Factor should have either two levels or be a vector with two distinct values. If `y_true` is presented as a vector, it will be coerced into a factor. Length of `y_true` has to correspond to number of samples in `x_test`.

Value

A list of parameters where

`actual_err0`	Error rate for class 0.
`actual_err1`	Error rate for class 1.
`actual_errTotal`	Error rate overall.
`actual_normrisk`	Risk value normilized to be between 0 and 1.
`actual_risk`	Risk value without normilization.

See Also

Other functions in the package: `abcrlda`, `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)
risk_calculate(model, train_data, train_label)
```

Index