Package ‘abcrlda’

May 28, 2020

**Type** Package

**Title** Asymptotically Bias-Corrected Regularized Linear Discriminant Analysis

**Version** 1.0.3

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

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

**Description**
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

**License** GPL-3

**URL** https://ieeexplore.ieee.org/document/8720003/,
https://dx.doi.org/10.1109/LSP.2019.2918485

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.1.0

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2020-05-28 04:20:02 UTC
R topics documented:

- abcrlda ................................................................. 2
- cross_validation ......................................................... 4
- da_risk_estimator ......................................................... 5
- grid_search ............................................................... 6
- predict.abcrlda ........................................................... 8
- risk_calculate ............................................................. 9

Index 11

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 $\gamma$ in the ABC-RLDA discriminant function given by:

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

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

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

Normalized 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


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
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 $\gamma$ in the ABC-RLDA discriminant function given by:

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

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

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 \( R = \varepsilon_0 \cdot C_{10} + \varepsilon_1 \cdot C_{01} \)

e_0 Error estimate for class 0.

e_1 Error estimate for class 1.

Reference


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)
Value

Calculates risk based on estimated class error rates and misclassification costs

$$\mathcal{R} = \varepsilon_0 \cdot C_{10} + \varepsilon_1 \cdot C_{01}$$

Reference


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.

$$\mathcal{R} = \varepsilon_0 \cdot C_{10} + \varepsilon_1 \cdot C_{01}$$

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

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

Usage

```r
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 evaluate 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

- **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

predict.abcrlda

Class Prediction for abcrlda objects

Description

Classifies observations based on a given abcrlda object.

Usage

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

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 inherited from the object variable.

Reference


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 normalized to be between 0 and 1.
actual_risk      Risk value without normalization.

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

abcrlda, 2, 5, 6, 8–10

cross_validation, 3, 4, 6, 8–10

da_risk_estimator, 3, 5, 5, 8–10

grid_search, 3, 5, 6, 6, 9, 10

predict.abcrlda, 3, 5, 6, 8, 8, 10

risk_calculate, 3, 5, 6, 8, 9, 9