Package ‘OTE’

October 12, 2022

Type Package

Title Optimal Trees Ensembles for Regression, Classification and Class Membership Probability Estimation

Version 1.0.1

Date 2020-04-18

Author Zardad Khan, Asma Gul, Aris Perperoglou, Osama Mahmoud, Werner Adler, Miftahuddin and Berthold Lausen

Maintainer Zardad Khan <zardadkhan@awkum.edu.pk>

Description Functions for creating ensembles of optimal trees for regression, classification (Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). (2019) <doi:10.1007/s11634-019-00364-9>) and class membership probability estimation (Khan, Z, Gul, A, Mahmoud, O, Miftahuddin, M, Perperoglou, A, Adler, W & Lausen, B (2016) <doi:10.1007/978-3-319-25226-1_34>) are given. A few trees are selected from an initial set of trees grown by random forest for the ensemble on the basis of their individual and collective performance. Three different methods of tree selection for the case of classification are given. The prediction functions return estimates of the test responses and their class membership probabilities. Unexplained variations, error rates, confusion matrix, Brier scores, etc. are also returned for the test data.

License GPL (>= 3)

Encoding UTF-8

Imports randomForest,stats

LazyData true

RoxygenNote 7.1.0

NeedsCompilation no

Repository CRAN

Date/Publication 2020-04-20 10:50:07 UTC

R topics documented:

`OTE-package` *Optimal Trees Ensembles for Regression, Classification and Class Membership Probability Estimation*

Description

Functions for creating ensembles of optimal trees for regression, classification and class membership probability estimation are given. A few trees are selected from an initial set of trees grown by random forest for the ensemble on the basis of their individual and collective performance. The prediction functions return estimates of the test responses/class labels and their class membership probabilities. Unexplained variations, error rates, confusion matrix, Brier scores, etc. for the test data are also returned. Three different methods for tree selection are given for the case of classification.

Details

Package:	OTE
Type:	Package
Version:	1.0.1
Date:	2020-04-18
License:	GPL-3

Author(s)

Zardad Khan, Asma Gul, Aris Perperoglou, Osama Mahmoud, Werner Adler, Miftahuddin and Berthold Lausen Maintainer: Zardad Khan <zardadkhan@awkum.edu.pk>

References

Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 1-20.

Body *Exploring Relationships in Body Dimensions*

Description

The Body data set consists of 507 observations on 24 predictor variables including age, weight, hight and 21 body dimensions. All the 507 observations are on individuals, 247 men and 260 women, in the age of twenties and thirties with a small number of old people. The class variable is gender having two categories male and female.

Usage

```
data(Body)
```

Format

A data frame with 507 observations recorded on the following 25 variables.

`Biacrom` The diameter of Biacrom taken in centimeter.

`Biiliac` "Pelvic breadth" measured in centimeter.

`Bitro` Bitrochanteric whole diameter measured in centimeter.

`ChestDp` The depth of Chest of a person in centimeter between sternum and spine at nipple level.

`ChestD` The diameter of Chest of a person in centimeter at nipple level.

`ElbowD` The sum of diameters of two Elbows in centimeter.

`WristD` Sum of two Wrists diameters in centimeter.

`KneeD` The sum of the diameters of two Knees in centimeter.

`AnkleD` The sum of the diameters of two Ankles in centimeter.

`ShoulderG` The wideness of shoulder in centimeter.

`ChestG` The circumference of chest centimeter taken at nipple line for males and just above breast tissue for females.

`WaistG` The circumference of Waist in centimeter taken as the average of contracted and relaxed positions at the narrowest part.

`AbdG` Girth of Abdomin in centimeter at umbilicus and iliac crest, where iliac crest is taken as a landmark.

`HipG` Girth of Hip in centimeter at level of bitrochanteric diameter.

`ThighG` Average of left and right Thigh girths in centimeter below gluteal fold.

`BicepG` Average of left and right Bicep girths in centimeter.

`ForearmG` Average of left and right Forearm girths, extended, palm up.

`KneeG` Average of left and right Knees girths over patella, slightly flexed position.

`CalfG` Average of right and left Calf maximum girths.

`AnkleG` Average of right and left Ankle minimum girths.

WristG Average of left and right minimum circumferences of Wrists.

Age Age in years

Weight Weight in kilogram

Height Height in centimeter

Gender Binary response with two categories; 1 - male, 0 - female

Source

Heinz, G., Peterson, L.J., Johnson, R.W. and Kerk, C.J. (2003), “Exploring Relationships in Body Dimensions”, Journal of Statistics Education , 11.

References

Hurley, C. (2012), “ gclus: Clustering Graphics”, R package version 1.3.1, https://CRAN.R-project.org/package=gclus.

Examples

```
data(Body)
str(Body)
```

Galaxy *Radial Velocity of Galaxy NGC7531*

Description

This data set is a record of radial velocity of a spiral galaxy that is measured at 323 points in its covered area of the sky. The positions of the measurements, that are in the range of seven slot crossing at the origin, are denoted by 4 variables.

Usage

```
data(Galaxy)
```

Format

A data frame with 324 observations recorded on the following 5 variables.

east.west It is the east-west coordinate where east is taken as negative, west is taken as positive and origin, (0,0), is close to the center of galaxy.

north.south It is the north-south coordinate where south is taken as negative, north is taken as positive and origin, (0,0), is near the center of galaxy.

angle It is the degrees of anti rotation (clockwise) from the slot horizon where the observation lies.

radial.position It is the signed distance from the center, (0,0), which is signed as negative if the east-west coordinate is negative.

velocity This is the response variable denoting the radial velocity(km/sec) of the galaxy.

Source

Buta, R. (1987), “The Structure and Dynamics of Ringed Galaxies, III: Surface Photometry and Kinematics of the Ringed Nonbarred Spiral NGC7531” The Astrophysical J. Supplement Ser. 64. 1–37.

Examples

```
data(Galaxy)
str(Galaxy)
```

`OTClass` *Train the ensemble of optimal trees for classification.*

Description

This function selects optimal trees for classification from a total of `t.initial` trees grown by random forest. Number of trees in the initial set, `t.initial`, is specified by the user. If not specified then the default `t.initial = 1000` is used.

Usage

```
OTClass(XTraining, YTraining, method=c("oob+independent","oob","sub-sampling"),
p = 0.1,t.initial = NULL,nf = NULL, ns = NULL, info = TRUE)
```

Arguments

`XTraining`	An `n x d` dimensional training data matrix/frame consiting of traing observation where `n` is the number of observations and `d` is the number of features.
`YTraining`	A vector of length `n` consisting of class labels for the training data. Should be binary (0,1).
`method`	Method used in the selection of optimal trees. `method="oob+independent"` used out-of-bag observation from the bootstrap sample taken for growing the individual tree for indidual tree assessment while an independent training data for their collective assessement. `method="oob"` use the out-of-bag observations both for individual and collective assessment. `method="sub-sampling"` uses a sub-sample of the training data for individual tree assessment as well as its contribution towards the ensemble.
`p`	Percent of the best `t.initial` trees to be selected on the basis of performance on out-of-bag observations.
`t.initial`	Size of the initial set of classification trees.
`nf`	Number of features to be sampled for spliting the nodes of the trees. If equal to `NULL` then the default `sqrt(number of features)` is executed.
`ns`	Node size: Minimal number of samples in the nodes. If equal to `NULL` then the default `1` is executed.
`info`	If `TRUE`, displays processing information.

Details

Large values are recommended for `t.initial` for better performance as possible under the available computational resources.

Value

A trained object consisting of the selected trees.

Note

Prior action needs to be taken in the case of missing values as the fuction can not handle them at the current version.

Author(s)

Zardad Khan <zkhan@essex.ac.uk>

References

Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 1-20.

Liaw, A. and Wiener, M. (2002) “Classification and regression by random forest” R news. 2(3). 18–22.

See Also

Predict.OTClass, OTReg, OTProb

Examples

```
#load the data

  data(Body)
  data <- Body

#Divide the data into training and test parts

  set.seed(9123)
  n <- nrow(data)
  training <- sample(1:n,round(2*n/3))
  testing <- (1:n)[-training]
  X <- data[,1:24]
  Y <- data[,25]

#Train OTClass on the training data

  Opt.Trees <- OTClass(XTraining=X[training,],YTraining = Y[training],
  t.initial=200,method="oob+independent")

#Predict on test data
```

```
  Prediction <- Predict.OTClass(Opt.Trees, X[testing,],YTesting=Y[testing])

#Objects returned

  names(Prediction)
  Prediction$Confusion.Matrix
  Prediction$Predicted.Class.Labels
```

OTProb *Train the ensemble of optimal trees for class membership probability estimation.*

Description

This function selects optimal trees for class membership probability estimation from a total of `t.initial` trees grown by random forest. Number of trees in the initial set, `t.initial`, is specified by the user. If not specified then the default `t.initial` = `1000` is used.

Usage

```
OTProb(XTraining, YTraining, p = 0.2, t.initial = NULL,
      nf = NULL, ns = NULL, info = TRUE)
```

Arguments

`XTraining` An `n x d` dimensional training data matrix/frame consiting of traing observation where `n` is the number of observations and `d` is the number of features.

`YTraining` A vector of length `n` consisting of class labels for the training data. Should be binary (0,1).

`p` Percent of the best `t.initial` trees to be selected on the basis of performance on out-of-bag observations.

`t.initial` Size of the initial set of probability estimation trees.

`nf` Number of features to be sampled for spliting the nodes of the trees. If equal to `NULL` then the default `sqrt(number of features)` is executed.

`ns` Node size: Minimal number of samples in the nodes. If equal to `NULL` then the default `5` is executed.

`info` If `TRUE`, displays processing information.

Details

Large values are recommended for `t.initial` for better performance as possible under the available computational resources.

Value

A trained object consisting of the selected trees.

Note

Prior action needs to be taken in case of missing values as the fuction can not handle them at the current version.

Author(s)

Zardad Khan <zkhan@essex.ac.uk>

References

Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 1-20.

Liaw, A. and Wiener, M. (2002) "Classification and regression by random forest" R news. 2(3). 18–22.

See Also

Predict.OTProb, OTReg, OTClass

Examples

```
#load the data

  data(Body)
  data <- Body

#Divide the data into training and test parts

  set.seed(9123)
  n <- nrow(data)
  training <- sample(1:n,round(2*n/3))
  testing <- (1:n)[-training]
  X <- data[,1:24]
  Y <- data[,25]

#Train OTClass on the training data

  Opt.Trees <- OTProb(XTraining=X[training,],YTraining = Y[training],t.initial=200)

#Predict on test data

  Prediction <- Predict.OTProb(Opt.Trees, X[testing,],YTesting=Y[testing])

#Objects returned

  names(Prediction)
```

```
Prediction$Brier.Score
Prediction$Estimated.Probabilities
```

OTReg *Train the ensemble of optimal trees for regression.*

Description

This function selects optimal trees for regression from a total of `t.initial` trees grown by random forest. Number of trees in the initial set, `t.initial`, is specified by the user. If not specified then the default `t.initial = 1000` is used.

Usage

```
OTReg(XTraining, YTraining, p = 0.2, t.initial = NULL,
      nf = NULL, ns = NULL, info = TRUE)
```

Arguments

`XTraining`	An `n x d` dimensional training data matrix/frame consiting of traing observation where `n` is the number of observations and `d` is the number of features.
`YTraining`	A vector of length `n` consisting of the values of the continuous response variable for the training data.
`p`	Percent of the best `t.initial` trees to be selected on the basis of performance on out-of-bag observations.
`t.initial`	Size of the initial set of regression trees.
`nf`	Number of features to be sampled for spliting the nodes of the trees. If equal to `NULL` then the default `sqrt(number of features)` is executed.
`ns`	Node size: Minimal number of samples in the nodes. If equal to `NULL` then the default `5` is executed.
`info`	If `TRUE`, displays processing information.

Details

Large values are recommended for `t.initial` for better performance as possible under the available computational resources.

Value

A trained object consisting of the selected trees for regression.

Note

Prior action needs to be taken in case of missing values as the fuction can not handle them at the current version.

Author(s)

Zardad Khan <zkhan@essex.ac.uk>

References

Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 1-20.

Liaw, A. and Wiener, M. (2002) “Classification and regression by random forest” R news. 2(3). 18–22.

See Also

`Predict.OTReg`, `OTProb`, `OTClass`

Examples

```
# Load the data

  data(Galaxy)
  data <- Galaxy

#Divide the data into training and test parts

  set.seed(9123)
  n <- nrow(data)
  training <- sample(1:n,round(2*n/3))
  testing <- (1:n)[-training]
  X <- data[,1:4]
  Y <- data[,5]

#Train OTReg on the training data

  Opt.Trees <- OTReg(XTraining=X[training,],YTraining = Y[training],t.initial=200)

#Predict on test data

  Prediction <- Predict.OTReg(Opt.Trees, X[testing,],YTesting=Y[testing])

#Objects returned

  names(Prediction)
  Prediction$Unexp.Variations
  Prediction$Pr.Values
  Prediction$Trees.Used
```

Predict.OTClass *Prediction function for the object returned by* OTClass

Description

This function provides prediction for test data on the trained OTClass object for classification.

Usage

```
Predict.OTClass(Opt.Trees, XTesting, YTesting)
```

Arguments

Opt.Trees — An object of class OptTreesEns.

XTesting — An m x d dimensional training data matrix/frame consiting of test observations where m is the number of observations and d is the number of features.

YTesting — Optional. A vector of length m consisting of class labels for the test data. Should be binary (0,1).

Value

A list with values

Error.Rate — Error rate of the clssifier for the observations in XTesting.

Confusion.Matrix — Confusion matrix based on the estimated class labels and the true class labels.

Estimated.Class — A vector of length m consisting of the estimated class labels for the observations in XTesting.

Author(s)

Zardad Khan <zkhan@essex.ac.uk>

References

Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 1-20.

Liaw, A. and Wiener, M. (2002) “Classification and regression by random forest” R news. 2(3). 18–22.

See Also

OTClass, OTReg, OTProb

Examples

```
#load the data

  data(Body)
  data <- Body

#Divide the data into training and test parts

  set.seed(9123)
  n <- nrow(data)
  training <- sample(1:n,round(2*n/3))
  testing <- (1:n)[-training]
  X <- data[,1:24]
  Y <- data[,25]

#Train OTClass on the training data

  Opt.Trees <- OTClass(XTraining=X[training,],YTraining = Y[training],
  t.initial=200, method="oob+independent")

#Predict on test data

  Prediction <- Predict.OTClass(Opt.Trees, X[testing,],YTesting=Y[testing])

#Objects returned

  names(Prediction)
  Prediction$Confusion.Matrix
  Prediction$Predicted.Class.Labels
```

Predict.OTProb *Prediction function for the object returned by* OTProb

Description

This function provides prediction for test data on the trained OTProb object for class membership probability estimation.

Usage

```
Predict.OTProb(Opt.Trees, XTesting, YTesting)
```

Arguments

Opt.Trees An object of class OptTreesEns.

XTesting An m x d dimensional training data matrix/frame consiting of test observations where m is the number of observations and d is the number of features.

YTesting Optional. A vector of length m consisting of class labels for the test data. Should be binary (0,1).

Value

A list with values

`Brier.Score` Brier Score based on the estimated probabilities and true class label in YTesting.

`Estimated.Probabilities`
A vector of length `m` consisting of the estimated class membership probabilities for the observation in XTesting

Author(s)

Zardad Khan <zkhan@essex.ac.uk>

References

Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 1-20.

Liaw, A. and Wiener, M. (2002) "Classification and regression by random forest" R news. 2(3). 18–22.

See Also

`OTProb`.

Examples

```
#load the data

  data(Body)
  data <- Body

#Divide the data into training and test parts

  set.seed(9123)
  n <- nrow(data)
  training <- sample(1:n,round(2*n/3))
  testing <- (1:n)[-training]
  X <- data[,1:24]
  Y <- data[,25]

#Train OTClass on the training data

  Opt.Trees <- OTProb(XTraining=X[training,],YTraining = Y[training],t.initial=200)

#Predict on test data

  Prediction <- Predict.OTProb(Opt.Trees, X[testing,],YTesting=Y[testing])

#Objects returned

  names(Prediction)
```

```
Prediction$Brier.Score
Prediction$Estimated.Probabilities
```

Predict.OTReg *Prediction function for the object returned by* OTReg

Description

This function provides prediction for test data on the trained OTReg object for the continuous response variable.

Usage

```
Predict.OTReg(Opt.Trees, XTesting, YTesting)
```

Arguments

Opt.Trees	An object of class OptTreesEns.
XTesting	An m x d dimensional training data matrix/frame consiting of test observations where m is the number of observations and *d* is the number of features.
YTesting	Optional. A vector of length m consisting of the values of the continuous response variable for the test data.

Value

A list with values

Unexp.Variations	Unexplained variations based on estimated response and given response.
Pr.Values	A vector of length m consisting of the estimated values for the response observations in XTesting

Author(s)

Zardad Khan <zkhan@essex.ac.uk>

References

Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2019). Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 1-20.

Liaw, A. and Wiener, M. (2002) “Classification and regression by random forest” R news. 2(3). 18–22.

See Also

OTProb, OTReg, OTClass

Examples

```
# Load the data

  data(Galaxy)
  data <- Galaxy

#Divide the data into training and test parts

  set.seed(9123)
  n <- nrow(data)
  training <- sample(1:n,round(2*n/3))
  testing <- (1:n)[-training]
  X <- data[,1:4]
  Y <- data[,5]

#Train oTReg on the training data

  Opt.Trees <- OTReg(XTraining=X[training,],YTraining = Y[training],t.initial=200)

#Predict on test data

  Prediction <- Predict.OTReg(Opt.Trees, X[testing,],YTesting=Y[testing])

#Objects returned

  names(Prediction)
  Prediction$Unexp.Variations
  Prediction$Pr.Values
  Prediction$Trees.Used
```

Index