Package 'GAMens'

October 12, 2022

Title Applies GAMbag, GAMrsm and GAMens Ensemble Classifiers for Binary Classification

Version 1.2.1

Author Koen W. De Bock, Kristof Coussement and Dirk Van den Poel

Maintainer Koen W. De Bock <kdebock@audencia.com>

Depends R (>= 2.4.0), splines, gam, mlbench, caTools

Description Implements the GAMbag, GAMrsm and GAMens ensemble classifiers for binary classification (De Bock et al., 2010) <doi:10.1016/j.csda.2009.12.013>. The ensembles implement Bagging (Breiman, 1996) <doi:10.1023/A:1010933404324>, the Random Subspace Method (Ho, 1998) <doi:10.1109/34.709601>, or both, and use Hastie and Tibshirani's (1990, ISBN:978-0412343902) generalized additive models (GAMs) as base classifiers. Once an ensemble classifier has been trained, it can be used for predictions on new data. A function for cross validation is also included.

License GPL (>= 2)

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2018-04-05 17:12:34 UTC

R topics documented:

GAMens *Applies the GAMbag, GAMrsm or GAMens ensemble classifier to a data set*

Description

Fits the GAMbag, GAMrsm or GAMens ensemble algorithms for binary classification using generalized additive models as base classifiers.

Usage

```
GAMens(formula, data, rsm_size = 2, autoform = FALSE, iter = 10, df = 4,
  bagging = TRUE, rsm = TRUE, fusion = "avgagg")
```

Arguments

formula	a formula, as in the gam function. Smoothing splines are supported as nonparametric smoothing terms, and should be indicated by s. See the documentation of s in the gam package for its arguments. The GAMens function also provides the possibility for automatic formula specification. See 'details' for more information.
data	a data frame in which to interpret the variables named in formula.
rsm_size	an integer, the number of variables to use for random feature subsets used in the Random Subspace Method. Default is 2. If rsm=FALSE, the value of rsm_size is ignored.
autoform	if FALSE (default), the model specification in formula is used. If TRUE, the function triggers automatic formula specification. See 'details' for more information.
iter	an integer, the number of base classifiers (GAMs) in the ensemble. Defaults to iter=10 base classifiers.
df	an integer, the number of degrees of freedom (df) used for smoothing spline estimation. Its value is only used when autoform = TRUE. Defaults to df=4. Its value is ignored if a formula is specified and autoform is FALSE.
bagging	enables Bagging if value is TRUE (default). If FALSE, Bagging is disabled. Either bagging, rsm or both should be TRUE
rsm	enables Random Subspace Method (RSM) if value is TRUE (default). If FALSE, RSM is disabled. Either bagging, rsm or both should be TRUE
fusion	specifies the fusion rule for the aggregation of member classifier outputs in the ensemble. Possible values are 'avgagg' (default), 'majvote', 'w.avgagg' or 'w.majvote'.

Details

The GAMens function applies the GAMbag, GAMrsm or GAMens ensemble classifiers (De Bock et al., 2010) to a data set. GAMens is the default with (bagging=TRUE and rsm=TRUE. For GAMbag, rsm should be specified as FALSE. For GAMrsm, bagging should be FALSE.

The GAMens function provides the possibility for automatic formula specification. In this case, dichotomous variables in data are included as linear terms, and other variables are assumed continuous, included as nonparametric terms, and estimated by means of smoothing splines. To enable automatic formula specification, use the generic formula [response variable name]~. in combination with autoform = TRUE. Note that in this case, all variables available in data are used in the model. If a formula other than [response variable name]~. is specified then the autoform option is automatically overridden. If autoform=FALSE and the generic formula [response variable name]~. is specified then the GAMs in the ensemble will not contain nonparametric terms (i.e., will only consist of linear terms).

Four alternative fusion rules for member classifier outputs can be specified. Possible values are 'avgagg' for average aggregation (default), 'majvote' for majority voting, 'w.avgagg' for weighted average aggregation, or 'w.majvote' for weighted majority voting. Weighted approaches are based on member classifier error rates.

Value

An object of class GAMens, which is a list with the following components:

GAMs	the member GAMs in the ensemble.
formula	the formula used tot create the GAMens object.
iter	the ensemble size.
df	number of degrees of freedom (df) used for smoothing spline estimation.
rsm	indicates whether the Random Subspace Method was used to create the GAMens object.
bagging	indicates whether bagging was used to create the GAMens object.
rsm_size	the number of variables used for random feature subsets.
fusion_method	the fusion rule that was used to combine member classifier outputs in the ensemble.
probs	the class membership probabilities, predicted by the ensemble classifier.
class	the class predicted by the ensemble classifier.
samples	an array indicating, for every base classifier in the ensemble, which observations were used for training.
weights	a vector with weights defined as (1 - error rate). Usage depends upon specification of fusion_method.

Author(s)

Koen W. De Bock <kdebock@audencia.com>, Kristof Coussement <K.Coussement@ieseg.fr> and Dirk Van den Poel <Dirk.VandenPoel@ugent.be>

References

De Bock, K.W. and Van den Poel, D. (2012): "Reconciling Performance and Interpretability in Customer Churn Prediction Modeling Using Ensemble Learning Based on Generalized Additive Models". Expert Systems With Applications, Vol 39, 8, pp. 6816–6826.

De Bock, K. W., Coussement, K. and Van den Poel, D. (2010): "Ensemble Classification based on generalized additive models". Computational Statistics & Data Analysis, Vol 54, 6, pp. 1535–1546.

Breiman, L. (1996): "Bagging predictors". Machine Learning, Vol 24, 2, pp. 123–140.

Hastie, T. and Tibshirani, R. (1990): "Generalized Additive Models", Chapman and Hall, London.

Ho, T. K. (1998): "The random subspace method for constructing decision forests". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 20, 8, pp. 832–844.

See Also

predict.GAMens, GAMens.cv

Examples

```
## Load data (mlbench library should be loaded)
library(mlbench)
data(Ionosphere)
IonosphereSub<-Ionosphere[,c("V1","V2","V3","V4","V5","Class")]

## Train GAMens using all variables in Ionosphere dataset
Ionosphere.GAMens <- GAMens(Class~., IonosphereSub ,4 , autoform=TRUE,
iter=10 )

## Compare classification performance of GAMens, GAMrsm and GAMbag ensembles,
## using 4 nonparametric terms and 2 linear terms
Ionosphere.GAMens <- GAMens(Class~s(V3,4)+s(V4,4)+s(V5,3)+s(V6,5)+V7+V8,
Ionosphere ,3 , autoform=FALSE, iter=10 )

Ionosphere.GAMrsm <- GAMens(Class~s(V3,4)+s(V4,4)+s(V5,3)+s(V6,5)+V7+V8,
Ionosphere ,3 , autoform=FALSE, iter=10, bagging=FALSE, rsm=TRUE )

Ionosphere.GAMbag <- GAMens(Class~s(V3,4)+s(V4,4)+s(V5,3)+s(V6,5)+V7+V8,
Ionosphere ,3 , autoform=FALSE, iter=10, bagging=TRUE, rsm=FALSE )

## Calculate AUCs (for function colAUC, load caTools library)
library(caTools)
GAMens.auc <- colAUC(Ionosphere.GAMens[[9]], Ionosphere["Class"]=="good",
plotROC=FALSE)
GAMrsm.auc <- colAUC(Ionosphere.GAMrsm[[9]], Ionosphere["Class"]=="good",
plotROC=FALSE)
GAMbag.auc <- colAUC(Ionosphere.GAMbag[[9]], Ionosphere["Class"]=="good",
plotROC=FALSE)
```

GAMens.cv *Runs v-fold cross validation with GAMbag, GAMrsm or GAMens ensemble classifier*

Description

In v-fold cross validation, the data are divided into v subsets of approximately equal size. Subsequently, one of the v data parts is excluded while the remainder of the data is used to create a GAMens object. Predictions are generated for the excluded data part. The process is repeated v times.

Usage

```
GAMens.cv(formula, data, cv, rsm_size = 2, autoform = FALSE, iter = 10,
  df = 4, bagging = TRUE, rsm = TRUE, fusion = "avgagg")
```

Arguments

formula	a formula, as in the gam function. Smoothing splines are supported as nonparametric smoothing terms, and should be indicated by s. See the documentation of s in the gam package for its arguments. The GAMens function also provides the possibility for automatic formula specification. See 'details' for more information.
data	a data frame in which to interpret the variables named in formula.
cv	An integer specifying the number of folds in the cross-validation.
rsm_size	an integer, the number of variables to use for random feature subsets used in the Random Subspace Method. Default is 2. If rsm=FALSE, the value of rsm_size is ignored.
autoform	if FALSE (by default), the model specification in formula is used. If TRUE, the function triggers automatic formula specification. See 'details' for more information.
iter	an integer, the number of base (member) classifiers (GAMs) in the ensemble. Defaults to iter=10 base classifiers.
df	an integer, the number of degrees of freedom (df) used for smoothing spline estimation. Its value is only used when autoform = TRUE. Defaults to df=4. Its value is ignored if a formula is specified and autoform is FALSE.
bagging	enables Bagging if value is TRUE (default). If FALSE, Bagging is disabled. Either bagging, rsm or both should be TRUE
rsm	enables Random Subspace Method (RSM) if value is TRUE (default). If FALSE, rsm is disabled. Either bagging, rsm or both should be TRUE
fusion	specifies the fusion rule for the aggregation of member classifier outputs in the ensemble. Possible values are 'avgagg' for average aggregation (default), 'majvote' for majority voting, 'w.avgagg' for weighted average aggregation based on base classifier error rates, or 'w.majvote' for weighted majority voting.

Value

An object of class GAMens.cv, which is a list with the following components:

foldpred	a data frame with, per fold, predicted class membership probabilities for the left-out observations.
pred	a data frame with predicted class membership probabilities.
foldclass	a data frame with, per fold, predicted classes for the left-out observations.
class	a data frame with predicted classes.
conf	the confusion matrix which compares the real versus predicted class memberships, based on the class object.

Author(s)

Koen W. De Bock <kdebock@audencia.com>, Kristof Coussement <K.Coussement@ieseg.fr> and Dirk Van den Poel <Dirk.VandenPoel@ugent.be>

References

De Bock, K.W. and Van den Poel, D. (2012): "Reconciling Performance and Interpretability in Customer Churn Prediction Modeling Using Ensemble Learning Based on Generalized Additive Models". Expert Systems With Applications, Vol 39, 8, pp. 6816–6826.

De Bock, K. W., Coussement, K. and Van den Poel, D. (2010): "Ensemble Classification based on generalized additive models". Computational Statistics & Data Analysis, Vol 54, 6, pp. 1535–1546.

Breiman, L. (1996): "Bagging predictors". Machine Learning, Vol 24, 2, pp. 123–140.

Hastie, T. and Tibshirani, R. (1990): "Generalized Additive Models", Chapman and Hall, London.

Ho, T. K. (1998): "The random subspace method for constructing decision forests". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 20, 8, pp. 832–844.

See Also

predict.GAMens, GAMens

Examples

```
## Load data: mlbench library should be loaded!)
library(mlbench)
data(Sonar)
SonarSub<-Sonar[,c("V1","V2","V3","V4","V5","V6","Class")]

## Obtain cross-validated classification performance of GAMrsm
## ensembles, using all variables in the Sonar dataset, based on 5-fold
## cross validation runs

Sonar.cv.GAMrsm <- GAMens.cv(Class~s(V1,4)+s(V2,3)+s(V3,4)+V4+V5+V6,
SonarSub ,5, 4 , autoform=FALSE, iter=10, bagging=FALSE, rsm=TRUE )

## Calculate AUCs (for function colAUC, load caTools library)
```

```
library(caTools)

GAMrsm.cv.auc <- colAUC(Sonar.cv.GAMrsm[[2]], SonarSub["Class"]=="R",
plotROC=FALSE)
```

`predict.GAMens` *Predicts from a fitted GAMens object (i.e., GAMbag, GAMrsm or GAMens classifier).*

Description

Generates predictions (classes and class membership probabilities) for observations in a dataframe using a GAMens object (i.e., GAMens, GAMrsm or GAMbag classifier).

Usage

```
## S3 method for class 'GAMens'
predict(object, data, ...)
```

Arguments

`object`	fitted model object of `GAMens` class.
`data`	data frame with observations to genenerate predictions for.
`...`	further arguments passed to or from other methods.

Value

An object of class `predict.GAMens`, which is a list with the following components:

`pred`	the class membership probabilities generated by the ensemble classifier.
`class`	the classes predicted by the ensemble classifier.
`conf`	the confusion matrix which compares the real versus predicted class memberships, based on the `class` object. Obtains value `NULL` if the testdata is unlabeled.

Author(s)

Koen W. De Bock <kdebock@audencia.com>, Kristof Coussement <K.Coussement@ieseg.fr> and Dirk Van den Poel <Dirk.VandenPoel@ugent.be>

References

De Bock, K.W. and Van den Poel, D. (2012): "Reconciling Performance and Interpretability in Customer Churn Prediction Modeling Using Ensemble Learning Based on Generalized Additive Models". Expert Systems With Applications, Vol 39, 8, pp. 6816–6826.

De Bock, K. W., Coussement, K. and Van den Poel, D. (2010): "Ensemble Classification based on generalized additive models". Computational Statistics & Data Analysis, Vol 54, 6, pp. 1535–1546.

Breiman, L. (1996): "Bagging predictors". Machine Learning, Vol 24, 2, pp. 123–140.

Hastie, T. and Tibshirani, R. (1990): "Generalized Additive Models", Chapman and Hall, London.

Ho, T. K. (1998): "The random subspace method for constructing decision forests". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 20, 8, pp. 832–844.

See Also

GAMens, GAMens.cv

Examples

```
## Load data, mlbench library should be loaded!)
library(mlbench)
data(Sonar)
SonarSub<-Sonar[,c("V1","V2","V3","V4","V5","V6","Class")]

## Select indexes for training set observations
idx <- c(sample(1:97,60),sample(98:208,70))

## Train GAMrsm using all variables in Sonar dataset. Generate predictions
## for test set observations.
Sonar.GAMrsm <- GAMens(Class~.,SonarSub[idx,], autoform=TRUE, iter=10,
bagging=FALSE, rsm=TRUE)
Sonar.GAMrsm.predict <- predict(Sonar.GAMrsm,SonarSub[-idx,])


## Load data mlbench library should be loaded!)
library(mlbench)
data(Ionosphere)
IonosphereSub<-Ionosphere[,c("V1","V2","V3","V4","V5","V6","V7","V8","Class")]
Ionosphere_s <- IonosphereSub[order(IonosphereSub$Class),]

## Select indexes for training set observations
idx <- c(sample(1:97,60),sample(98:208,70))


## Compare test set classification performance of GAMens, GAMrsm and
## GAMbag ensembles, using using 4 nonparametric terms and 2 linear terms in the
## Ionosphere dataset
Ionosphere.GAMens <- GAMens(Class~s(V3,4)+s(V4,4)+s(V5,3)+s(V6,5)+V7+V8,
IonosphereSub[idx,], autoform=FALSE, iter=10, bagging=TRUE, rsm=TRUE)

Ionosphere.GAMens.predict <- predict(Ionosphere.GAMens,
IonosphereSub[-idx,])
```

```
Ionosphere.GAMrsm <- GAMens(Class~s(V3,4)+s(V4,4)+s(V5,3)+s(V6,5)+V7+V8,
IonosphereSub[idx,], autoform=FALSE, iter=10, bagging=FALSE, rsm=TRUE)

Ionosphere.GAMrsm.predict <- predict(Ionosphere.GAMrsm,
IonosphereSub[-idx,])

Ionosphere.GAMbag <- GAMens(Class~s(V3,4)+s(V4,4)+s(V5,3)+s(V6,5)+V7+V8,
IonosphereSub[idx,], autoform=FALSE, iter=10, bagging=TRUE, rsm=FALSE)

Ionosphere.GAMbag.predict <- predict(Ionosphere.GAMbag,
IonosphereSub[-idx,])

## Calculate AUCs(for function colAUC, load caTools library)
library(caTools)
GAMens.auc <- colAUC(Ionosphere.GAMens.predict[[1]],
IonosphereSub[-idx,"Class"]=="good", plotROC=FALSE)

GAMrsm.auc <- colAUC(Ionosphere.GAMrsm.predict[[1]],
Ionosphere[-idx,"Class"]=="good", plotROC=FALSE)

GAMbag.auc <- colAUC(Ionosphere.GAMbag.predict[[1]],
IonosphereSub[-idx,"Class"]=="good", plotROC=FALSE)
```

Index