Package 'EBcoBART'

July 21, 2025

Title Co-Data Learning for Bayesian Additive Regression Trees

Version 1.1.1

Description Estimate prior variable weights for Bayesian Additive Regression Trees (BART). These weights correspond to the probabilities of the variables being selected in the splitting rules of the sum-of-trees.
Weights are estimated using empirical Bayes and external information on the explanatory variables (co-data).
BART models are fitted using the 'dbarts' 'R' package.
See Goedhart and others (2023) <doi:10.48550/arXiv.2311.09997> for details.

License GPL (>= 3)

Encoding UTF-8

URL https://github.com/JeroenGoedhart/EBcoBART

RoxygenNote 7.3.1

Imports dbarts, loo, posterior, univariateML, extraDistr, graphics

Depends R (>= 2.10)

LazyData true

NeedsCompilation no

Author Jeroen M. Goedhart [aut, cre, cph] (ORCID:
<https://orcid.org/0000-0003-0134-1897>),
Thomas Klausch [aut],
Mark A. van de Wiel [aut],
Vincent Dorie [ctb] (Author of 'dbarts' 'R' package and auxiliary function getDepth),
Hanarth Fonds [fnd]

Maintainer Jeroen M. Goedhart <jeroengoed@gmail.com>

Repository CRAN

Date/Publication 2025-01-14 16:00:01 UTC

Contents

`Bloodplatelet` *Bloodplatelet*

Description

Contains not standardized messenger-RNA expression measurements, derived from blood platelets, which are used to classify breast cancer versus non-small- cell lung cancer patients. For the 500 m-RNA variables, co-data is available. Co-data is defined by estimated p-values (- logit scale) of all the 500 m-RNA for three different classification tasks: 1) colorectal cancer vs. control patients, 2) pancreas cancer vs. control patients, and 3) pancreas cancer vs. colorectal cancer. Co-data is therefore informative if different cancer classification tasks have similar important m-RNA variables. See Novianti and others (2017) doi:10.1093/bioinformatics/btw837 for details on the complete data set, from which this data is derived.

Usage

```
data(Bloodplatelet)
```

Format

A list object with five objects:

Xtrain Data frame with 101 rows (samples) and 140 columns (variables). Explanatory variables used for fitting BART. Variable names are present.

Y Numeric of length 100. Binary training response (0: Breast cancer, 1: non-small-cell lung cancer)

CoData Matrix with 500 rows and 4 columns. Auxiliary information on the 500 variables. Contains, for each variable, estimated p-values from three different classification tasks. P-values are -logit transformed. An intercept is included to the co-data matrix.

Author(s)

Jeroen M. Goedhart, <j.m.goedhart@amsterdamumc.nl>

Mark A van de Wiel

References

P. W. Novianti, B.C. Snoek, S. Wilting, and M. A. Van De Wiel, Better diagnostic signatures from RNAseq data through use of auxiliary co-data 2017 Bioinformatics, Vol. 33, No. 10, p. 1572-1574

Dat_EBcoBART *Convenience function to correctly specify co-data matrix if X contains factor variables.*

Description

The R package dbarts uses dummy encoding for factor variables so the co-data matrix should contain co-data information for each dummy. If co-data is only available for the factor as a whole (e.g. factor belongs to a group), use this function to set-up the co-data in the right-format for the EBcoBART function.

Usage

```
Dat_EBcoBART(X, CoData)
```

Arguments

X	Explanatory variables. Should be a data.frame. The function is only useful when X contains factor variables.
CoData	The co-data model matrix with co-data information on explanatory variables in X. Should be a matrix, so not a data.frame. If grouping information is present, please encode this yourself using dummies with dummies representing which group a explanatory variable belongs to. The number of rows of the co-data matrix should equal the number of columns of X.

Value

A list object with X: the explanatory variables with factors encoded as dummies and CoData: the co-data matrix with now co-data for all dummies.

Author(s)

Jeroen M. Goedhart, <j.m.goedhart@amsterdamumc.nl>

Examples

```
p <- 15
n <- 30
X <- matrix(runif(n * p),nrow = n, ncol = p) #all continuous variables
Fact <- factor(sample(1:3, n, replace = TRUE)) # factor variables
X <- cbind.data.frame(X, Fact)
G <- 4   #number of groups for co-data
Co <- rep(1:G, rep(ncol(X)/G,G)) # first 4 covariates in group 1,
#2nd 4 covariates in group 2, etc..
Example <- data.frame(factor(Co))
Example <- stats::model.matrix(~ 0 + ., Example) # encode the grouping structure
# with dummies
Dat <- Dat_EBcoBART(X = X, CoData = Example)
```

```
X <- Dat$X
CoData <- Dat$CoData
```

EBcoBART *Learning prior covariate weights for BART models with empirical Bayes and co-data.*

Description

Function that estimates the prior probabilities of variables being selected in the splitting rules of Bayesian Additive Regression Trees (BART). Estimation is performed using empirical Bayes and co-data, i.e. external information on the explanatory variables.

Usage

```
EBcoBART(
  Y,
  X,
  model,
  CoData,
  nIter = 10,
  EB_k = FALSE,
  EB_alpha = FALSE,
  EB_sigma = FALSE,
  Prob_Init = c(rep(1/ncol(X), ncol(X))),
  verbose = FALSE,
  ndpost = 5000,
  nskip = 5000,
  nchain = 5,
  keepevery = 1,
  ntree = 50,
  alpha = 0.95,
  beta = 2,
  k = 2,
  sigest = stats::sd(Y) * 0.667,
  sigdf = 10,
  sigquant = 0.75
)
```

Arguments

Y	Response variable that can be either continuous or binary. Should be a numeric.
X	Explanatory variables. Should be a matrix. If X is a data.frame and contains factors, you may consider the function Dat_EBcoBART
model	What type of response variable Y. Can be either continuous or binary

CoData	The co-data model matrix with co-data information on explanatory variables in X. Should be a matrix, so not a data.frame. If grouping information is present, please encode this yourself using dummies with dummies representing which group a explanatory variable belongs to. The number of rows of the co-data matrix should equal the number of columns of X. If no CoData is available, but one aims to estimate either prior para- meter k, alpha or sigma, please specify CoData == NULL.
nIter	Number of iterations of the EM algorithm
EB_k	Logical (T/F). If true, the EM algorithm also estimates prior parameter k (of leaf node parameter prior). Defaults to False. Setting to true increases computational time.
EB_alpha	Logical (T/F). If true, the EM algorithm also estimates prior parameter alpha (of tree structure prior). Defaults to False. Setting to true increases computational time.
EB_sigma	Logical (T/F). If true, the EM algorithm also estimates prior parameters of the error variance. To do so, the algorithm estimates the degrees of freedom (sigdf) and the quantile (sigest) at which sigquant of the probability mass is placed. Thus, the specified sigquant is kept fixed and sigdf and sigest are updated. Defaults to False.
Prob_Init	Initial vector of splitting probabilities for explanatory variables X. Length should equal number of columns of X (and number of rows in CoData). Defaults to 1/p, i.e. equal weight for each variable.
verbose	Logical. Asks whether algorithm progress should be printed. Defaults to FALSE.
ndpost	Number of posterior samples returned by dbarts after burn-in. Same as in dbarts. Defaults to 5000.
nskip	Number of burn-in samples. Same as in dbarts. Defaults to 5000.
nchain	Number of independent mcmc chains. Same as in dbarts. Defaults to 5.
keepevery	Thinning. Same as in dbarts. Defaults to 1.
ntree	Number of trees in the BART model. Same as in dbarts. Defaults to 50.
alpha	Alpha parameter of tree structure prior. Called base in dbarts. Defaults to 0.95. If EB_alpha is TRUE, this parameter will be the starting value.
beta	Beta parameter of tree structure prior. Called power in dbarts. Defaults to 2.
k	Parameter for leaf node parameter prior. Same as in dbarts. Defaults to 2. If EB_k is TRUE, this parameter will be the starting value.
sigest	Only for continuous response. Estimate of error variance used to set scaled inverse Chi^2 prior on error variance. Same as in dbarts. Defaults to 0.667*var(Y). #' If EB_sigma is TRUE, this parameter will be the starting value.
sigdf	Only for continuous response. Degrees of freedom for error variance prior. Same as in dbarts. Defaults to 10. If EB_sigma is TRUE, this parameter will be the starting value.
sigquant	Only for continuous response. Quantile at which sigest is placed Same as in dbarts. Defaults to 0.75. If EB_sigma is TRUE, this parameter will be fixed, only sigdf and sigest will be updated.

Value

An object with the estimated variable weights, i.e the probabilities that variables are selected in the splitting rules. Additionally, the final co-data model is returned. If EB is set to TRUE, estimates of k and/or alpha and/or (sigdf, sigest) are also returned. The returned object is of class S3 for which print(), summary(), and plot() are available. Function print() prints convergence details of the algorithm, summary() prints prior parameter estimates of EBcoBART, and plot() plots the estimated prior variable weights (including vertical line for equal variable weights).

The prior parameter estimates can then be used in your favorite BART R package that supports manually setting the splitting variable probability vector (dbarts and BARTMachine).

Author(s)

Jeroen M. Goedhart, <j.m.goedhart@amsterdamumc.nl>

References

dbarts

Jerome H. Friedman. "Multivariate Adaptive Regression Splines." The Annals of Statistics, 19(1) 1-67 March, 1991.

Hugh A. Chipman, Edward I. George, Robert E. McCulloch. "BART: Bayesian additive regression trees." The Annals of Applied Statistics, 4(1) 266-298 March 2010.

Jeroen M. Goedhart, Thomas Klausch, Jurriaan Janssen, Mark A. van de Wiel. "Co-data Learning for Bayesian Additive Regression Trees." arXiv preprint arXiv:2311.09997. 2023 Nov 16.

Examples

```
###################################
### Binary response example ######
###################################
# For continuous response example, see README.
# Use data set provided in R package
# We set EB = T indicating that we also estimate
# tree structure prior parameter alpha
# and leaf node prior parameter k

data("Lymphoma")
Xtr <- as.matrix(Lymphoma$Xtrain) # Xtr should be matrix object
Ytr <- Lymphoma$Ytrain
Xte <- as.matrix(Lymphoma$Xtest) # Xte should be matrix object
Yte <- Lymphoma$Ytest
CoDat <- Lymphoma$CoData
CoDat <- stats::model.matrix(~., CoDat) # encode grouping by dummies
#(include intercept)

set.seed(4) # for reproducible results
Fit <- EBcoBART(Y = Ytr, X = Xtr, CoData = CoDat,
                nIter = 2,         # Low! Only for illustration
                model = "binary",
                EB_k = TRUE, EB_alpha = TRUE,
```

```
                EB_sigma = FALSE,
                verbose = TRUE,
                ntree = 5,         # Low! Only for illustration
                nchain = 3,
                nskip = 500,       # Low! Only for illustration
                ndpost = 500,      # Low! Only for illustration
                Prob_Init = rep(1/ncol(Xtr), ncol(Xtr)),
                k = 2, alpha = .95, beta = 2)
EstProbs <- Fit$SplitProbs # estimated prior weights of variables
alpha_EB <- Fit$alpha_est
k_EB <- Fit$k_est
print(Fit)
summary(Fit)

# The prior parameter estimates EstProbs, alpha_EB,
# and k_EB can then be used in your favorite BART fitting package
# We use dbarts:

FinalFit <- dbarts::bart(x.train = Xtr, y.train = Ytr,
                         x.test = Xte,
                         ntree = 5,         # Low! Only for illustration
                         nchain = 3,        # Low! Only for illustration
                         nskip = 200,       # Low! Only for illustration
                         ndpost = 200,      # Low! Only for illustration
                         k = k_EB, base = alpha_EB, power = 2,
                         splitprobs = EstProbs,
                         combinechains = TRUE, verbose = FALSE)
```

Lymphoma *Lymphoma*

Description

Contains training data and test data to predict 2 year progression free survival (yes/no) based on four types of variables: copy number variation, point mutations, translocations, and clinical. For the variables, auxiliary information (co-data) is available which may be used to give more weight to certain variables in the prediction model. This data set is used in the manuscript "Co-data Learning for Bayesian Additive Regression Trees"

Usage

```
data(Lymphoma)
```

Format

A list object with five objects:

Xtrain Dataframe with 101 rows (samples) and 140 columns (variables). Explanatory variables used for fitting BART. Variable names are anonymized.

Ytrain Numeric of length 101. Binary training response (0: 2 year progression free survival, 1: disease came back within 2 years)

Xtest Dataframe with 83 rows (samples) and 140 columns (variables). Explanatory variables used for fitting BART. Variable names are anonymized.

Ytest Numeric of length 83 Binary training response (0: 2 year progression free survival, 1: disease came back within 2 years)

CoData Dataframe with 140 rows and 2 columns. Auxiliary information on the 140 variables. Contains a grouping structure indicating which type a variable is (copy number variation (CNV), mutation, translocation, or clinical), and p values (logit scale) for each variable obtained from a previous study

Author(s)

Jeroen M. Goedhart, <j.m.goedhart@amsterdamumc.nl>

Jurriaan Janssen

References

Jeroen M. Goedhart, Thomas Klausch, Jurriaan Janssen, Mark A. van de Wiel. "Co-data Learning for Bayesian Additive Regression Trees." arXiv preprint arXiv:2311.09997. 2023 Nov 16.

Index