Package ‘CERFIT’

October 12, 2022

Version 0.1.0
Title Causal Effect Random Forest of Interaction Tress
Description Fits a Causal Effect Random Forest of Interaction Tress (CERFIT) which is a modification of the Random Forest algorithm where each split is chosen to maximize subgroup treatment heterogeneity. Doing this allows it to estimate the individualized treatment effect for each observation in either randomized controlled trial (RCT) or observational data. For more information see X. Su, A. T. Peña, L. Liu, and R. A. Levine (2018) <doi:10.48550/arXiv.1709.04862>.
License GPL (>= 2)
Encoding UTF-8
LazyData true
RoxygenNote 7.1.2
LinkingTo Rcpp, RcppArmadillo
Imports partykit, CBPS, randomForest, twang, Rcpp, stats, glmnet
Depends R (>= 2.10)
NeedsCompilation yes
Author Justin Thorp [aut, cre], Luo Li [aut], Juanjuan Fan [aut]
Maintainer Justin Thorp <jjtthorp@gmail.com>
Repository CRAN
Date/Publication 2022-06-01 13:00:10 UTC

R topics documented:

CERFIT ................................................................. 2
educational ............................................................. 4
MinDepth ............................................................... 5
predict.CERFIT ....................................................... 6
warts ...................................................................... 7

Index 8
CERFIT

Fits a Random Forest of Interactions Trees

Description

Estimates an observations individualized treatment effect for RCT and observational data. Treatment can be an binary, categorical, ordered, or continuous variable. Currently if response is binary useRes must be set equal to TRUE.

Usage

CERFIT(
  formula,
  data,
 .ntrees,
  subset = NULL,
  search = c("exhaustive", "sss"),
  method = c("RCT", "observational"),
  PropForm = c("randomForest", "CBPS", "GBM", "HI"),
  split = c("t.test"),
  mtry = NULL,
  nsplit = NULL,
  nsplit.random = FALSE,
  minsplit = 20,
  minbucket = round(minsplit/3),
  maxdepth = 30,
  a = 50,
  sampleMethod = c("bootstrap", "subsample", "subsampleByID", "allData"),
  useRes = TRUE,
  scale.y = FALSE
)

Arguments

formula  Formula to build CERFIT. Categorical predictors must be listed as a factor. e.g., Y ~ x1 + x2 | treatment
.data    Data to grow a tree.
.ntrees  Number of Trees to grow
subset   A logical vector that controls what observations are used to grow the forest. The default value will use the entire dataframe
.search  Method to search through candidate splits
.method  For observational study data, method="observational"; for randomized study data, method="RCT".
.PropForm Method to estimate propensity score
.split   Impurity measure splitting statistic
**CERFIT**

- **mtry**: Number of variables to consider at each split
- **nsplit**: Number of cut points selected
- **nsplit.random**: Logical: indicates if process to select cut points are random
- **minsplit**: Number of observations required to continue growing tree
- **minbucket**: Number of observations required in each child node
- **maxdepth**: Maximum depth of tree
- **a**: Sigmoid approximation variable (for "sss" which is still under development)
- **sampleMethod**: Method to sample learning sample. Default is bootstrap. Subsample takes a subsample of the original data. SubsamplebyID samples by an ID column and uses all observations that have that ID. allData uses the entire data set for every tree.
- **useRes**: Logical indicator if you want to fit the CERFIT model to the residuals from a linear model
- **scale.y**: Logical, standardize y when creating splits (For "sss" to increase stability)

**Details**

This function implements Random Forest of Interaction Trees proposed in Su (2018). Which is a modification of the Random Forest algorithm where instead of a split being chosen to maximize prediction accuracy each split is chosen to maximized subgroup treatment heterogeneity. It chooses the best split by maximizing the test statistic for $H_0 : \beta_3 = 0$ in the following linear model

$$Y_i = \beta_0 + \beta_1 I(X_{ij} < c) + \beta_2 I(Z = 1) + \beta_3 I(X_{ij} < c)I(Z = 1) + \epsilon_i$$

Where $X_{ij}$ represents the splitting variable and $Z = 1$ represents treatment. So, by maximizing the test statistic for $\beta_3$ we are maximizing the treatment difference between the nodes.

The above equation only works when the data comes from a randomized controlled trial. But we can modify it to gives us unbiased estimates of treatment effect in observational studies Li et al. (2022). To do that we add propensity score into the linear model.

$$Y_i = \beta_0 + \beta_1 I(X_{ij} < c) + \beta_2 I(Z = 1) + \beta_3 I(X_{ij} < c)I(Z = 1) + \beta_4 \epsilon_i + \epsilon_i$$

Where $\epsilon_i$ represents the propensity score. The CERIT function will estimate propensity score automatically when the method argument is set to observational.

To control how this function estimates propensity score you can use the PropForm argument. Which can take four possible values randomForest, CBPS, GBM and HI. randomForest uses the randomForest package to use a random forest to estimate propensity score, CBPS uses Covariate balancing propensity score to estimate propensity score GBM uses generalized boosted regression models to estimate propensity score, and HI is for continuous treatment and estimates the general propensity score. Some of these options only work for certain treatment types. Full list below

- **binary**: GBM, CBPS, randomForest
- **categorical**: GBM, CBPS
- **ordered**: GBM, CBPS
- **continuous**: CBPS, HI
Value

Returns a fitted CERFIT object which is a list with the following elements

- RandFor: The Random forest of interaction trees
- trt.type: A string containing the treatment type of the data used to fit the model. Can be binary, multiple, ordered or continuous.
- response.type: A string representing the response type of the data. Can be binary or continuous.
- useRes: A logical indicator that is TRUE if the model was fit on the residuals of a linear model
- data: The data used to fit the model also contains the propensity score if method was set to observational

References


Examples

```r
fit <- CERFIT(Result_of_Treatment ~ sex + age + Number_of_Warts + Area + Time + Type | treatment, data = warts, ntrees = 30, method = "RCT", mtry = 2)
```

---

**educational**

*Observational Educational Dataset*

**Description**

A simulated dataset containing the grades and other attributes of 1000 simulated students

**Usage**

`educational`
MinDepth

Format
A data frame with 1000 rows and 7 variables:

SAT_MATH  SAT Math Score
HSGPA  High School GPA
AGE  Age of Student
GENDER  Gender of Student
URM  Under Represented Minority
A  Treatment Variable
Y  Students Final Grade

Source

MinDepth

Calculate Variable Importance

Description
Calculates the average minimal depth of each predictor used to fit a CERFIT object. It calculates Variables importance by using a Variables average minimal depth. variable’s with a lower average minimal depth are more important.

Usage
MinDepth(cerfit)

Arguments
cerfit  A fitted CERFIT object

Details
The depth of the root node is zero and if a variable does not appear at any split in a tree it is assigned maxdepth + 1 for that tree.

Value
Returns a named vector with the name of each predictor used to fit the CERFIT object and its corresponding average minimal depth across all trees
Examples

fit <- CERFIT(Result_of_Treatment ~ sex + age + Number_of_Warts + Area + Time + Type | treatment,
data = warts,
ntrees = 30,
method = "RCT",
mtry = 2)
importance <- MinDepth(fit)

predict.CERFIT

Get predictions from a CERFIT object

Description

Get predictions from a CERFIT object

Usage

## S3 method for class 'CERFIT'
predict(
  object,
  newdata = NULL,
  gridval = NULL,
  prediction = c("overall", "by iter"),
type = c("response", "ITE", "node", "opT"),
  alpha = 0.5,
  ...
)

Arguments

object A fitted CERFIT object
newdata New data to make predictions from. IF not provided will make predictions on training data
gridval For continuous treatment. Controls for what values of treatment to predict
prediction Return prediction using all trees ("overall") or using first i trees ("by iter")
type Choose what value you wish to predict. Response will predict the response. ITE will predict the Individualized treatment effect. Node will predict the node. And opT will predict the optimal treatment for each observation.
alpha For continuous treatment it is the mixing parameter for the elastic net regularization in each node. When equal to 0 it is ridge regression and when equal to 1 it is lasso regression.
... Additional Arguments
Value

The return value depends on the type argument. If type is response, the function will return a matrix with \( n \) rows and the number of columns equal to the level of treatment. If type is ITE, then it returns a matrix with \( n \) rows and a number of columns equal to one minus the levels of treatment. And if type is opT then it returns a matrix with \( n \) rows and two columns. With the first column denoting the optimal treatment and the second column denoting the optimal response.

Examples

```r
fit <- CERFIT(Result_of_Treatment ~ sex + age + Number_of_Warts + Area + Time + Type | treatment,
data = warts,
ntrees = 30,
method = "RCT",
mtry = 2)
ite <- predict(fit,type = "ITE")
```

---

**warts**

*Randomized Controlled Trial Warts Dataset*

Description

A dataset comparing immunotherapy to cryotherapy treatments and their effectiveness of removing warts.

Usage

`warts`

Format

A data frame with 180 rows and 8 variables:

- **sex** Patients Sex
- **age** Patients Age
- **Time** Time Elapsed Before Treatment
- **Number_of_Warts** Number of Warts
- **Type** Type of Wart
- **Area** Wart Surface Area
- **Result_of_Treatment** Treatment Outcome
- **treatment** 0 for immunotherapy and 1 for cryotherapy

Source

Index

* datasets
  educational, 4
  warts, 7
CERFIT, 2
educational, 4
MinDepth, 5
predict.CERFIT, 6
warts, 7