bart.R

Eliyahu

2023-05-02

```
# load packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(BART)
```

```
## Warning: package 'BART' was built under R version 4.2.3
```

```
## Loading required package: nlme
##
## Attaching package: 'nlme'
##
## The following object is masked from 'package:dplyr':
##
##     collapse
##
## Loading required package: nnet
## Loading required package: survival
```

```
library(here)
```

```
## here() starts at C:/Users/Eliyahu/OneDrive - Johns Hopkins/theory_stat2/project
```

```
# set seed
set.seed(1)

# load data
diabetes <- read_csv(here("data", "diabetes_binary_health_indicators_BRFSS2015.csv"))
```

```
## Rows: 253680 Columns: 22
## -- Column specification --------------------------------------------------------
```

```
## Delimiter: ","
## dbl (22): Diabetes_binary, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# this data is too big for BART to be efficient, so we subset 5000 rows
diabetes <- diabetes %>% slice_sample(n= 5000)

# set the feature and target sets
x <- diabetes %>%
    select(-c(Diabetes_binary)) %>%
    as.data.frame()
y <- diabetes %>%
    select(Diabetes_binary)%>%
    deframe()

# set the smoker variable
x_treat1 <- x %>%
    mutate(Smoker = 1)

x_treat0 <- x %>%
    mutate(Smoker = 0)

# BART fit
bartfit <- gbart(x.train = x, y.train = y, type = 'pbart', ntree = 50,printevery = 1000)
```

```
## *****Calling gbart: type=2
## *****Data:
## data:n,p,np: 5000, 21, 0
## y1,yn: 0.000000, 0.000000
## x1,x[n*p]: 1.000000, 3.000000
## *****Number of Trees: 50
## *****Number of Cut Points: 1 ... 7
## *****burn,nd,thin: 100,10000,10
## *****Prior:beta,alpha,tau,nu,lambda,offset: 2,0.95,0.212132,3,1,-1.09664
## *****Dirichlet:sparse,theta,omega,a,b,rho,augment: 0,0,1,0.5,1,21,0
## *****printevery: 1000
##
## MCMC
## done 0 (out of 10100)
## done 1000 (out of 10100)
## done 2000 (out of 10100)
## done 3000 (out of 10100)
## done 4000 (out of 10100)
## done 5000 (out of 10100)
## done 6000 (out of 10100)
## done 7000 (out of 10100)
## done 8000 (out of 10100)
## done 9000 (out of 10100)
## done 10000 (out of 10100)
## time: 98s
## trcnt,tecnt: 1000,0
```

```r
# predictions
predict1 <- predict(bartfit, x_treat1)
```

```
## *****In main of C++ for bart prediction
## tc (threadcount): 1
## number of bart draws: 1000
## number of trees in bart sum: 50
## number of x columns: 21
## from x,np,p: 21, 5000
## ***using serial code
```

```r
predict0 <- predict(bartfit, x_treat0)
```

```
## *****In main of C++ for bart prediction
## tc (threadcount): 1
## number of bart draws: 1000
## number of trees in bart sum: 50
## number of x columns: 21
## from x,np,p: 21, 5000
## ***using serial code
```

```r
# ate estimation
e_ate <- mean(predict1$prob.test.mean - predict0$prob.test.mean)
print(e_ate)
```

```
## [1] 0.01480413
```