Illustrative applicationEstimating Optimal Decision Trees for Treatment Assignment: The Case of K > 2 Treatment Alternatives

This is a notebook to reproduce the illustrative application of the paper “Optimal Decision Trees for Treatment Assignment: The Case of K > 2 Treatment Alternatives”.

# the R-packages and the source file below are used;  
# make sure to first install these R-packages  
library(devtools)  
library(quint)  
library(partykit)  
library(ranger)  
library(combinat)  
library(styler)  
source("https://raw.githubusercontent.com/KULeuven-PPW-OKPIV/multivalued/main/Multivalued%20code%20snippet.R")

For this application, the BCRP data set (available in the quint r-package) is used. We start by loading this data.

dt <- na.omit(bcrp)  
dt$wcht1 <- factor(dt$wcht1)  
dt$nationality <- factor(dt$nationality)  
dt$marital <- factor(dt$marital)  
dt$cond <- factor(dt$cond)  
head(dt)

## physt1 cesdt1 physt3 cesdt3 negsoct1 uncomt1 disopt1 comorbid age  
## 1 37.65374 14 52.62905 4 9 28 14 6 29.48392  
## 2 53.64822 10 51.18797 14 7 36 10 2 44.66256  
## 3 63.84140 8 66.45392 9 6 29 15 1 43.09925  
## 4 38.72757 2 45.99656 4 5 30 17 13 46.93498  
## 5 55.85700 0 60.66603 1 10 32 22 1 41.09514  
## 6 53.06922 0 54.83928 0 8 24 21 3 44.44627  
## wcht1 nationality marital trext cond  
## 1 1 1 0 0.2589759 3  
## 2 1 1 1 0.5557208 1  
## 3 1 1 0 0.2589759 2  
## 4 1 1 1 0.5557208 2  
## 5 0 1 1 0.5557208 2  
## 6 1 1 1 0.2589759 3

We first use the AIPWE as an estimator of the misclassification costs. To do so, we need an outcome- and propensity model. For the propensity model, we first use empirical proportions. The outcome model (with as response the difference in physical functioning between 9 months follow-up and baseline), includes main effects of all covariates, and their interaction with treatment.

# Outcome model  
model1 <- lm(I(physt3 - physt1) ~ cond \* (1 + cesdt1 + physt1 + negsoct1 + uncomt1 + disopt1 + comorbid + age + wcht1 + nationality + marital + trext), data = dt)  
  
## make a prediction using the model for each observation and each treatment alternative.  
data1 <- dt  
N <- nrow(dt)  
data1[, "cond"] <- as.factor(rep(1, N))  
u1 <- predict.glm(model1, as.data.frame(data1))  
data2 <- dt  
data2[, "cond"] <- as.factor(rep(2, N))  
u2 <- predict.glm(model1, as.data.frame(data2))  
data3 <- dt  
data3[, "cond"] <- as.factor(rep(3, N))  
u3 <- predict.glm(model1, as.data.frame(data3))  
  
# Propensity (empirical proportions)  
PRaipwe <- table(dt$cond) / N  
  
# AIPWE estimates  
cond <- dt$cond  
Y <- dt$physt3 - dt$physt1  
c1 <- (Y \* (cond == 1) / PRaipwe[1]) - (((cond == 1) - PRaipwe[1]) / PRaipwe[1] \* u1)  
c2 <- (Y \* (cond == 2) / PRaipwe[2]) - (((cond == 2) - PRaipwe[2]) / PRaipwe[2] \* u2)  
c3 <- (Y \* (cond == 3) / PRaipwe[3]) - (((cond == 3) - PRaipwe[3]) / PRaipwe[3] \* u3)  
est.aipwe <- cbind(c1, c2, c3)  
  
# optimal treatment (maximum aipwe estimate)  
Zaipwe <- sapply(1:nrow(dt), function(b) which.max(est.aipwe[b, ]))  
dataCaipwe <- cbind(as.factor(Zaipwe), dt[, c(1, 2, 5:14)])  
  
# individual loss matrix (for each observation and each treatment alternative, the cost compared to assignment to the optimal treatment)  
indiv.loss <- t(sapply(1:nrow(dataCaipwe), function(b) max(est.aipwe[b, ]) - est.aipwe[b, ]))  
  
print(head(indiv.loss))

## c1 c2 c3  
## [1,] 17.543821 20.06833 0.000000  
## [2,] 17.209547 0.00000 6.178263  
## [3,] 19.907418 0.00000 18.150354  
## [4,] 6.269371 0.00000 16.502371  
## [5,] 12.729999 0.00000 10.830230  
## [6,] 3.114457 0.00000 4.766833

Next, we estimate the optimal treatment regime using the estimated misclassification costs (with AIPWE), by means of the user defined splitting functions in rpart. We use five different seeds for pruning.

## estimate optimal treatment regime  
fit <- rpart(  
 as.numeric(Zaipwe) ~ cesdt1 + physt1 + negsoct1 + uncomt1 + disopt1 +  
 comorbid + age + wcht1 + nationality + marital + trext,  
 data = dataCaipwe, weights = 1:nrow(dataCaipwe), method = multirisk,  
 ## multirisk is the user defined splitting function  
 parms = list(indiv.loss = t(sapply(1:nrow(dataCaipwe), function(b) {  
 max(est.aipwe[b, ]) - est.aipwe[b, ]  
 })))  
) ## use individual loss matrix estimated using AIPWE  
  
  
# Pruning and printing based on five different seeds  
trees\_aipwe <- list(c(0))  
for (i in 1:5) {  
 set.seed(111 + i)  
 xfit <- xpred.rpart(fit, xval = 20)  
 xerror <- colMeans(sapply(1:ncol(xfit), function(m) {  
 sapply(  
 1:nrow(xfit),  
 function(k) indiv.loss[k, xfit[k, m]]  
 )  
 }))  
 xerror  
 cp1 <- fit$cptable[which.min(xerror), 1]  
 fitpr <- prune(fit, cp = cp1)  
 trees\_aipwe[[i]] <- fitpr  
 print(trees\_aipwe[[i]])  
}

## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1369.418 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1369.418 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1369.41800 2   
## 2) comorbid< 4.5 173 864.54230 2   
## 4) age< 34.70226 11 59.69881 3 \*  
## 5) age>=34.70226 162 757.15290 2 \*  
## 3) comorbid>=4.5 51 328.97180 1   
## 6) comorbid< 6.5 28 139.22480 3 \*  
## 7) comorbid>=6.5 23 97.49449 1 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1369.4180 2   
## 2) comorbid< 4.5 173 864.5423 2 \*  
## 3) comorbid>=4.5 51 328.9718 1 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1369.418 2 \*

Next, we use the theoretical proportions of 1/3 per treatment alternative as the propensity model and repeat the same procedure.

# Propensity (theoretical probabilities)  
PRaipwe <- c(1 / 3, 1 / 3, 1 / 3)  
  
# AIPWE estimates  
c1 <- (Y \* (cond == 1) / PRaipwe[1]) - (((cond == 1) - PRaipwe[1]) / PRaipwe[1] \* u1)  
c2 <- (Y \* (cond == 2) / PRaipwe[2]) - (((cond == 2) - PRaipwe[2]) / PRaipwe[2] \* u2)  
c3 <- (Y \* (cond == 3) / PRaipwe[3]) - (((cond == 3) - PRaipwe[3]) / PRaipwe[3] \* u3)  
est.aipwe <- cbind(c1, c2, c3)  
  
# optimal treatment (maximum aipwe estimate)  
Zaipwe <- sapply(1:nrow(dt), function(b) which.max(est.aipwe[b, ]))  
dataCaipwe <- cbind(as.factor(Zaipwe), dt[, c(1, 2, 5:14)])  
  
# individual loss matrix (for each observation and each treatment alternative, the cost compared to assignment to the optimal treatment)  
indiv.loss <- t(sapply(1:nrow(dataCaipwe), function(b) max(est.aipwe[b, ]) - est.aipwe[b, ]))  
  
## estimate optimal treatment regime  
fit <- rpart(  
 as.numeric(Zaipwe) ~ cesdt1 + physt1 + negsoct1 + uncomt1 + disopt1 +  
 comorbid + age + wcht1 + nationality + marital + trext,  
 data = dataCaipwe, weights = 1:nrow(dataCaipwe), method = multirisk,  
 ## multirisk is the user defined splitting function  
 parms = list(indiv.loss = t(sapply(1:nrow(dataCaipwe), function(b) {  
 max(est.aipwe[b, ]) - est.aipwe[b, ]  
 })))  
) ## use individual loss matrix estimated using AIPWE  
  
  
# pruning and printing using five different seeds  
trees\_aipwe <- list(c(0))  
for (i in 1:5) {  
 set.seed(111 + i)  
 xfit <- xpred.rpart(fit, xval = 20)  
 xerror <- colMeans(sapply(1:ncol(xfit), function(m) {  
 sapply(  
 1:nrow(xfit),  
 function(k) indiv.loss[k, xfit[k, m]]  
 )  
 }))  
 xerror  
 cp1 <- fit$cptable[which.min(xerror), 1]  
 fitpr <- prune(fit, cp = cp1)  
 trees\_aipwe[[i]] <- fitpr  
 print(trees\_aipwe[[i]])  
}

## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1368.68000 2   
## 2) comorbid< 4.5 173 865.50890 2 \*  
## 3) comorbid>=4.5 51 330.43990 1   
## 6) comorbid< 6.5 28 139.14150 3 \*  
## 7) comorbid>=6.5 23 94.90943 1 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1368.68 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1368.6800 2   
## 2) comorbid< 4.5 173 865.5089 2 \*  
## 3) comorbid>=4.5 51 330.4399 1 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1368.68 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 1368.68 2 \*

Finally, we use a Random Forest to estimate the misclassification costs. We use 3 different seeds for estimating the random forest-based misclassification costs. For estimating the optimal treatment regimes, we again use 5 seeds for pruning.

trees\_rf <- list(c(0))  
N <- nrow(dt)  
resp <- dt$physt3 - dt$physt1  
  
# Use three different seeds for random forest  
k <- 1  
for (i in 1:3) {  
 set.seed(1 + i)  
 RF <- ranger(x = dt[, c(1, 2, 5:14)], y = resp, always.split.variables = "cond")  
 data1 <- dt[, c(1, 2, 5:14)]  
 data1[, "cond"] <- as.factor(rep(1, N))  
 data2 <- dt[, c(1, 2, 5:14)]  
 data2[, "cond"] <- as.factor(rep(2, N))  
 data3 <- dt[, c(1, 2, 5:14)]  
 data3[, "cond"] <- as.factor(rep(3, N))  
 total <- data.frame(rbind(data1, data2, data3))  
 preds <- predict(RF, total)  
 est.rf <- as.matrix(data.frame(  
 u1 = preds$predictions[1:224],  
 u2 = preds$predictions[225:448],  
 u3 = preds$predictions[449:672]  
 ))  
  
 Zaipwe <- sapply(1:nrow(dt), function(b) which.max(est.rf[b, ]))  
 dataCaipwe <- cbind(as.factor(Zaipwe), dt[, c(1, 2, 5:14)])  
  
 fit <- rpart(as.numeric(Zaipwe) ~ cesdt1 + physt1 + negsoct1 + uncomt1 + disopt1 + comorbid + age + wcht1 + nationality + marital + trext,  
 data = dataCaipwe, weights = 1:nrow(dataCaipwe), method = multirisk,  
 parms = list(indiv.loss = t(sapply(  
 1:nrow(dataCaipwe),  
 function(b) max(est.rf[b, ]) - est.rf[b, ]  
 )))  
 )  
  
 indiv.loss <- t(sapply(1:nrow(dataCaipwe), function(b) {  
 max(est.rf[b, ]) - est.rf[b, ]  
 }))  
  
 # pruning and printing using five different seeds  
 for (j in 1:5) {  
 set.seed(1 + j)  
 xfit <- xpred.rpart(fit, xval = 20)  
 xerror <- colMeans(sapply(1:ncol(xfit), function(m) {  
 sapply(1:nrow(xfit), FUN = function(k) {  
 indiv.loss[k, xfit[k, m]]  
 })  
 }))  
 xerror  
 cp1 <- fit$cptable[which.min(xerror), 1]  
 fitpr <- prune(fit, cp = cp1)  
 print(fitpr)  
 trees\_rf[[k]] <- fitpr  
 k <- k + 1  
 }  
}

## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 47.20286 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 47.2028600 2   
## 2) age< 50.73101 216 41.1973200 2   
## 4) comorbid< 6.5 195 36.2891100 2 \*  
## 5) comorbid>=6.5 21 4.6783790 1   
## 10) physt1< 43.38614 7 0.6824365 2 \*  
## 11) physt1>=43.38614 14 0.3194469 1 \*  
## 3) age>=50.73101 8 2.9069860 3 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 47.2028600 2   
## 2) age< 50.73101 216 41.1973200 2   
## 4) comorbid< 6.5 195 36.2891100 2 \*  
## 5) comorbid>=6.5 21 4.6783790 1   
## 10) physt1< 43.38614 7 0.6824365 2 \*  
## 11) physt1>=43.38614 14 0.3194469 1 \*  
## 3) age>=50.73101 8 2.9069860 3 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 47.20286 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 47.2028600 2   
## 2) age< 50.73101 216 41.1973200 2   
## 4) comorbid< 6.5 195 36.2891100 2 \*  
## 5) comorbid>=6.5 21 4.6783790 1   
## 10) physt1< 43.38614 7 0.6824365 2 \*  
## 11) physt1>=43.38614 14 0.3194469 1 \*  
## 3) age>=50.73101 8 2.9069860 3 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 49.61104 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 49.61104 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 49.61104 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 49.61104 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 49.61104 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 51.75149 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 51.75149 2 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 51.751490 2   
## 2) age< 50.73101 216 45.302770 2 \*  
## 3) age>=50.73101 8 2.403449 3 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 51.7514900 2   
## 2) age< 50.73101 216 45.3027700 2   
## 4) comorbid< 6.5 195 41.1383300 2 \*  
## 5) comorbid>=6.5 21 3.4938550 1   
## 10) physt1< 43.38614 7 0.1994531 2 \*  
## 11) physt1>=43.38614 14 0.2060276 1 \*  
## 3) age>=50.73101 8 2.4034490 3 \*  
## n= 224   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 224 51.7514900 2   
## 2) age< 50.73101 216 45.3027700 2   
## 4) comorbid< 6.5 195 41.1383300 2 \*  
## 5) comorbid>=6.5 21 3.4938550 1   
## 10) physt1< 43.38614 7 0.1994531 2 \*  
## 11) physt1>=43.38614 14 0.2060276 1 \*  
## 3) age>=50.73101 8 2.4034490 3 \*