Homework: Ensemble Learning

This homework sheet will test your knowledge of ensemble learning using R.

0

a) Load the file `Heart.csv` into R as follows. This dataset provides information on the risk factors for heart disease.

```
heart <- read.csv("Heart.csv", header=TRUE, sep=",")

# Remove NA observations
heart <- na.omit(heart)

# Remove obsolete variable "X"
heart$X <- NULL

# Convert column "AHD" into factor
levels(heart$AHD)[1] <- 0
levels(heart$AHD)[2] <- 1
heart$AHD <- factor(heart$AHD, labels=c("No", "Yes"))

# Convert column "Sex" into factor
heart$Sex <- as.factor(heart$Sex)
heart$Sex <- factor(heart$Sex, labels=c("female", "male"))

# Number of observations
nrow(heart)

## [1] 297

# List of column names
colnames(heart)

##  [1] "Age"       "Sex"       "ChestPain" "RestBP"    "Chol"
##  [6] "Fbs"       "RestECG"   "MaxHR"     "ExAng"     "Oldpeak"
## [11] "Slope"     "Ca"        "Thal"      "AHD"

# View first rows of the dataset
head(heart)

##   Age    Sex    ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak
## 1  63   male      typical    145  233   1       2   150     0     2.3
## 2  67   male asymptomatic    160  286   0       2   108     1     1.5
## 3  67   male asymptomatic    120  229   0       2   129     1     2.6
## 4  37   male   nonanginal    130  250   0       0   187     0     3.5
## 5  41 female   nontypical    130  204   0       2   172     0     1.4
## 6  56   male   nontypical    120  236   0       0   178     0     0.8
##   Slope Ca       Thal AHD
## 1     3  0      fixed  No
## 2     2  3     normal Yes
## 3     2  2 reversable Yes
## 4     3  0     normal  No
## 5     1  0     normal  No
## 6     1  0     normal  No
```

The columns are as follows:

Column Name	Interpretation
Age	Age
Sex	Sex (1 = male; 0 = female)
ChestPain	Chest pain (typical, asymptotic, nonanginal, nontypical)
RestBP	Resting blood pressure
Chol	Serum cholestoral in mg/dl
Fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
RestECG	Resting electrocardiographic results
MaxHR	Maximum heart rate achieved
ExAng	Exercise induced angina (1 = yes; 0 = no)
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
Ca	Number of major vessels colored by flourosopy (0 - 3)
Thal	(3 = normal; 6 = fixed defect; 7 = reversable defect)
AHD	Diagnosis of heart disease (1 = yes; 0 = no)

b) We are interested in what determines heart disease (`AHD`). Therefore, split the data into a test (20 %) and training (80 %) set. Learn a decision tree on the training set. 4

c) Learn a random forest on the training set. Set the number of trees to 50 and the number of variables to choose from at each node to 2. Store the variable importance metric for all variables as we need these in the subsequent exercise. 5

d) Plot and interpret the estimated mean squared errors when varying the number of trees. 6

e) Plot the five most important variables. 1

f) For both the decision tree and the random forest, predict `AHD` for the test set. Compare the performances by looking at the ROC curve. 8

g) Fit a model using `gamboost`. This time, only consider the variables `Thal`, `Sex`, `ChestPain` and `Chol`, whereby the latter has a smoothing effect. Set the number of iterations to 100. 4

h) Refine the previous model by finding the optimal number of iterations with help of cross-validation. 2

i) Plot and interpret the partial effects for the previous subset of variables. 8

j) Now fit a linear model and another one using `glmboost` with the previous subset of variables. For the latter, set the number of iterations to 100. 3

k) Calculate and compare the following three measures for the `gam`, the `glm` and the `lm` model: 8

1. Akaike Information Criterion

2. Mean squared error of prediction for the test set
3. Prediction accuracy for the test set

l) Fit a model using `adaBoost` to the training set. Set the number of iterations to 100. 2

m) Update the ada object from the previous exercise to have additional testing errors. Print the final confusion matrix and plot MSEs, as well as the variable importance. 3

n) Which of the previous models has the highest accuracy on the test set? Compare the following approaches: decision tree, random forest, boosted GAM, boosted GLM, linear model, AdaBoost. 4