Fourth homework assignment

Professional Master in Economics

Hedibert Freitas Lopes

Bayesian Learning

Due date: 7:30pm, June 29th, 2021

Prepare one (and only one) PDF file with your solutions and send the file to Igor Martins (`igorfbm@al.insper.edu.br`)
Assignments will be delivered in pairs

Bayesian learning for binary regression models. For this last homework assignment, you will implement the following models to predicting *tomorrow's rain* in Australia (as well as comparing in-sample and out-of-sample fit): GLM, Bayes GLM, GLMNET (ridge and lasso penalties), CART, random forest and BART. The data can be downloaded from Kaggle. I provide below a short R script for reading and cleaning the data as well as building the response vector y and the matrix of potential predictors X. Your job is to return our best model. Please, do provided detailed explanation throughout our derivations[1].

```
# Reading the data
data = read.csv("weatherAUS.csv")

# Selecting only a handful of regions (to avoid slowing down computation!)
ind = data[,2]=="Melbourne" | data[,2]=="Sydney" | data[,2]=="Brisbane" | data[,2]=="Perth" |
      data[,2]=="Adelaide" | data[,2]=="Canberra" | data[,2]=="Darwin" | data[,2]=="Robart"

data = data[ind,]
attach(data)
n = nrow(data)

# The response variable
y = rep(0,n)
y[RainTomorrow=="Yes"]=1

# A bunch of continuous regressors
par(mfrow=c(2,6))
boxplot(WindSpeed9am~RainTomorrow,outline=FALSE)
boxplot(WindSpeed3pm~RainTomorrow,outline=FALSE)
boxplot(Humidity9am~RainTomorrow,outline=FALSE)
boxplot(Humidity3pm~RainTomorrow,outline=FALSE)
boxplot(Pressure9am~RainTomorrow,outline=FALSE)
boxplot(Pressure3pm~RainTomorrow,outline=FALSE)
boxplot(MinTemp~RainTomorrow,outline=FALSE)
boxplot(MaxTemp~RainTomorrow,outline=FALSE)
boxplot(Rainfall~RainTomorrow,outline=FALSE)
boxplot(Evaporation~RainTomorrow,outline=FALSE)
boxplot(Sunshine~RainTomorrow,outline=FALSE)
boxplot(WindGustSpeed~RainTomorrow,outline=FALSE)

X = cbind(WindSpeed9am,WindSpeed3pm,Humidity9am,Humidity3pm,Pressure9am,Pressure3pm,
          MinTemp,MaxTemp,Rainfall,Evaporation,Sunshine,WindGustSpeed)
n = nrow(X)


# Handling (read: excluding!)  rows with missing values
missobs = rep(0,n)
for (i in 1:n)
  missobs[i] = sum(is.na(X[i,]))
y = y[missobs==0]
X = X[missobs==0,]
n = nrow(X)
p = ncol(X)
for (i in 1:p)
  X[,i] = scale(X[,i])
```

---

[1] For hints, please check the worked ICU example in our course webpage.