Exercise week 6

Gherardo Varando, gherardo.varando@math.ku.dk

06/01/2020

CORIS data

Load the data with

```
coris <- read.table("coris.dat", skip = 4, sep = ",",
col.names = c("row.names", "sbp", "tobacco",
"ldl", "adiposity",
"famhist", "typea", "obesity",
"alcohol",
"age", "chd"))[,-1]
```

The goal is to predict the presence of coronary heart disease (`chds`).

Exercise 1

Ex 1.1 Fit the full logistic regression model using all the predictors. Obtain estimations of the accuracy (that is, the percentage of correctly classified) using both leave-one-out and 10-fold cross validation.

Use backward stepwise selection for logistic regression, using AIC to obtain a simpler model as we did in exercises of week 5, estimate the accuracy using both leave-one-out and 10-fold cross validation.

Compare now the two models, which one is to prefer based on the accuracy estimations? To be able to compare the two accuracy estimations you should use the same groups in the 10-fold cross-validation.

Ex 1.2 Perform stepwise forward selection using accuracy estimated with 5-fold cross-validation to score the candidate models. That is similarly to the stepwise forward selection with AIC, we start from the logistic model using just the intercept and we try to add the possible predictors, for each candidate model we obtain the accuracy estimated with 5-fold cross-validation and we move to the model that obtain the best increase in accuracy. We the repeat after no increase in the estimated accuracy is possible. We return the final model selected and the estimation of the accuracy using 5-fold cross-validation.

The wine quality dataset

We load both red and white wine datasets and we transform the quality index to a binary good-bad variable.

```
wines_red <- read.csv("winequality-red.csv", sep =";")
wines_white <- read.csv("winequality-white.csv", sep =";")

good <- wines_red$quality > 5
wines_red$quality <- "bad"
wines_red[good, "quality"] <- "good"
wines_red[,"quality"] <- as.factor(wines_red[, "quality"])
good <- wines_white$quality > 5
wines_white$quality <- "bad"
wines_white[good, "quality"] <- "good"
wines_white[,"quality"] <- as.factor(wines_white[, "quality"])
```

Exercise 2

The goal is to predict the quality of wines from the other variables in the dataset

Ex 1.1 Fit a logistic regression models using all the predictors and the data for the red wines. Compute the accuracy of the model on the red wines and on the white wines.

Ex 2.2 Fit a logistic regression model using the white wines data and compute the accuracy over the white and red wines.

Ex 2.3 Do you think the model fitted using the red wines data is useful for the white wines ? Compare the information from `summary` from the two models, in particular the relevant covariates of the two models are the same?

Exercise 3

Ex 3.1 Use now both red and white wines data and perform stepwise forward selection with AIC to select a logistic regression model for the binary quality variable.

Ex 3.2 Estimate the accuracy of the model using 10-fold cross validation on the red and white wines data.

Ex 3.3 We want now to estimate the accuracy of the model trained both on red and white wines only over the red wines, to do so just randomly select some (200) red wines observations and train the model on the remaining red and white wines. Then test the model over the 200 red wines selected and compute

the accuracy. Repeat the process a number of times (10, 50, 100) and average the obtained accuracy.