

Machine Learning - Lab 4

Statistical learning

Souhaib BEN TAIEB

9 March 2020

Note: We use the acronym ISLR for “An Introduction to Statistical Learning with Applications in R”.

Exercise 1

Do the exercise 1 in chapter 2.4 of ISLR.

Exercise 2

Do the exercise 2 in chapter 2.4 of ISLR.

Exercise 3

Do the exercise 3 in chapter 2.4 of ISLR.

Exercise 4

Do the exercise 5 in chapter 2.4 of ISLR.

Exercise 5

Do the exercise 7 in chapter 2.4 of ISLR. See page 39 in ISLR for the K-Nearest Neighbors.

Exercise 6

Do some exploratory data analysis on the **Wage** data set (available in the ISLR package).

- Tabulate education and marital status
- Tabulate education and race
- Tabulate marital status race
- Plot marital status as a function of age
- Try other combinations

Exercise 7

- Fit a spline curve to the relationship between wage and age using two degrees of freedom (**df=2**).
- Experiment with different values of **df** (degrees of freedom)
- Select one that you think is about right.

Exercise 8

Now we will test which value of **df** minimizes the MSE on some test data.

First, we randomly split the **Wage** data set into training and test sets, with 2000 observations in the training data and the remaining 1000 observations in the test data.

```
library(ISLR)
idx <- sample(1:nrow(Wage), size=2000)
train <- Wage[idx,]
test <- Wage[-idx,]
```

- Using a loop, compute the training and test MSE for $df = 1, 2, \dots, 20$, and store it in two vectors `trainingMSE` and `testMSE`.
- Plot both `trainingMSE` and `testMSE` as a function of `df`.
- Which value of `df` gives the minimum training MSE?
- Which value of `df` gives the minimum test MSE?
- Plot a vertical line at your “guessed” value of `df`. How close is it to the optimal?
- Do you get the same results if you repeat the exercise on different splits of training and test data? Why?

Exercise 9

- Repeat the previous analysis, but use the full linear model including the other variables in the data set.
- How much better is the test MSE once you include the other predictor variables?
- Check your model by plotting the residuals as a function of each predictor variable. Do you see anything unusual in the residual plots?