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.

(a) better performance

(b) worse performance

(c) better performance

(d) worse performance

Exercise 2

Do the exercise 2 in chapter 2.4 of ISLR.

(a) regression and inference
(b) classification and prediction
(c) regression and prediction

Exercise 3

Do the exercise 3 in chapter 2.4 of ISLR.

See Figures 2.9, 2.10, 2.11 and 2.12 in ISLR.

Exercise 4

Do the exercise 5 in chapter 2.4 of ISLR.

Very flexible methods provide a better fit (with a lower bias), but can overfit the data and have a larger variance.

Less flexible methods typically have a small variance but a high bias.

Which one to choose between a more flexible or a less flexible approach? This depends on the underlying data generating process. If the true underlying function to estimate is linear for example, then a less flexible approach would be more appropriate. However, if it is highly nonlinear, then a more flexible approach would be needed.

Exercise 5

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

(a) $32\sqrt{10}\sqrt{5}\sqrt{2}\sqrt{3}$

(b) Green. Observation #5 is the closest neighbor for K = 1.
(c) Red. Observations #2, #5, #6 are the closest neighbors for K = 3. 2 is Red, 5 is Green, and 6 is Red.
(d) Small. A small K would be flexible for a non-linear decision boundary, whereas a large K would try to fit a more linear boundary because it takes more points into consideration.

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

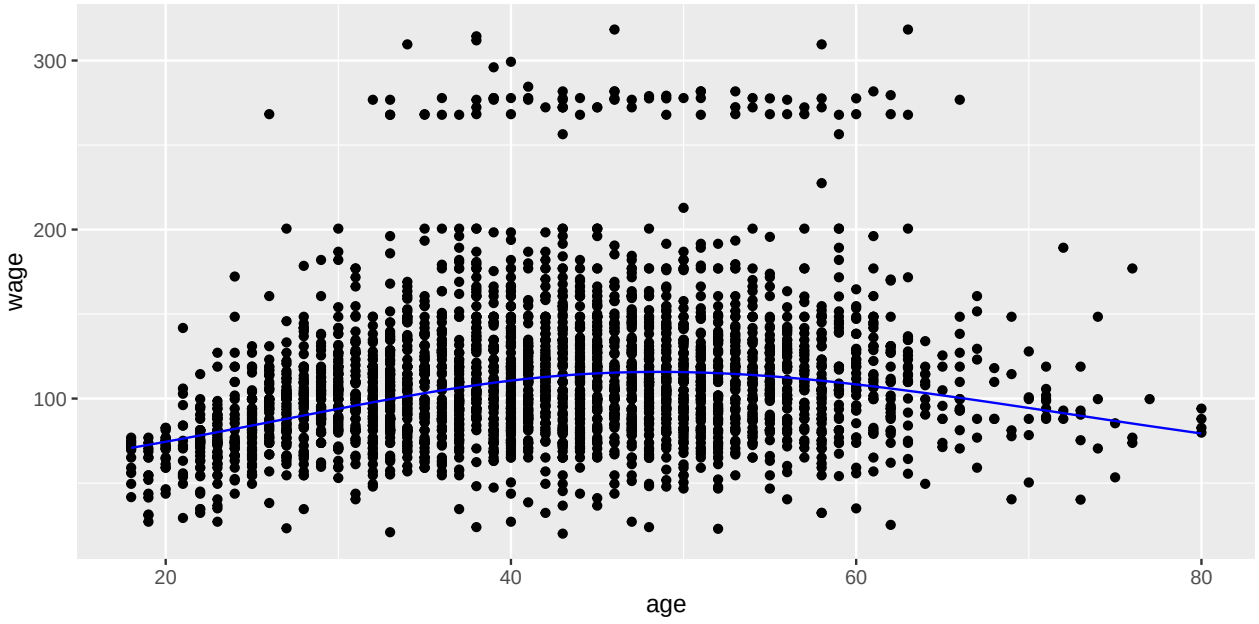
Exercise 7

- 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.

```
library(ISLR)
library(splines)
library(ggplot2)
p <- qplot(age, wage, data=Wage)

fit <- lm(log(wage) ~ ns(age, df=2), data=Wage)
Wage$fc <- exp(fitted(fit))

p + geom_line(aes(age, fc), data=Wage, col='blue')
```



Exercise 8

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

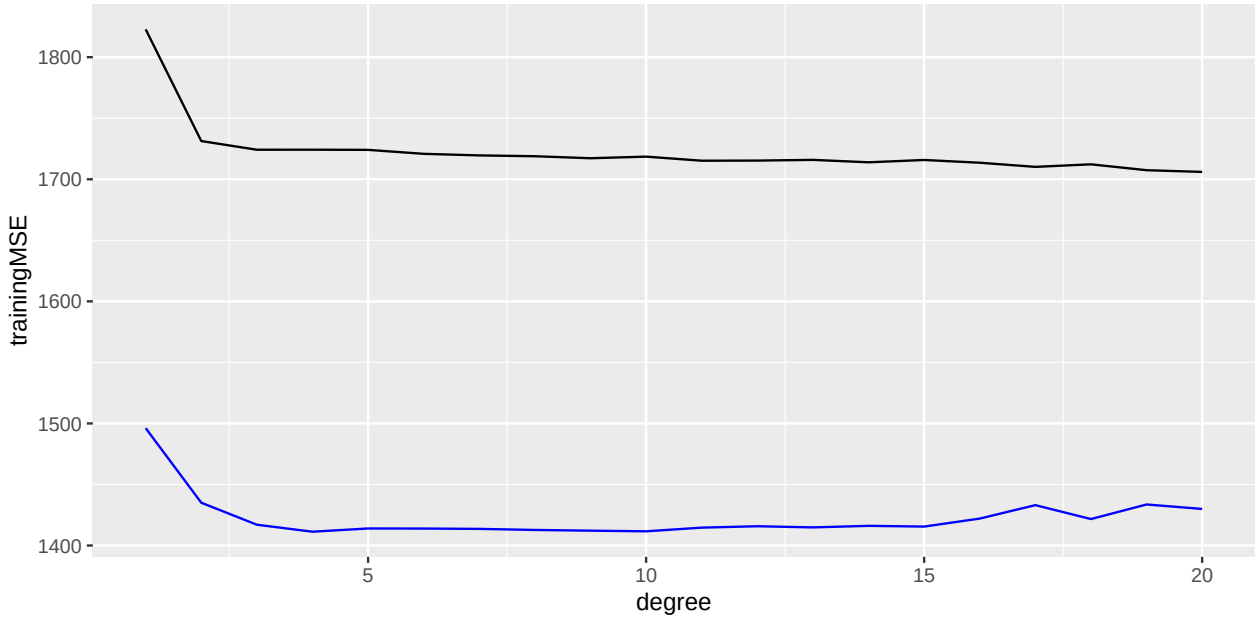
First, we randomly split the `Wage` data set into training and test sets, with 2000 observations in the training data and the remaining 1000 observations in the test data.

```
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, ..., 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?

```
# MSE on training and test sets
trainingMSE <- testMSE <- numeric(20)
for(i in 1:20)
{
  fit <- lm(log(wage) ~ ns(age, df=i), data=train)
  trainingMSE[i] <- mean((train$wage - exp(fitted(fit)))^2)
  testMSE[i] <- mean((test$wage - exp(predict(fit,newdata=test)))^2)
}

qplot(degree, trainingMSE, geom="line",
      data=data.frame(degree=1:20, trainingMSE, testMSE)) +
  geom_line(aes(degree, testMSE), col='blue')
```



Exercise 9

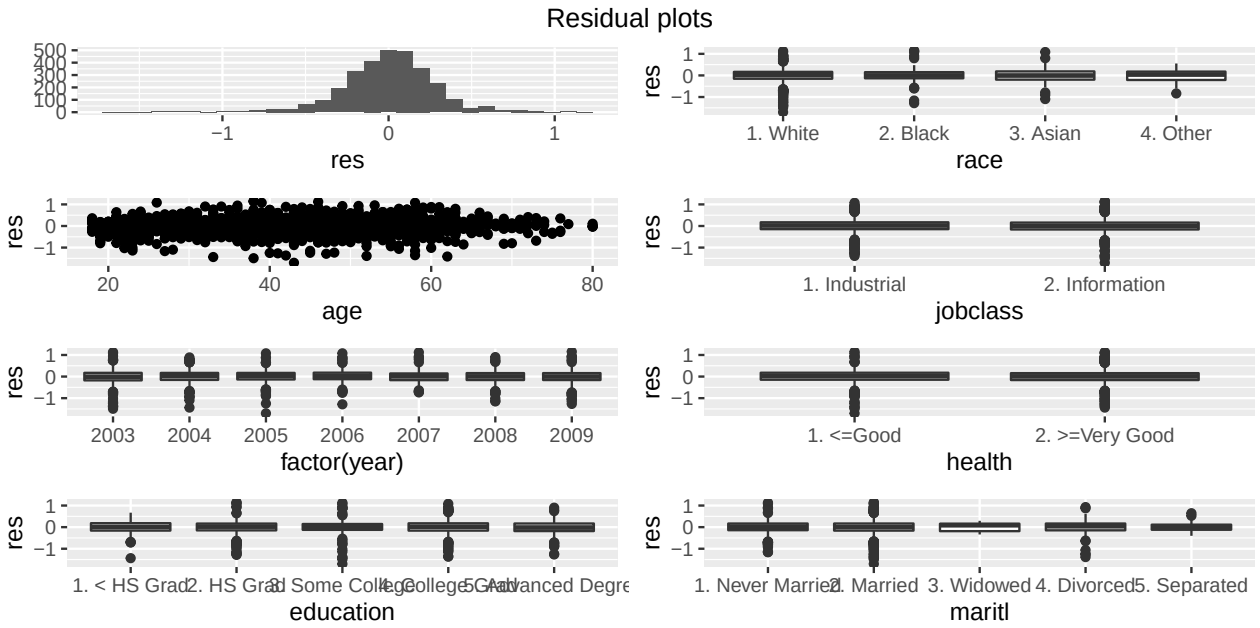
- Repeat the previous analysis, but use the full linear model including the other variables in the data set.
- How much better is the test MSE once you include the other predictor variables?

- Check your model by plotting the residuals as a function of each predictor variable. Do you see anything unusual in the residual plots?

```
fit <- lm(log(wage) ~ year + ns(age, df=5) + education + race + jobclass + health + maritl, data=Wage)
```

```
library(gridExtra)
res <- residuals(fit)
resplots <- list()
resplots[[1]] <- qplot(res)
resplots[[2]] <- qplot(age,res, data=Wage)
resplots[[3]] <- qplot(factor(year),res, data=Wage, geom="boxplot")
resplots[[4]] <- qplot(education,res, data=Wage, geom="boxplot")
resplots[[5]] <- qplot(race,res, data=Wage, geom="boxplot")
resplots[[6]] <- qplot(jobclass,res, data=Wage, geom="boxplot")
resplots[[7]] <- qplot(health,res, data=Wage, geom="boxplot")
resplots[[8]] <- qplot(maritl,res, data=Wage, geom="boxplot")

marrangeGrob(resplots, ncol=2, nrow=4, top="Residual plots")
```




```
res <- residuals(fit)
outliers <- subset(Wage, abs(res) > 1.5)
```