Data Mining Test (Lab)

23/11/2018

Time: 1 hour 10 mins

In the TESTO folder, you can find the RMarkdown file "consegna.Rmd". Write your answers (R code and text) there, then

1. use the Knit button to generate an HTML file
2. name the HTML file with your badge number
3. upload the HTML file to the CONSEGNA folder

Other formats will not be accepted.

Exercize 1 (ISLR, Chapter 5, Applied Exercize 8)

Points 3

We will now perform cross-validation on a simulated data set. Generate a simulated data set as follows:

```
set.seed(1)
x = rnorm(100)
y = x - 2*x^2 + rnorm(100)
```

a. Set the random seed `set.seed(123)`, and then **print in output only** the LOOCV errors that result from fitting the following five models using least squares:

i) $Y = \beta_0 + \epsilon$

ii) $Y = \beta_0 + \beta_1 X + \epsilon$

iii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

iv) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

v) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

```
# write here the R code
```

b. Which of the models has the smallest LOOCV error? Is this what you expected? Explain your answer.

Write here your answer.

Exercize 2 (ISLR, Chapter 8, Conceptual Exercize 5)

Points 2

Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of Pr(Class is Red$|X$):

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach. The second approach is to classify based on the average probability.

a. Write the R code to compute the majority vote and the average probability and **print in output only** the results.

```
# write here the R code
```

b. What is the final classification under each of these two approaches?

Write here your answer.

Exercize 3 (ISLR, Chapter 6, Applied Exercize 10)

Points 3

Load the Boston data set from the `MASS` library. The response variable is `medv`.

a. Split the data set into a training set containing the first 300 observations and a test set containing the last 206 observations. Use the `regsubsets()` function from the `leaps` library to perform best subset selection on the training set. **Print in output only** the test MSE associated with the best model of each size.

```
# write here the R code
```

b. For which model size does the test MSE take on its minimum value? Is this what you expected? Explain your answer.

Write here your answer.

Exercize 4

Points 3

Consider a Fixed-X setting where the response is generated according to the model

$$y_i = f(x_i) + \varepsilon_i$$

where

- $x_i = -1 + \frac{(i-1)}{10}, \quad i = 1, \ldots, n$
- $n = 21$
- the true regression function is $f(x_i) = (x_i - 3)(x_i - 2)(x_i - 1)x_i(x_i + 1)(x_i + 2)(x_i + 3)$
- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ with $\sigma = 2$.

a. Consider a polynomial regression model of degree d. **Print in output** the test prediction error $\text{ErrF} = \mathbb{E}(\text{MSE}_{\text{Te}})$ for $d = 1, 2, \ldots, 10$.

```
# write here the R code
```

b. Which is the degree that minimize ErrF? Is this what you expected? Explain your answer.

Write here your answer.