Data Mining (Lab)

22/02/2019

Time: 1 hour 30 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

Points 4

One way to mitigate the shortcomings of regression tree models (consider the R function **rpart**) is by bootstrap aggregation, or bagging. In bagging, you draw bootstrap samples (random samples with replacement) from your data. From each sample, you build a regression tree model. The final model is the average of all the individual regression trees.

a. Write a function called *predict.bag* which takes in input

- *ntree*, i.e. the number of boostrap samples
- *train* i.e. the training data
- *fml* i.e. the regression tree model formula
- *newdata* i.e. the test data

and gives in output the predicted values.

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

b. Split the Boston data into training and test as follows:

```
library(MASS)
set.seed(123)
istrain =  sample(c(T,F), nrow(Boston), rep=T)
train = Boston[istrain,]
test = Boston[!istrain,]
```

Apply your bagging function *predict.bag* with

- 100 bootstrap samples
- training and test Boston data
- `"medv ~ ."` as the model formula

Use `set.seed(123)` before running the function.

Print in output the MSE on the test set for

- the regression tree model `rpart(medv ~ ., train)`
- the bagging model

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

Comment the results.

Write here your answer.

Exercize 2 (ISLR, Chapter 6, Applied Exercize 10)

Points 4

Generate a simulated data set as follows:

```
set.seed(123)
n = 1000
p = 20
X = matrix(rnorm(p*n), ncol=p)
beta = c(2,rep(1,p))
y = beta[1] + X%*%beta[-1] + rnorm(n,mean=0,sd=100)
```

where the true coefficients are $\beta_0 = 1$ (intercept) and $\beta_1 = \ldots = \beta_{20} = 1$.

Split your data set into a training set containing the first 100 observations and a test set containing the last 900 observations.

Perform best subset selection on the training set to obtain the best model of size k (i.e. the model including the intercept term and k selected predictors) for $k = 1, 2, \ldots, 10$.

Note that if the jth predictor is not included in the best model of size k, then $\hat{\beta}_j^k = 0$, where $\hat{\beta}_j^k$ is the jth coefficient estimate for the best model of size k, $j = 0, 1, \ldots, 20$.

Print in output the scatterplot displaying $d_k = \sqrt{\sum_{j=0}^{20}(\beta_j - \hat{\beta}_j^k)^2}$ (y-axis) for a range of values of k from 1 to 10 (x-axis).

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

Comment on what you observe.

Write here your answer.

Exercize 3

Points 4

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

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

where

- $x_i = i, \quad i = 1, \ldots, n$
- $n = 100$
- the true regression function is $f(x_i) = 1 + x_i$
- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ with $\sigma = 10$.

Consider a polynomial regression model of degree $d = 99$. **Print in output** the test prediction error $\text{ErrF} = \mathbb{E}(\text{MSE}_{\text{Te}})$.

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