Business Analytics - ETC3250 2018 - Lab 7

Cross-validation

Souhaib Ben Taieb

12 April 2018

Exercise 1

Understand all the steps in the proof of the leave-one-out cross-validation (LOOCV) statistic for linear models available at https://github.com/bsouhaib/BA2018/blob/master/slides/week6/loocv-proof.pdf

Assignment - Question

Consider a simple regression procedure applied to a dataset with 1000 predictors and 200 samples:

1. Find the 5 predictors having the largest correlation with the response
2. Apply linear regression using only these 5 predictors

You will simulate the wrong way and the right way to perform cross validation.

(a) Write a function that estimate the test error of this proedure using 10-folds cross-validaiton, the right way (as explained in Lecture 12)

(b) Write a function that estimate the test error of this proedure using 10-folds cross-validaiton, the wrong way (as explained in Lecture 12)

(c) Produce 100 samples from the data generating process below. For each sample, run the functions in (a) and (b). Then, produce a boxplot of the cross-validate errors. Briefly describe what you observe.

```
p <- 1000
n <- 200
X  <- matrix(rnorm(n*p),n)
y  <- runif(n)
```

The previous data generating process assume that there are 1000 standard normal predictors, which are uncorrelated from the response, which is uniform in {0,1}.

TURN IN

- Your `.Rmd` file (which should knit without errors and without assuming any packages have been pre-loaded)
- Your Word (or pdf) file that results from knitting the Rmd.
- DUE: April 22, 11:55pm (late submissions not allowed), loaded into moodle