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)

```
# SOLUTION
PickBest <- function(DT, m = 5) {
  X <- DT[, -which(colnames(DT) == "y")]
  y <- DT[, "y"]
  p <- ncol(X)
  correlations <- numeric(p)
  for (i in seq(p)) {
    correlations[i] <- abs(cor(X[,i], y))
  }
  sorted <- sort(correlations,index.return=T,decreasing=T)
  bestm <- sorted$ix[seq(m)]
  bestvar <- colnames(X)[bestm]
  return(bestvar = bestvar)
}

rightCV <- function(DT, K) {
  n <- nrow(DT)
  DT <- DT[sample(n),]
  nk <- floor(n/K)
  err <- numeric(n)

  for(k in seq(1, K)){
    ind_k_fold <- seq((k - 1) * nk + 1, k * nk)

    DT_train <- DT[-ind_k_fold, ]
    bestvar <- PickBest(DT_train, 3)
    var_selected <- which(colnames(DT_train) %in% c(bestvar, "y") )
    DT_train <- DT[-ind_k_fold, var_selected]
    fit <- lm("y ~ .", data = DT_train)
```

```
    DT_test <- DT[ind_k_fold, var_selected]

    pred <- predict(fit, newdata = DT_test)
    err[ind_k_fold] <- (DT_test$y - pred)^2
  }
  return(mean(err))
}
```

(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)

```
# SOLUTION
wrongCV <- function(DT, K) {
  n <- nrow(DT)
  DT <- DT[sample(n),]
  nk <- floor(n/K)
  err <- numeric(n)

  bestvar <- PickBest(DT, 3)
  var_selected <- which(colnames(DT) %in% c(bestvar, "y") )
  DT <- DT[, var_selected]

  for(k in seq(1, K)){
    ind_k_fold <- seq((k - 1) * nk + 1, k * nk)

    DT_train <- DT[-ind_k_fold, ]
    fit <- lm("y ~ .", data = DT_train)
    DT_test <- DT[ind_k_fold, ]

    pred <- predict(fit, newdata = DT_test)
    err[ind_k_fold] <- (DT_test$y - pred)^2
  }
  return(mean(err))
}
```

(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.

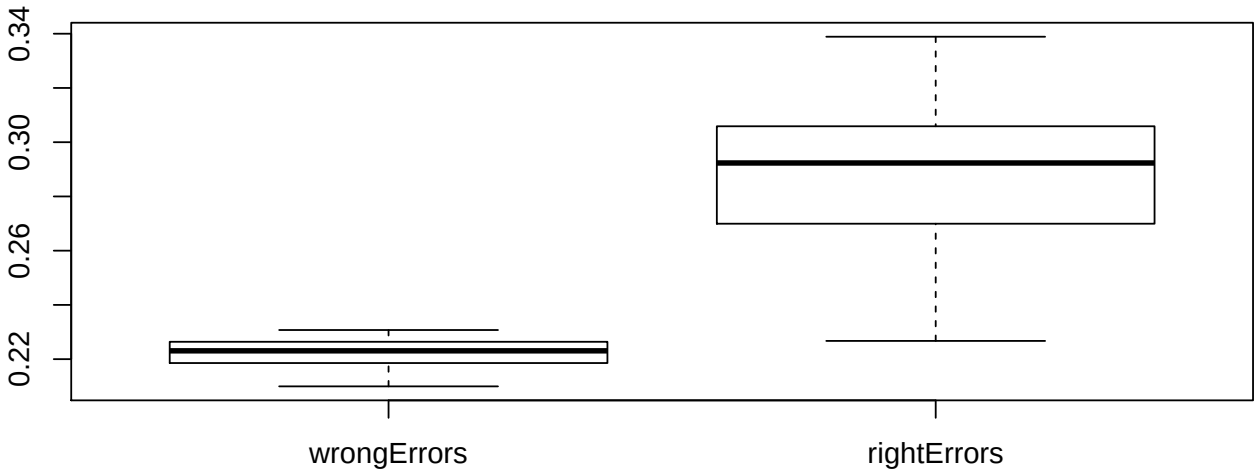
```
p <- 1000
n <- 200
X  <- matrix(rnorm(n*p),n)
y  <- runif(n)
```

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}.

```
# SOLUTION
p <- 1000
n <- 200
M <- 20 # M <- 100
wrongErrors <- rightErrors <- numeric(M)
for (i in seq(M)) {
  X  <- matrix(rnorm(n*p),n)
  y  <- runif(n) < 0.5
  DT <- data.frame(X = X, y = y)
```

```
  wrongErrors[i] <- wrongCV(DT, 5)
  rightErrors[i] <- rightCV(DT, 5)
}
errors <- data.frame(wrongErrors = wrongErrors, rightErrors = rightErrors)

boxplot(errors)
```



We can clearly see that the wrongCV underestimate the true expected test error.

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