DSCI353-353m-453: LE3: Resampling, Model Selection

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

17 February, 2023

Contents

3.1.1 LE3, 2 parts, 9 points.

- LE3A, 3 questions, 4 points, ISLR Chapter 7 and 10.
  - 3A-1 = 2.0 points

  - 3A-2 = 2.0 points
  - 3A-3 = 1.0 points
- LE3B, 6 questions, 4 point, ISLR Chapters 5 and 6.
  - 3B-1:5 = 1.0 point total
  - 3B-6 = 2.0 point total
  - 1 point Code Style

Details

- Due Thursday, February 23
  - Midnight
- The grading is done on how you show your thinking,
  - explain yourself and
  - show your R code and
  - the output you got from your code.
- Code style is important
  - Follow Rstudio code diagnostics notices
  - And the Google R Style Guide

To be done as an Rmd file,

- where you turn in
  - the Rmd file and
  - the compiled pdf showing your work.
  - and the R script of IntroR.R

You will want to produce a report type format

- (html and pdf type document) to turn in.
- And not an ioslides or beamer (slide type) compiled output.
  - These are presentation formats, and can be fussy

Also are you backing up your git repo

- in a second and third location,
- to avoid corruption problems?

3.1.2 LE3A: ISLR Chapters 7 and 10

3.1.2.1 3A-1. ISLR 7.10 (2.0 points)

This question relates to the *College* data set.

3.1.2.1.1 (a)

Split the data into a training set and a test set

- Using out-of-state tuition as the response
  - and the other variables as the predictors,
- perform forward step-wise selection on the training set
  - in order to identify a satisfactory model
  - that uses just a subset of the predictors.

You'll want to show the learning curves

- that show Cp, AIC and/or $AdjR^2$
  - as a function of the number of variables chosen,
  - to demonstrate the best model.

```
# Put your code here, with comments and good style and syntax
library(ISLR)
library(leaps)
```

```r
attach(College)
set.seed(2)
```

ANSWER:

3.1.2.1.2 (b) Fit a GAM on the training data

- using out-of-state tuition as the response
- and the features selected in the previous step as the predictors.
- Plot the results, and explain your findings.

```r
# Put your code here, with comments and good style and syntax

library(gam)
```

```
## Loading required package: splines

## Loading required package: foreach

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##     accumulate, when

## Loaded gam 1.22
```

ANSWER:

3.1.2.1.3 (c) Evaluate the model obtained on the test set

- and explain the results obtained.

```r
# Put your code here, with comments and good style and syntax
```

ANSWER: We obtain a test R-squared of

3.1.2.1.4 (d) Non-linear relationship

For which variables,

- if any,
- is there evidence of a non-linear relationship with the response?

```r
# Put your code here, with comments and good style and syntax
```

ANSWER:

3.1.2.2 3A-2. GAM with Smoothing Spline (2 points) Use GAMs to model how the economy is affected by weather.

3.1.2.2.1 (a) Generate 100 rnorm data points

- to make your economy and weather dataframe

3.1.2.2.2 (b) Do a linear model on this dataframe

- and plot the data
- and the abline

```
# Put your code here, with comments and good style and syntax
```

3.1.2.2.3 (c) Now use a smoothing spline

- to locally smooth your weather
- And calculate another lm linear model.

Using a bs spline and 3 knots

- indicate where the knots in your spline are
- What is a bspline, a natural spline and a cubic spline?

```
library(splines)
```

ANSWER:

3.1.2.2.4 (d) Explain your lm call

- and how the smoothing spline gets incorporated?
- What do the arguments for the spline mean?

ANSWER:

3.1.2.2.5 (e) Plot the data,

- your smoothing spline model (use the predict function)

```
# Put your code here, with comments and good style and syntax
```

3.1.2.2.6 (f) Now use a GAM

- You can use the gam package, or the mgcv package
- Explain your gam call, its arguments and their meaning

```
# Put your code here, with comments and good style and syntax
library(mgcv)
```

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.

##
## Attaching package: 'mgcv'

## The following objects are masked from 'package:gam':
##
##     gam, gam.control, gam.fit, s
```

ANSWER: k - the dimension of the basis used to represent the smooth term. bs - "cr" for a cubic regression spline.

3.1.2.3 3A-3. Real World Data (1.0 points)

Here, we are looking at the wine quality of various red wines which can be predicted based off of a number of features (acidity, pH, density, etc).

Read in the data and make a linear model to predict wine quality based off of all of the predictors

```
winequality_red <- read_delim("./data/winequality-red.csv",
    ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 1599 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## dbl (12): fixed acidity, volatile acidity, citric acid, residual sugar, chlo...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

We will use step-wise AIC in order to assist in our variable selection process. Remember that R-squared always decreases as you add more variables, so we need to use a different estimator of prediction error (AIC).

Use step-wise AIC in order to select the variables most important towards predicting wine quality.

We can use bootstrapping in order to gain inference from a population based on a sample of data. This technique can be used to estimate the standard error of any statistic and to obtain a confidence interval for it.

Using the boot package in R, measure the 95% confidence interval of the r-squared for the quality of wine in our wine quality dataset (use the boot::boot() function to bootstrap the data, and then use boot.ci() to get the confidence interval).

Measure the confidence interval twice. Once for quality ~ all the variables, then again for quality ~ the variables you selected through step-wise AIC.

```
# Bootstrapping function: function used for the statistic call in the boot
# function

rsq_function <- function(formula, data, indices) {
  d <- data[indices,] #allows boot to select sample
  fit <- lm(formula, data = d) #fit regression model
  return(summary(fit)$r.square) #return R-squared of model
}

myBootstrap1 <- winequality_red %>%
  boot(statistic = rsq_function, R = 1500, formula = quality ~ .)

# Bootstrap here
# formula is the formula from your linear model y ~ x

# Example : myBootstrap <- boot(data = _, statistic = _, R = _, formula = _)

# Use boot.ci to calculate the confidence intervals

myBootstrap2 <- winequality_red %>%
  boot(statistic = rsq_function, R = 1500,
       formula = quality ~ `volatile acidity` + chlorides + `free sulfur dioxide` +
    `total sulfur dioxide` + pH + sulphates + alcohol)

boot.ci(myBootstrap1, index = 1, type = 'basic')
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = myBootstrap1, type = "basic", index = 1)
##
## Intervals :
## Level      Basic
## 95%   ( 0.3185,  0.3923 )
## Calculations and Intervals on Original Scale
```

```
boot.ci(myBootstrap2, index = 1, type = 'basic')
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = myBootstrap2, type = "basic", index = 1)
##
## Intervals :
## Level      Basic
## 95%   ( 0.3187,  0.3949 )
## Calculations and Intervals on Original Scale
```

```
#
```

ANSWER:

3.1.3 LE3B: ISLR Chapters 5 and 6

3.1.3.1 3B-1. ISLR 5.1 (0.2 point possible)

3.1.3.1.1 (a) When we fit a model to data, which is typically larger?

1. Test Error
2. Training Error

ANSWER:

3.1.3.1.2 (b) What are reasons why test error could be LESS than training error?

1. By chance, the test set has easier cases than the training set.
2. The model is highly complex, so training error systematically overestimates test error
3. The model is not very complex, so training error systematically overestimates test error

ANSWER:

3.1.3.2 3B-2. ISLR 5.2 (0.2 points) Suppose we want to use cross-validation to estimate the error of the following procedure:

- Step 1: Find the k variables most correlated with y
- Step 2: Fit a linear regression using those variables as predictors

We will estimate the error for each k from 1 to p, and then choose the best k.

True or false: a correct cross-validation procedure will possibly choose a different set of k variables for every fold.

ANSWER:

3.1.3.3 3B-3. ISLR 5.3 (0.2 points) Suppose that we perform forward stepwise regression and use cross-validation to choose the best model size.

Using the full data set to choose the sequence of models is the WRONG way to do crossvalidation (we need to redo the model selection step within each training fold).

If we do crossvalidation the WRONG way, which of the following is true?

1. The selected model will probably be too complex
2. The selected model will probably be too simple

ANSWER:

3.1.3.4 3B-4. ISLR 5.4 (0.2 points) One way of carrying out the bootstrap is to average equally over all possible bootstrap samples from the original data set (where two bootstrap data sets are different if they have the same data points but in different order). Unlike the usual implementation of the bootstrap, this method has the advantage of not introducing extra noise due to resampling randomly.

To carry out this implementation on a data set with n data points, how many bootstrap data sets would we need to average over?

(You can use "^" to denote power, as in "n^2")

ANSWER:

3.1.3.5 3B-5. ISLR 5.5 (0.2 points) If we have n data points, what is the probability that a given data point does not appear in a bootstrap sample?

ANSWER:

3.1.3.6 3B-6. ISLR 6.10 (2.0 point possible) We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

3.1.3.6.1 (a) Generate a data set with p = 20 features, n = 1,000 observations, and an associated quantitative response vector generated according to the model

$$Y = X\beta + \epsilon$$

where β has some elements that are exactly equal to zero.

```
set.seed(1)
num_predictors = 20
num_observations = 1000

x <- matrix(rnorm(num_observations  * num_predictors ), num_observations, num_predictors)

beta <- rnorm(num_predictors)
beta[sample(1:num_predictors, sample(1:num_predictors, 1))] = 0

eps <- rnorm(num_predictors)

y <- x %*% beta + eps
```

3.1.3.6.2 (b) Split your data set into a training set containing 100 observations and a test set containing 900 observations.

```r
# Put your code here, with comments and good style and syntax
```

3.1.3.6.3 (c) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.

```r
# Put your code here, with comments and good style and syntax
library(leaps)
```

3.1.3.6.4 (d) Plot the test set MSE associated with the best model of each size.

```r
# Put your code here, with comments and good style and syntax
```

3.1.3.6.5 (e) For which model size does the test set MSE take on its minimum value? Comment on your results. If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

```r
# Put your code here, with comments and good style and syntax
```

ANSWER:

3.1.3.6.6 (f) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

```r
# Put your code here, with comments and good style and syntax
```

ANSWER:

3.1.3.6.7 (g) Create a plot displaying $\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$ for a range of values of r, where $\hat{\beta}_j^r$ is the jth coefficient estimate for the best model containing r coefficients.

Comment on what you observe.

```r
# Put your code here, with comments and good style and syntax
```

How does this compare to the test MSE plot from (d)?

```r
# Put your code here, with comments and good style and syntax
```

ANSWER:

3.1.3.7 Links <!– # Keep a complete change log history at bottom of file. # Complete Change Log History # v0.00.00 - 1405-07 - Nick Wheeler made the blank script ##########