1808-dsc351-351m-451-homework6-ISLR-ModVarSelect

Roger H. French, JiQi Liu

08 November, 2018

Contents

Use ISLR Chapter 6, Linear Model Selection & Regularization And ISLR Chapter 7, Moving Beyond Linearity for this assignment

1 Questions, 5 points total.

1 ISLR Exercise 6.10 (6 parts a-f, 5 points total)

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.

1.1 (a)

A data set (HW6-testdata.csv) has been generated

- with p = 10 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.

The β values

- are randomly generated coefficients
- which are with the p features (X1-10)
  - to produce the y response.

Read in and describe the data set.

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

1.2 (b)

Split your data set

- into a training set containing 100 observations
- and a test set containing 900 observations
  - using the caret package.

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

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

1.3 (c)

Perform best subset selection on the training set,

- and plot the training set MSE
- associated with the best model of each size.

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

1.4 (d)

Plot the test set MSE

- associated with the best model of each size.

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

1.5 (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.

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

1.6 (f)

In this data set,

- 3 of the β coefficients were made to be 0, = how is this reflected in your test MSE fit?

Comment on the coefficient values

- what are their signs and magnitudes?

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

2 Cites

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. 1st ed. 2013, Corr. 5th printing 2015 edition. Springer Texts in Statistics. New York: Springer, 2013.