HW8

Ben Tankus

2/23/2021

1. (2 points) (ISLR 2.4 Exercise #1, page 52) For each of the following parts, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answers.

(a) The sample size n is extremely large, and the number of predictors p is small.
(b) The number of predictors p is extremely large, and the number of observations n is small.
(c) The relationship between the predictors and response is highly non-linear.
(d) The variance of the error terms, i.e., sigma

2 = Var(epsilon), is extremely high. 2. (4 points) (ISLR 5.4 Exercise #8, page 201) We will now perform cross-validation on a simulated dataset.

(a) Generate a simulated data set as follows:

```
set.seed(1)
x <- rnorm(100)
y <- x - 2*x^2 + rnorm(100)
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

(b) Create a scatter plot of X against Y using the data you generated above. Comment on what you see.

(c) Set a random seed, and then compute the leave-one-out cross-validation (LOOCV) errors that result from fitting the following four models using least squares:

- Y = beta0 + beta1X + epsilon
- Y = beta0 + beta1X + beta2X2 + epsilon
- Y = beta0 + beta1X + beta2X2 + beta3X3 + epsilon
- Y = beta0 + beta1X + beta2X2 + beta3X3 + beta4X4 + epsilon Note that you may find it helpful to use the data.frame() function to create a single data set containing both X and Y .

(d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why or why not?

(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

3. (4 points) (Based on ISLR 6.8 Exercise #11, page 264 — Predicting crime rates in Boston data.) The Boston data set is in the MASS package, you'll need to load that first.

```r
library(MASS)
?Boston
```

```
## starting httpd help server ... done
```

```r
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

Your job is to build a regression model to predict the crime rate (crim) in Boston suburbs based on the other provided variables. Your solution should include: • A brief exploratory analysis (some summary statistics, and a few plots of any obvious relationships). • A description of the set of regression models you considered. • A description of how the models were evaluated. • A summary of one (or a few) models that based on your analysis are the best among those you considered.