Stat 427/627 Statistical Machine Learning

Homework 1

Due: Friday, May 24

Contents

Total: $8 + 4 + 4 + 8 + 16 = 40$

1 Exercise 2.4.1. Flexible or Inflexible (p. 52. 8 points)

For each part, 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 answer**.

(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 is extremely high.

2 Exericse 2.4.2.(a,b) Classification or Regression (p. 52, 4 pts)

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 300 publicly-traded firms in the US. For each firm we record profit, number of employees, industry, average stock price over the last year, and the CEO salary. We are interested in understanding which factors affect CEO salary.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

(c) Skipped.

3 Exercise 2.4.4 (modified) Examples of SML. (p. 53, 4 pts)

Identify some examples of potential real-life applications for statistical learning. In each example, describe the response and the predictors. Also state the goal - inference or prediction.

(a) Describe ***one*** real-life applications in which classification might be useful.
(b) Describe ***one*** real-life applications in which regression might be useful.

4 Exercise 3.7.4. Training and Test Residual Sums of Squares (p. 122, 8 pts)

Residual Sums of Squares (RSS) is also referred to as the "Sum of Squares of Errors'' (SSE).

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, to include a separate cubic regression, i.e.,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

(a) Suppose the true relationship between X and Y is linear, i.e., $Y = \beta_0 + \beta_1 X$. Consider the **training** Residual sum of squares (RSS) for the linear regression, and also the **training** RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using **test** rather than training RSS.

(c) Suppose the true relationship between X and Y is **not** linear, but we don't know how far it is from linear. Consider the **training** RSS for the linear regression, and also the **training** RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using **test** rather than training RSS.

5 Application: Predict the number of college application. (16 pts)

Predict the number of applications received based on the other variables in the `College` data set in {`ISLR2`}.

Use the following code to split the data set randomly. Only use the subset `college.data` in this exercise.

```
library(ISLR2)
data("College")

my.college <- College[-484, ] # remove an extreme case.
#my.college <- College[College$Apps <=16000, ] # remove several extreme case.
train.pct <- 0.78
set.seed(2024)
Z <- sample(nrow(my.college), floor(train.pct*nrow(my.college)))
college.data <- my.college[Z, ]
holdout.data <- my.college[-Z, ]
```

(a) (2 pts) Read the help file of the data file and determine the response variable for this study. Prepare a histograms of the response with and without case 484 in the original `College` data set. Why would I remove it from the analysis for this excise?

(b) (2 pts) Fit a first order linear regression with all available predictors. Be sure to use `college.data` data frame. (In `lm()` and `glm()` function, `y ~ .` fit a model between `y` and all other variables in the data. `y ~ 1` fits a model with only intercept without any predictor.)

(c) (4 pt) Compute the variance inflation factor and comment on the severity of the collinearity of the data. Why is "collinearity" a concern, even if the model is correct?

(d) (8 pts) Find the best least squares regression model(s), using $adjusted - R^2$, BIC, and Cp criteria with best subset algorithm. Also use the stepwise variable selection algorithm using AIC criterion. Note that the "best" model(s) may be different depending on the criterion. (Hint: In the `regsubsets()` function, set `nvmax = 17` to consider models with up to 17 predictors.)

—— This is the end of HW 1. ——