Introduction to Statistical Learning with R (ISLR) Chapter2 Excercises Answers

Peiyun Zhou

05/13/2019

2.4 Exercises

Q1. For each of parts (a) through (d), 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.

- An inflexible method will be better than an flexible method.

(b) The number of predictors p is extremely large, and the number of observations n is small.

- A flexible method is better than an infledxible method.

(c) The relationship between the predictors and response is highly non-linear.

- A flexible method is better than an infledxible method. Because the non-linear relationship between the predictors and response violates the basic assumption for linear regression.

(d) The variance of the error terms is extremely high.

- A flexible method is better than an infledxible method. Because one basic assumption for linear regression is the variance of the errors should be normally distributed.

Q2. 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 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

- This is a regression problem, because we are interested in making inference about what factors affect CEO salary. In this case, N=500 and p value should be smaller than 0.05.

(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.

- This is a classification problem. We are interested in making prediction. N=20 and p values should be smaller than 0.05.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

- This is a regression problem. We are interested in making inference about the relationship between the % change in USC/Euro exchange rates and the weekly changes in world stock markets. N=52 (total weeks in a year) and p values should be smaller than 0.05.

Q3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training er- ror, test error, and Bayes (or irreducible) error curves, on a sin- gle plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent

- Check Figure 2.12 on Page 36

Q4. You will now think of some real-life applications for statistical learning

(a) the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

(b) Explain why each of the five curves has the shape displayed in part (a)

Q4: You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. The goal is to predict whether a credit card transaction is fraud or not. The dependent varaible is whether Predictors are previous categories of the credit card transaction, whether this is a
2. The goal aims to classify students who be admitted to a college. The predictors are students' SAT, GPAs, age, schools, and gender. The dependent varaible is whether the student has been admitted.
3. The goal aims to classify whether the users will buy the new video game or not. The predictors are how many hours the users spend on playing video games; whether the new vedio game is their favorite type.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Examine how educational levels, gender, and majors affect annual income.
2. Study the effects of temperature and amounts of rain on the corn production amount.
3. Explore how team and age affect basketball players' successful shooting rates.

(c) Describe three real-life applications in which cluster analysis might be useful.

1.
2.
3.

Q5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

- "If we are mainly interested in inference, then restrictive models are much more interpretable. In contrast, very flexible approaches, such as the splines and the boosting methods discussed can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response." (P.25)

Q6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a para- metric approach to regression or classification (as opposed to a non- parametric approach)? What are its disadvantages?

- The advantage of a parametric approach: "it reduces the problem of estimating f down to one of the esitmating a set of parameters." (P.21). "The potential disadvantage is that the model we choose will usually not match the true unknown form of f." (P.22)
- None-parametric approch has a major advantage over parametric approaches: "by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f." (P.23) But it do suffer from a major disadvantage: "since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f." (P.23)

Q8 College Datasets

```
college<-read.csv("College.csv",sep=",")
rownames(college)=college[,1]
names(college)
```

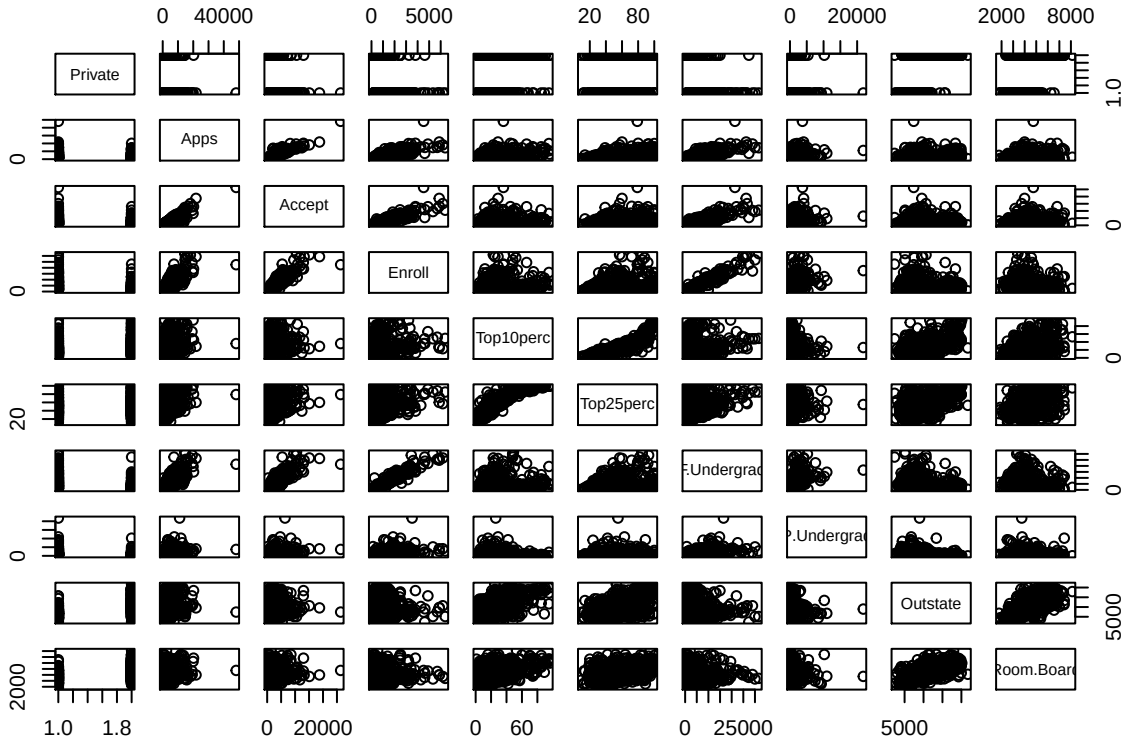
```
##  [1] "X"           "Private"     "Apps"        "Accept"      "Enroll"
##  [6] "Top10perc"   "Top25perc"   "F.Undergrad" "P.Undergrad" "Outstate"
## [11] "Room.Board"  "Books"       "Personal"    "PhD"         "Terminal"
## [16] "S.F.Ratio"   "perc.alumni" "Expend"      "Grad.Rate"
```

```
college=college[,-1]
summary(college)
```

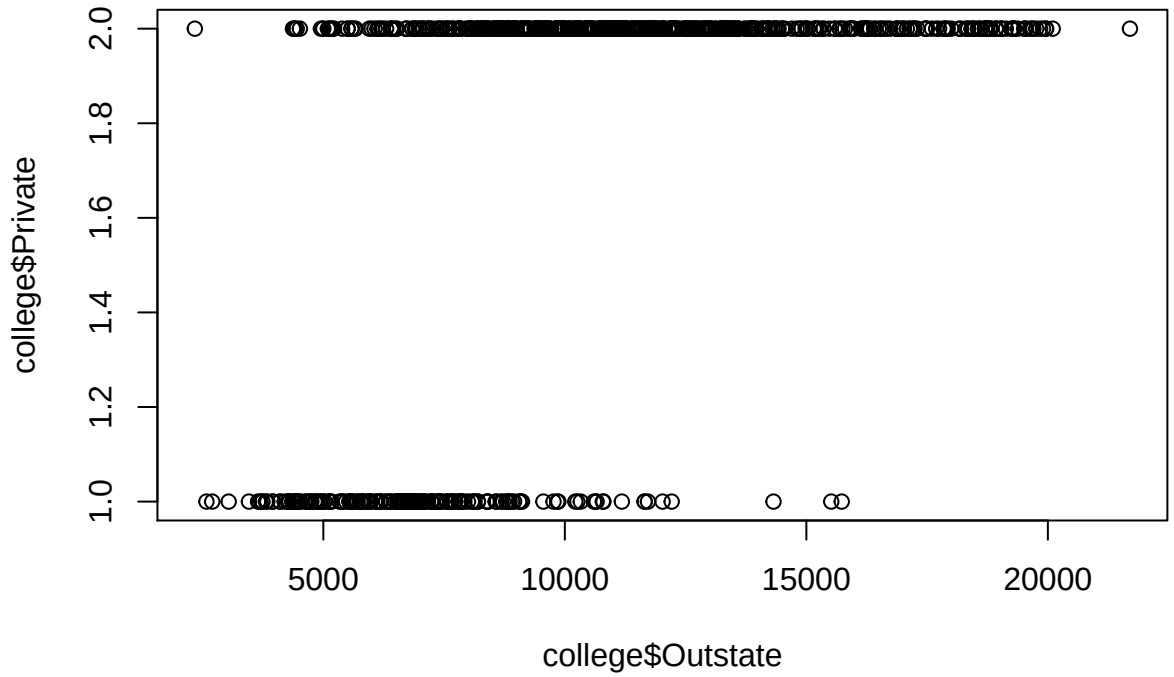
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
```

```
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

```
pairs(college[,1:10])
```




```
plot(college$Outstate, college$Private)
```




```
Elite=rep("No",nrow(college))
Elite[college$Top10perc >50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college ,Elite)
summary(college$Elite)
```

```
##  No Yes
## 699  78
```