

Chapter2-Exercises

Anthony Chau

12/8/2019

Contents

Conceptual Exercises	2
Question 1	2
Question 2	3
Question 3	4
Question 4	5
Question 5	7
Question 6	8
Question 7	9
Applied Exercises	9
Question 8	10
Question 9	11
Question 10	11

Conceptual Exercises

Question 1

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.
- (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. $\sigma^2 = \text{Var}()$, is extremely high.

Solution:

- (a) We would expect the performance of a flexible statistical learning method to be better than an inflexible method because the flexible methods can model the data more accurately given the large amount of observations. Also, since the number of predictors is small, the flexible methods would not estimate a large number of parameters - increasing overall interpretability.
- (b) The performance of flexible methods is worse than inflexible methods. Given the large number of predictors and low number of observations, flexible methods would most likely result in overfitting as well as have low interpretability.
- (c) The performance of flexible methods is better than inflexible methods because the flexible methods can better capture non-linear relationships over inflexible methods (such as linear regression).
- (d) Flexible methods would perform better than inflexible methods because the flexible methods could include more potential variables that could be useful in predicting the response. The inclusion of those variables can offset the presence of a high variance for the error terms.

Question 2

2. 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.
 - (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) 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.

Solution:

- (a) This scenario is a regression problem because we are predicting a quantitative variable, CEO salary. We are mainly interested in inference - understanding the factors that affect CEO salary. For this case, $n = 500$ and $p = 3$.
- (b) This scenario is a classification problem because we are predicting a qualitative variable - whether the product is a success or a failure. We are mainly interested in predicting - correctly determining if a new product will be a success or a failure. For this case, $n = 20$ and $p = 13$.
- (c) This scenario is a regression problem because we are predicting a quantitative variable, USD/Euro exchange rate. We are mainly interested in prediction - what the USD/Euro exchange rate is given some predictor variables. For this case, $n = 52$ (52 weeks in a year) and $p = 4$.

Question 3

3. We now revisit the bias-variance decomposition.
- (a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent 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).

Solution:

(a)

(b)

Question 4

4. 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.
 - (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.
 - (c) Describe three real-life applications in which cluster analysis might be useful.

Solution:

- (a) One application could be to determine if an applicant will default on a loan. The response is Yes/No, indicating whether the applicant defaulted on a loan. Potential explanatory variables could include credit score, age, income, employment status, number of dependents, etc. The goal of this application is prediction - it is important to predict if an applicant will default on a loan!

Another application could be to determine if a person would move up in socioeconomic status. The response is Yes/No, indicating whether the applicant moved up (Yes) or down (No) from their SES status when they were born. Potential explanatory variables could be parent's education level, parent's employment, household income, born in city or rural area, number of siblings, etc. The main goal of this application is inference. We would like to understand the relationship between the response and the explanatory variables.

Yet another application is to determine if a basketball team will win the championship game. The response is Yes/No, indicating if the team will win the championship game. Potential explanatory variables could include: season record, average number of points scored during season, whether team won a championship in previous year(s), etc. This application could be both for inference and prediction. It is an interesting question to pose: what leads to basketball teams winning championship games? On the other hand, you may be betting with your buddies on the championship game, so your money is on the line!

- (b) One application of regression is to predict the yearly sales for a clothing company (response variable). Potential explanatory variables could include: past year sales, advertising expenditure, number of physical stores, number of unique visitors to website, etc.

Another application of regression is to model the relationship between age of death (response) with the following explanatory variables: average minutes spent exercising per week, time spent engaging in mental activities, average number of servings of fruits and vegetables, smoker status, weight, etc. This application focuses on inferences - understanding the relationship between response and predictors.

Another application of regression could be to model the number of delayed flights at an airport in a day(response). Potential predictors could include: average daily passengers total, location of airport, domestic/international airport status, average number of thunderstorms/severe weather incidents occurred per year, number of average daily nonstop passengers, number of TSA checkpoints, etc. The main interest of this application is inference: what are the key factors which contribute to long airport delays.

- (c) One application of cluster analysis is market segmentation of customers when a company wants to introduce a product to a new market. Cluster analysis can help the company gain insight into which customers would be most likely to be interested in the product or buy more of the product.

Another application of cluster analysis is to identify regions with similar weathers by using info such as latitude/longitude, humidity, temperature, sea level, altitude, etc.

Lastly, another application of cluster analysis could be to group together students within a school based on criterion such as age, ethnicity, family income, GPA, etc.

Question 5

5. 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?

Solution:

The advantage of a very flexible model is that the model will have low bias because it can model the data precisely. Another advantage of flexible models is that they can model more complex relationship between variables. Undoubtedly, any real-world relationships are not constant or linear.

A disadvantage of a very flexible model is the large number of parameters that need to be estimated. This makes the model less interpretable. This problem is exacerbated if we have a low number of observations but a high number of predictor variables. Another disadvantage of a flexible model is that it will most likely have high variance. The model will have a low train MSE, but will have a high test MSE since the model will struggle to accurately predict given brand new observations. The model was taught to adhere too closely to the training data which causes it to struggle with new test data.

A more flexible approach is preferred if the main goal is accurate prediction. Perhaps, accurate prediction would help a company's bottom line or provide a critical diagnosis on a patient.

On the other hand, a less flexible approach is preferred if the main goal is inference. That is, we value interpretability and seek to understand the relationship between variables.

Question 6

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

Solution:

Parametric statistical methods make assumptions about the functional form of f and reduce the problem of estimating f to solving for a set of parameters. For example, linear regression is a parametric method which assumes that f is linear in its predictors X . To implement linear regression, one can use least squares to solve for β .

Nonparametric statistical methods do not make assumptions about the functional form of f . Instead, they attempt to estimate f through mathematical techniques.

An advantage of a parametric method is that it is easy to estimate a set of parameters compared to estimating an entire function for nonparametric methods.

A disadvantage of a parametric method is that the true form of f may be completely different than what the stated assumptions of the parametric method. This is where nonparametric methods shine as they can estimate a greater variety of functional forms of f . However, because of the high flexibility of nonparametric methods, these methods can result in high variance and overfitting. Nonparametric methods require a very large number of observations to produce accurate estimate because they cannot reduce problem to estimating a small set of coefficients.

Question 7

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors. (a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$. (b) What is our prediction with $K = 1$? Why? (c) What is our prediction with $K = 3$? Why? (d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

Solution:

The dataset is recreated below.

```
df <- data.frame(X1 = c(0, 2, 0, 0, -1, 1), X2 = c(3, 0, 1, 1, 0, 1),
                 X3 = c(0, 0, 3, 2, 1, 1),
                 Y = c("Red", "Red", "Red", "Green", "Green", "Red"))
```

```
# Test Point
```

```
TestPoint <- c(0, 0, 0)
```

```
df <- rbind(df, TestPoint)
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = 0): invalid factor level, NA
## generated
```

```
# calculate euclidean distance
# last row is euclidean distance of test point to all observations
# in training set
dist(df[1:7, 1:3], method = "euclidean")
```

```
##           1           2           3           4           5           6
## 2 3.605551
## 3 3.605551 3.741657
## 4 2.828427 3.000000 1.000000
## 5 3.316625 3.162278 2.449490 1.732051
## 6 2.449490 1.732051 2.236068 1.414214 2.236068
## 7 3.000000 2.000000 3.162278 2.236068 1.414214 1.732051
```

- (b) When $K = 1$, our prediction is Green. Because the closest observation to the test point is observation 5 and it is Green.
- (c) When $K = 3$, our prediction is Green. The three closest observations to the test point are observations 4, 5, and 6. Observations 4 and 5 are Green and observation 6 is Red. So the estimated probability for the Green Class is $2/3$ and K-Nearest Neighbors predicts that the test point belongs to the Green class.

Applied Exercises

Question 8

Solution:

Question 9

Solution:

Question 10

Solution: