Solutions for

"An Introduction to Statistical Learning: with Applications in R"

by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.

Contents

1 Preface

Solutions only for conceptual exercises.

2 Statistical Learning

2.1 ex1

Question: 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. $2 = \text{Var}()$, is extremely high.

Solution:

(a) The flexible method is better than inflexible.

(b) The flexible method is worse than inflexible.

(c) The flexible method is better than inflexible.

(d) The flexible method is worse than inflexible.

2.2 ex2

Question: 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 interesting in predicting the % change in the US dollar 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 dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

Solution:

(a) Regression, inference, n = 500, p = 4.

(b) Classification, prediction, n = 20, p = 14.

(c) Regression, prediction, n = 52 (# of weeks in the year), p = 4.

2.3 ex3

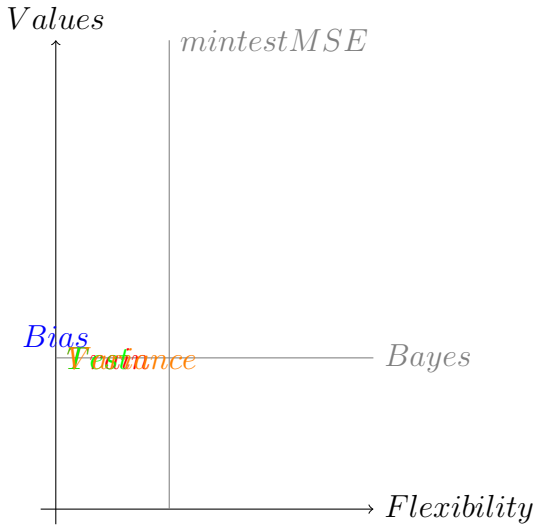
Question: We now revisit the bias-variance decomposition.

(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) The blue line represents the bias curve.
The orange line represents the variance curve.
The red line represents the training error curve.
The green line represents the test error curve.
The horizontal grey line represents the Bayes error curve.
The vertical grey line indicates the flexibility level corresponding to the smallest test MSE.



(b) .

2.4 ex4

Question: 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) The Classification applications:

1. In this case the response is and the predictors are .
2. In this case the response is and the predictors are .
3. In this case the response is and the predictors are .

(b) The Regression applications:

1. In this case the response is and the predictors are .
2. In this case the response is and the predictors are .
3. In this case the response is and the predictors are .

(c) The Cluster analysis applications:

1. In this case the response is and the predictors are .
2. In this case the response is and the predictors are .
3. In this case the response is and the predictors are .

2.5 ex5

Question: 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:

2.6 ex6

Question: 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 non-parametric approach)? What are its disadvantages?

Solution:

2.7 ex7

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

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

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$X_1 = X_2 = X_3 = 0$.

(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 non-linear, then would we expect the best value for K to be large or small? Why?

Solution:

3 Linear Regression

3.1 ex1

Question: Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

Solution:

3.2 ex2

Question: Carefully explain the differences between the KNN classifier and KNN regression methods.

Solution:

3.3 ex3

Question: Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0$ = 50, $\hat{\beta}_1$= 20, $\hat{\beta}_2$ = 0.07, $\hat{\beta}_3$ = 35, $\hat{\beta}_4$ = 0.01, $\hat{\beta}_5$ = -10.

(a) Which answer is correct, and why?

i For a fixed value of IQ and GPA, males earn more on average than females.

ii For a fixed value of IQ and GPA, females earn more on average than males.

iii For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Solution:

3.4 ex4

Question: 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, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. 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 that 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.

Solution:

3.5 ex5

Question: Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form $\hat{y}_i = x_i\hat{\beta}$, where

$$\beta = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right).$$

Show that we can write

$$\hat{y_{i'}} = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

. What is $a_{i'}$?

Note: *We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.*

Solution:

3.6 ex6

Question: Using the least squares coefficient estimates, argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Solution:

3.7 ex7

Question: It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y. Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Solution:

4 Classification

5 Resampling Methods

6 Linear Model Selection and Regularization

7 Moving Beyond Linearity

8 Tree-Based Methods

9 Support Vector Machines

10 Unsupervised Learning

List of Tables

List of Figures