Solutions Manual to

An Introduction to Statistical Learning

with Applications in R

Original Text by

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

이명규 지음
Myeongkyu Lee
mgklee@kaist.ac.kr

Contents

Chapter 2

Statistical Learning

2.7 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, $X_1 = X_2 = X_3 = 0$.

(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.

(a) The Euclidean distance is given by $\sqrt{X_1^2 + X_2^2 + X_3^2}$.

Obs.	Distance
1	3
2	2
3	$\sqrt{10}$
4	$\sqrt{5}$
5	$\sqrt{2}$
6	$\sqrt{3}$

(b) If $K = 1$, we just take the class of the nearest neighbor, Observation 5. Hence, our prediction is **Green**.

(c) The three nearest neighbors are Observations 2, 5, and 6. The Red class occurs twice while the Green class occurs once. Hence, our prediction is **Red**.

(d) According to our textbook, on Page 41, "As K grows, the method becomes less flexible and produces a decision boundary that is close to linear." Therefore, we would expect the best value for K to be **small**.

■

Chapter 3

Linear Regression

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

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

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

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

(b) Predict the salary of a college graduate 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.

(a) Fix X_1 and X_2. We have

$$\begin{aligned}\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_3 X_3 + \widehat{\beta}_4 X_1 X_2 + \widehat{\beta}_5 X_1 X_3 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_4 X_1 X_2 + \left(\widehat{\beta}_3 + \widehat{\beta}_5 X_1\right) X_3 \\ &= \text{const.} + \begin{cases} 35 - 10X_1 & \text{for college graduates,} \\ 0 & \text{for high school graduates.} \end{cases}\end{aligned}$$

If $X_1 > 3.5$, then high school graduates earn more, on average, than college graduates. Therefore, **iii** is correct.

(b) $X_1 = 4.0$ and $X_2 = 110$ gives

$$\widehat{Y} = 50 + 20 \cdot 4.0 + 0.07 \cdot 110 + 35 \cdot 1 + 0.01 \cdot 4.0 \cdot 110 - 10 \cdot 4.0 \cdot 1 = 137.1.$$

(c) **False**. The scale of each predictor matters. The value of the coefficient itself does not give much information about an interaction effect.

■

3.7 It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\overline{x} = \overline{y} = 0$.

Solution. Without loss of generality, we may assume that $\overline{x} = \overline{y} = 0$. By (3.4), $\widehat{\beta} = \dfrac{X \cdot Y}{\|X\|_2^2}$. By (3.17) and (3.18),

$$\begin{aligned} R^2 &= \frac{\|Y\|_2^2 - \left\|Y - X\widehat{\beta}\right\|_2^2}{\|Y\|_2^2} \\ &= \frac{2\widehat{\beta} X \cdot Y - \widehat{\beta}^2 \|X\|_2^2}{\|Y\|_2^2} \\ &= \frac{(X \cdot Y)^2}{\|X\|_2^2 \|Y\|_2^2} \\ &= \mathrm{Cor}(X, Y)^2. \end{aligned}$$

■

Chapter 4

Classification

4.10 Equation 4.32 derived an expression for $\log \dfrac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}$ in the setting where $p>1$, so that the mean for the kth class, μ_k, is a p-dimensional vector, and the shared covariance $\mathbf{\Sigma}$ is a $p \times p$ matrix. However, in the setting with $p=1$, (4.32) takes a simpler form, since the means $\mu_1, \cdots, \mu_K$ and the variance σ^2 are scalars. In this simpler setting, repeat the calculation in (4.32), and provide expressions for a_k and b_{kj} in terms of π_k, π_K, μ_k, μ_K, and σ^2.

Solution. Repeating the calculation in (4.32) gives

$$\begin{aligned}
\log \frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)} &= \log \frac{\pi_k f_k(x)}{\pi_K f_K(x)} \\
&= \log \frac{\pi_k}{\pi_K} - \frac{1}{2\sigma^2}\left[(x-\mu_k)^2 - (x-\mu_K)^2\right] \\
&= \log \frac{\pi_k}{\pi_K} - \frac{\mu_k^2 - \mu_K^2}{2\sigma^2} + \frac{\mu_k - \mu_K}{\sigma^2} \cdot x.
\end{aligned}$$

Therefore, the coefficients are given by

$$a_k = \log \frac{\pi_k}{\pi_K} - \frac{\mu_k^2 - \mu_K^2}{2\sigma^2} \quad \text{and} \quad b_k = \frac{\mu_k - \mu_K}{\sigma^2}.$$

■

4.11 Work out the detailed forms of a_k, b_{kj}, and c_{kjl} in (4.33). Your answer should involve π_k, π_K, μ_k, μ_K, $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_K$.

Solution. By similar calculations,

$$\begin{aligned}
&\log \frac{\Pr(Y=k \mid X=\mathbf{x})}{\Pr(Y=K \mid X=\mathbf{x})} = \log \frac{\pi_k f_k(\mathbf{x})}{\pi_K f_K(\mathbf{x})} \\
&= \log \frac{\pi_k}{\pi_K} - \log \frac{|\mathbf{\Sigma}_k|^{1/2}}{|\mathbf{\Sigma}_K|^{1/2}} - \frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) - (\mathbf{x}-\boldsymbol{\mu}_K)^\top \mathbf{\Sigma}_K^{-1}(\mathbf{x}-\boldsymbol{\mu}_K)\right] \\
&= \log \frac{\pi_k}{\pi_K} - \frac{1}{2}\log \frac{|\mathbf{\Sigma}_k|}{|\mathbf{\Sigma}_K|} \\
&\quad - \frac{1}{2}\left[\mathbf{x}^\top \left(\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_K^{-1}\right)\mathbf{x} - 2\mathbf{x}^\top \left(\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \mathbf{\Sigma}_K^{-1}\boldsymbol{\mu}_K\right) + \boldsymbol{\mu}_k^\top \mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_K^\top \mathbf{\Sigma}_K^{-1}\boldsymbol{\mu}_K\right].
\end{aligned}$$

Therefore, the coefficients are given by

$$
\begin{aligned}
a_k &= \log \frac{\pi_k}{\pi_K} - \frac{1}{2}\left(\log \frac{|\boldsymbol{\Sigma}_k|}{|\boldsymbol{\Sigma}_K|} + \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_K^\top \boldsymbol{\Sigma}_K^{-1}\boldsymbol{\mu}_K\right), \\
b_{kj} &= \left(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \boldsymbol{\Sigma}_K^{-1}\boldsymbol{\mu}_K\right)_j, \\
c_{kjl} &= -\frac{1}{2}\left(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_K^{-1}\right)_{jl}.
\end{aligned}
$$

■

Chapter 5

Resampling Methods

5.1 Using basic statistical properties of the variance, as well as single-variable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize $\mathrm{Var}(\alpha X + (1-\alpha)Y)$.

Solution. We assume $X \neq Y$. Let

$$\begin{aligned} f(\alpha) &= \mathrm{Var}(\alpha X + (1-\alpha)Y) \\ &= \alpha^2 \,\mathrm{Var}(X) + (1-\alpha)^2 \,\mathrm{Var}(Y) + 2\alpha(1-\alpha)\,\mathrm{Cov}(X,Y). \end{aligned}$$

Since

$$\begin{aligned} f'(\alpha) &= 2\alpha \,\mathrm{Var}(X) - 2(1-\alpha)\,\mathrm{Var}(Y) + 2(1-2\alpha)\,\mathrm{Cov}(X,Y) \\ &= 2\alpha[\mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y)] - 2\,\mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y), \end{aligned}$$

the only critical point of f is

$$\alpha^* = \frac{\mathrm{Var}(Y) - \mathrm{Cov}(X,Y)}{\mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y)}.$$

Because $f''(\alpha^*) = 2[\mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y)] = 2\,\mathrm{Var}(X-Y) > 0$, α^* does indeed minimize $\mathrm{Var}(\alpha X + (1-\alpha)Y)$. ■

Chapter 6

Linear Model Selection and Regularization

6.4 Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase λ from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

Solution. Recall the bias-variance trade-off from Equation (2.7) that

$$\mathbb{E}\left(y_0-\widehat{f}(x_0)\right)^2=\mathrm{Var}\left(\widehat{f}(x_0)\right)+\left[\mathrm{Bias}\left(\widehat{f}(x_0)\right)\right]^2+\mathrm{Var}(\epsilon).$$

(a) **iii.** The training RSS is minimized by the least squares coefficient estimates, that is, when $\lambda=0$. As λ increases, the ridge coefficient estimates shrink towards zero, so the training RSS can only increase.

(b) **ii.** See Figure 6.5. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias. When $\lambda\to\infty$, then all of the ridge coefficient estimates converge to zero; this corresponds to the *null model* that contains no predictors, that is, the model predicts with only the intercept β_0.

Initially, a rapid decrease in variance outweighs a tiny increase in bias, so the test RSS decreases. Beyond some point, (squared) bias grows much faster than variance declines. Eventually, the test RSS will start increasing. This reasoning also applies to (c) and (d).

(c) **iv.**

(d) **iii.**

(e) **v.** The irreducible error is independent of the model.

■

Chapter 7

Moving Beyond Linearity

7.1 It was mentioned in the chapter that a cubic regression spline with one knot at ξ can be obtained using a basis of the form x, x^2, x^3, $(x-\xi)^3_+$, where $(x-\xi)^3_+ = (x-\xi)^3$ if $x > \xi$ and equals 0 otherwise. We will now show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3_+$$

is indeed a cubic regression spline, regardless of the values of β_0, β_1, β_2, β_3, β_4.

(a) Find a cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, d_1 in terms of β_0, β_1, β_2, β_3, β_4.

(b) Find a cubic polynomial

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$

such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of β_0, β_1, β_2, β_3, β_4. We have now established that $f(x)$ is a piecewise polynomial.

(c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ.

(d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continuous at ξ.

(e) Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at ξ.

Therefore, $f(x)$ is indeed a cubic spline.

Hint: Parts (d) and (e) of this problem require knowledge of single-variable calculus. As a reminder, given a cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3,$$

the first derivative takes the form

$$f_1'(x) = b_1 + 2c_1 x + 3d_1 x^2,$$

and the second derivative takes the form

$$f_1''(x) = 2c_1 + 6d_1 x.$$

Solution.

(a) For all $x \leq \xi$,

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3 \\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, \\ f_1(x) &= a_1 + b_1 x + c_1 x^2 + d_1 x^3. \end{aligned}$$

Hence, $a_1 = \beta_0$, $b_1 = \beta_1$, $c_1 = \beta_2$, and $d_1 = \beta_3$.

(b) For all $x > \xi$,

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3 \\ &= \beta_0 - \beta_4 \xi^3 + \left(\beta_1 + 3\beta_4 \xi^2\right) x + (\beta_2 - 3\beta_4 \xi) x^2 + (\beta_3 + \beta_4) x^3, \\ f_2(x) &= a_2 + b_2 x + c_2 x^2 + d_2 x^3. \end{aligned}$$

Hence, $a_2 = \beta_0 - \beta_4 \xi^3$, $b_2 = \beta_1 + 3\beta_4 \xi^2$, $c_2 = \beta_2 - 3\beta_4 \xi$, and $d_2 = \beta_3 + \beta_4$.

(c) $f_1(\xi) = \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3$ and

$$\begin{aligned} f_2(\xi) &= \beta_0 - \beta_4 \xi^3 + \left(\beta_1 + 3\beta_4 \xi^2\right) \xi + (\beta_2 - 3\beta_4 \xi) \xi^2 + (\beta_3 + \beta_4) \xi^3 \\ &= \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3. \end{aligned}$$

(d) $f_1'(\xi) = \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2$ and

$$\begin{aligned} f_2'(\xi) &= \beta_1 + 3\beta_4 \xi^2 + 2(\beta_2 - 3\beta_4 \xi) \xi + 3(\beta_3 + \beta_4) \xi^2 \\ &= \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2. \end{aligned}$$

(e) $f_1''(\xi) = 2\beta_2 + 6\beta_3 \xi$ and $f_2''(\xi) = 2(\beta_2 - 3\beta_4 \xi) + 6(\beta_3 + \beta_4) \xi = 2\beta_2 + 6\beta_3 \xi$.

■

7.2 Suppose that a curve $\widehat{g}$ is computed to smoothly fit a set of n points using the following formula:

$$\widehat{g} = \underset{g}{\arg\min} \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(m)}(x) \right]^2 dx \right),$$

where $g^{(m)}$ represents the mth derivative of g (and $g^{(0)} = g$). Provide example sketches of $\widehat{g}$ in each of the following scenarios.

(a) $\lambda = \infty$, $m = 0$.

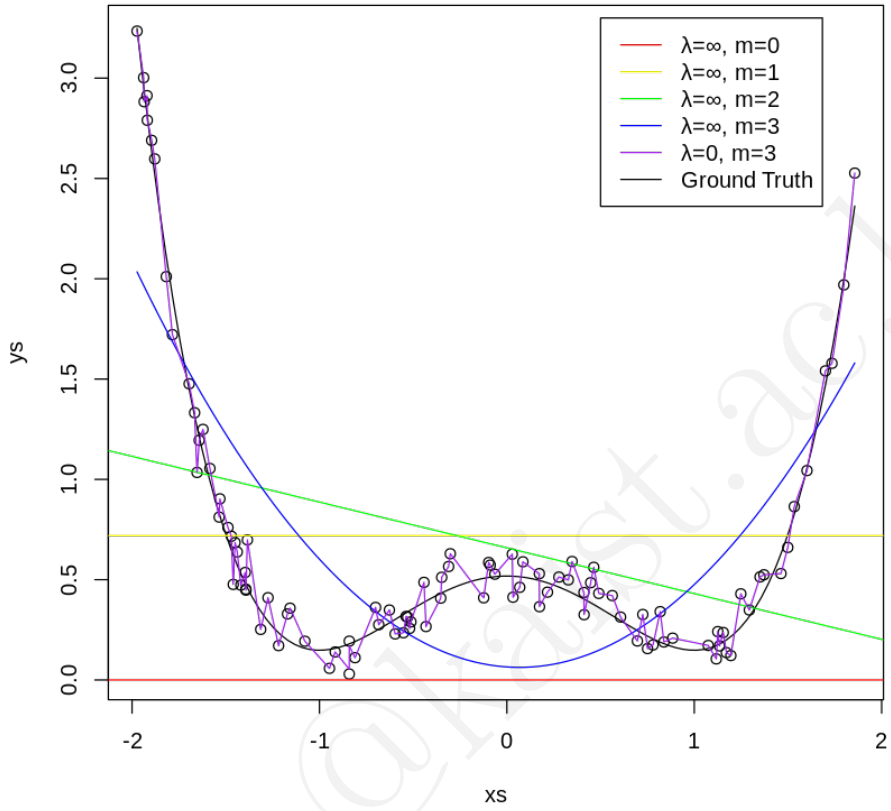
(b) $\lambda = \infty$, $m = 1$.

(c) $\lambda = \infty$, $m = 2$.

(d) $\lambda = \infty$, $m = 3$.

(e) $\lambda = 0$, $m = 3$.

Solution. $\lambda = \infty$ forces $g^{(m)}(x) = 0$, so we can fit a polynomial curve of degree $m-1$ for $m \geq 1$. ($m = 0$ simply implies $g(x) = 0$.) For $\lambda = 0$, we can interpolate the set of n points.



■

7.4 Suppose we fit a curve with basis functions

$$b_1(X) = I(0 \leq X \leq 2) - (X-1)I(1 \leq X \leq 2),$$
$$b_2(X) = (X-3)I(3 \leq X \leq 4) + I(4 < X \leq 5).$$

We fit the linear regression model

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon,$$

and obtain coefficient estimates $\widehat{\beta}_0 = 1$, $\widehat{\beta}_1 = 1$, $\widehat{\beta}_2 = 3$. Sketch the estimated curve between $X = -2$ and $X = 6$. Note the intercepts, slopes, and other relevant information.

Solution. The estimated curve is given by

$$\widehat{Y} = \begin{cases} 2 & \text{if } 0 \leq X \leq 1, \\ 3 - X & \text{if } 1 < X \leq 2, \\ 3X - 8 & \text{if } 3 \leq X \leq 4, \\ 4 & \text{if } 4 < X \leq 5, \\ 1 & \text{elsewhere.} \end{cases}$$

■

Chapter 8

Tree-Based Methods

8.2 It is mentioned in Section 8.2.3 that boosting using depth-one trees (or stumps) leads to an *additive* model: that is, a model of the form

$$f(X) = \sum_{j=1}^{p} f_j(X_j).$$

Explain why this is the case. You can begin with (8.12) in Algorithm 8.2.

Solution. Each stump $\widehat{f}^b$ has two terminal nodes, so it is of the form

$$\widehat{f}^b(X_{j_b}) = \widehat{y}_{b_1} I(X_{j_b} < s_b) + \widehat{y}_{b_2} I(X_{j_b} \geq s_b).$$

By (8.12) in Algorithm 8.2,

$$\begin{aligned}
\widehat{f}(X) &= \lambda \sum_{b=1}^{B} \widehat{f}^b(X_{j_b}) \\
&= \lambda \sum_{b=1}^{B} \left[\widehat{y}_{b_1} I(X_{j_b} < s_b) + \widehat{y}_{b_2} I(X_{j_b} \geq s_b) \right] \\
&= \lambda \sum_{j=1}^{p} \sum_{b \in \{i | j_i = j\}} \left[\widehat{y}_{b_1} I(X_{j_b} < s_b) + \widehat{y}_{b_2} I(X_{j_b} \geq s_b) \right].
\end{aligned}$$

■

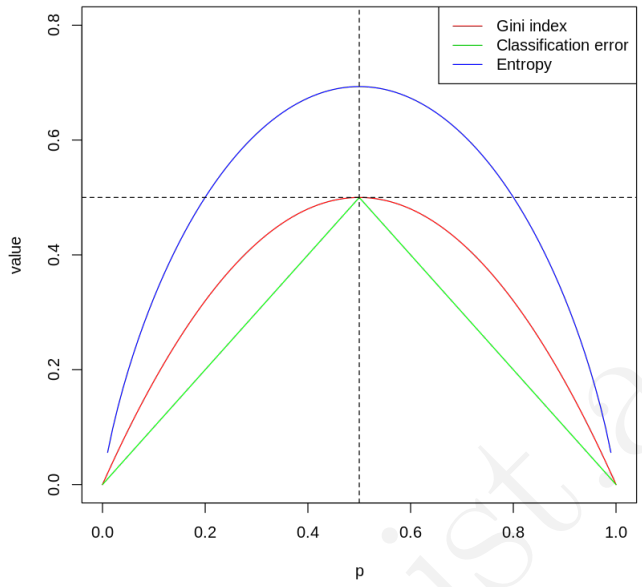
8.3 Consider the Gini index, classification error, and entropy in a simple classification setting with two classes. Create a single plot that displays each of these quantities as a function of $\widehat{p}_{m1}$. The x-axis should display $\widehat{p}_{m1}$, ranging from 0 to 1, and the y-axis should display the value of the Gini index, classification error, and entropy.
Hint: In a setting with two classes, $\widehat{p}_{m1} = 1 - \widehat{p}_{m2}$. You could make this plot by hand, but it will be much easier to make in `R`.

Solution. For the binary classification, the Gini index G, classification error E, and entropy D are, respectively,

$$\begin{aligned}
G &= \widehat{p}_{m1}\left(1 - \widehat{p}_{m1}\right) + \widehat{p}_{m2}\left(1 - \widehat{p}_{m2}\right) = 2\widehat{p}_{m1}\left(1 - \widehat{p}_{m1}\right), \\
E &= 1 - \max\left\{\widehat{p}_{m1}, \widehat{p}_{m2}\right\} = \min\left\{\widehat{p}_{m1}, 1 - \widehat{p}_{m1}\right\},
\end{aligned}$$

$$D = -\widehat{p}_{m1}\log\widehat{p}_{m1} - \widehat{p}_{m2}\log\widehat{p}_{m2} = -\widehat{p}_{m1}\log\widehat{p}_{m1} - (1-\widehat{p}_{m1})\log(1-\widehat{p}_{m1}).$$

The plot is as follows.



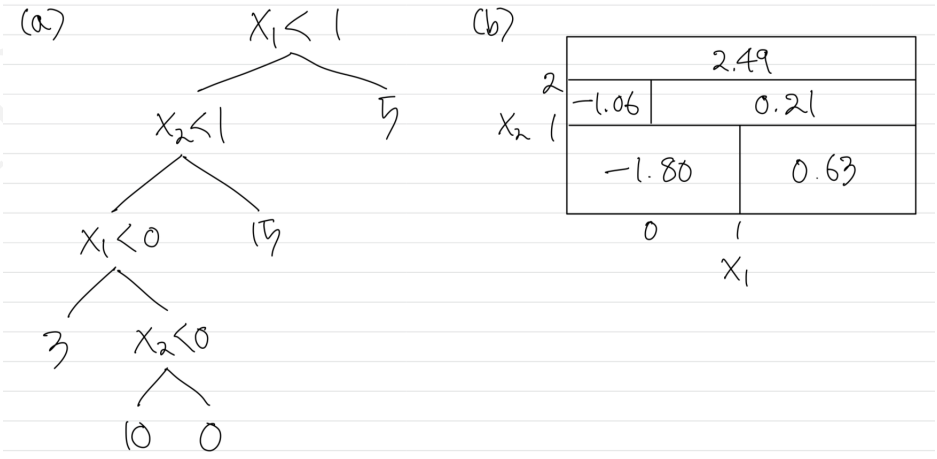
■

8.4 This question relates to the plots in Figure 8.14.

(a) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 8.14. The numbers inside the boxes indicate the mean of Y within each region.

(b) Create a diagram similar to the left-hand panel of Figure 8.14, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

Solution.



■

8.5 Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(\text{Class is Red} \mid X)$:

$$0.1,\ 0.15,\ 0.2,\ 0.2,\ 0.55,\ 0.6,\ 0.6,\ 0.65,\ 0.7,\ \text{and } 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

Solution.

1. Consider the majority vote—four votes for the green and six votes for the red. Hence, the final classification is **red**.

2. The average probability is

$$\frac{0.1 + 0.15 + 0.2 + 0.2 + 0.55 + 0.6 + 0.6 + 0.65 + 0.7 + 0.75}{10} = 0.45.$$

   Hence, the final classification is **green**.

■

Chapter 9

Support Vector Machines

9.2 We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0+\beta_1 X_1+\beta_2 X_2 = 0$. We now investigate a non-linear decision boundary.

(a) Sketch the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

(b) On your sketch, indicate the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 \leq 4.$$

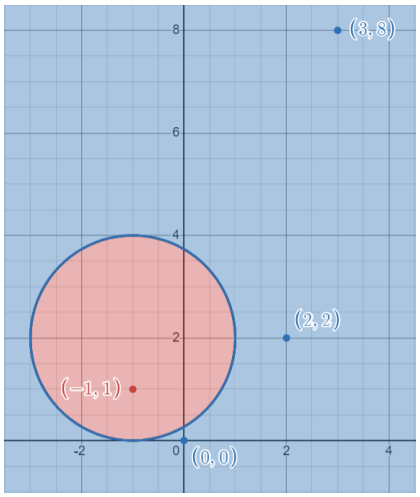
(c) Suppose that a classifier assigns an observation to the blue class if

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

and to the red class otherwise. To what class is the observation $(0, 0)$ classified? $(-1, 1)$? $(2, 2)$? $(3, 8)$?

(d) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_1^2, X_2, and X_2^2.

Solution. For parts (a) through (c), the desired plot is as follows.



For part (d), the decision boundary is

$$\begin{aligned}(1+X_1)^2+(2-X_2)^2=4 &\iff X_1^2+2X_1+X_2^2-4X_2+1=0\\ &\iff \begin{bmatrix}1 & 2 & 1 & -4 & 1\end{bmatrix}^\top \begin{bmatrix}1 & X_1 & X_1^2 & X_2 & X_2^2\end{bmatrix}=0,\end{aligned}$$

which is obviously linear in terms of X_1, X_1^2, X_2, and X_2^2, but not in terms of X_1 and X_2. ■

Chapter 10

Deep Learning

10.4 Consider a CNN that takes in 32×32 grayscale images and has a single convolution layer with three 5×5 convolution filters (without boundary padding).

(a) Draw a sketch of the input and first hidden layer similar to Figure 10.8.

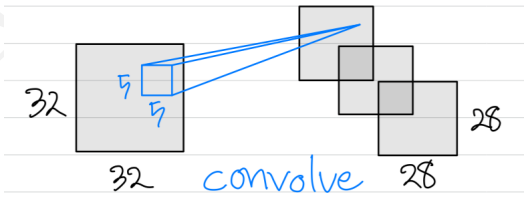
(b) How many parameters are in this model?

(c) Explain how this model can be thought of as an ordinary feed-forward neural network with the individual pixels as inputs, and with constraints on the weights in the hidden units. What are the constraints?

(d) If there were no constraints, then how many weights would there be in the ordinary feed-forward neural network in (c)?

Solution.

(a) Without boundary padding, the dimension of hidden layers is 28×28. $(32 - 5 + 1 = 28)$



(b) Counting bias terms, there are $(5 \cdot 5 + 1) \cdot 3 = 78$ parameters.

(c) There are $32 \cdot 32 = 1024$ nodes in the input layer and $28 \cdot 28 \cdot 3 = 2352$ nodes in the hidden layer. Each node in the hidden layer is connected to $5 \cdot 5 = 25$ nodes in the input layer, and all weights (including the bias) are the same throughout the part of the hidden layer. (3 parts in the hidden layer)

(d) If the network is fully connected, then there would be $(1024 + 1) \cdot 2352 = 2\,410\,800$ weights, including bias terms.

■

Chapter 11

Survival Analysis and Censored Data

11.4 This problem makes use of the Kaplan–Meier survival curve displayed in Figure 11.9. The raw data that went into plotting this survival curve is given in Table 11.4. The covariate column of that table is not needed for this problem.

(a) What is the estimated probability of survival past 50 days?

(b) Write out an analytical expression for the estimated survival function. For instance, your answer might be something along the lines of

$$\widehat{S}(t) = \begin{cases} 0.8 & \text{if } t < 31, \\ 0.5 & \text{if } 31 \le t < 77, \\ 0.22 & \text{if } 77 \le t. \end{cases}$$

(The previous equation is for illustration only: it is not the correct answer!)

Solution.

(a) The estimated probability of survival past 50 days is $4/5 \cdot 3/4 = 3/5$.

(b) According to Table 11.4, we have the following table:

Y	q_k	r_k	$\widehat{S}(t)$
20.2	0	6	1
26.5	1	5	4/5
37.2	1	4	$4/5 \cdot 3/4 = 3/5$
57.3	1	3	$3/5 \cdot 2/3 = 2/5$
89.8	0	2	2/5
90.8	0	1	2/5

Therefore, the estimated survival function is

$$\widehat{S}(t) = \begin{cases} 1 & \text{if } t < 26.5, \\ 0.8 & \text{if } 26.5 \le t < 37.2, \\ 0.6 & \text{if } 37.2 \le t < 57.3, \\ 0.4 & \text{if } 57.3 \le t. \end{cases}$$

■