

Data Mining: Theories, Algorithms, and Examples

SOLUTION MANUAL

Chapter 2

Exercise 2.1

Equation 2.25 is used to estimate the parameters of the linear regression model in equation 2.23 or 2.24:

$$y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i. \quad (2.23)$$

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.24)$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}'\mathbf{y}) \quad (2.25)$$

For the linear regression model in this exercise:

$$y_i = \beta_0 + \beta_1 \sqrt{x_i} + \varepsilon_i,$$

we define:

$$x_{i,1} = \sqrt{x_i} = \sqrt{\text{Launch Temperature}}$$

$$y_i = \text{Number of } O - \text{rings with stress}.$$

Using the data set in Table 2.1, we have:

$$\mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} 1 & 8.12 \\ 1 & 8.37 \\ 1 & 8.31 \\ 1 & 8.25 \\ 1 & 8.19 \\ 1 & 8.49 \\ 1 & 8.54 \\ 1 & 8.37 \\ 1 & 7.55 \\ 1 & 7.94 \\ 1 & 8.37 \\ 1 & 8.83 \\ 1 & 8.19 \\ 1 & 7.28 \\ 1 & 8.18 \\ 1 & 8.66 \\ 1 & 8.37 \\ 1 & 9.00 \\ 1 & 8.72 \\ 1 & 8.89 \\ 1 & 8.66 \\ 1 & 8.72 \\ 1 & 7.62 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}'\mathbf{y}) = \begin{bmatrix} 23 & 191.59 \\ 191.59 & 1600 \end{bmatrix}^{-1} \begin{bmatrix} 7 \\ 54.40 \end{bmatrix} \\ &= \begin{bmatrix} 16.99 & -2.03 \\ -2.03 & 0.24 \end{bmatrix} \begin{bmatrix} 7 \\ 54.40 \end{bmatrix} = \begin{bmatrix} 8.27 \\ -0.96 \end{bmatrix}. \end{aligned}$$

$$y_i = \beta_0 + \beta_1\sqrt{x_i} + \varepsilon_i = 8.27 - 0.96\sqrt{x_i}.$$

SSE is computed using Equation 2.6:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Instance	Launch	Number of O-rings	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
	Temperature	with Stress		
	x_i	y_i		
1	66	0	0.47	58.57

2	70	1	0.24	66.07
3	69	0	0.30	64.18
4	68	0	0.35	62.29
5	67	0	0.41	60.42
6	72	0	0.12	69.91
7	73	0	0.07	71.85
8	70	0	0.24	66.07
9	57	1	1.02	42.61
10	63	1	0.65	53.10
11	70	1	0.24	66.07
12	78	0	-0.21	81.73
13	67	0	0.41	60.42
14	53	2	1.28	35.99
15	67	0	0.41	60.42
16	75	0	-0.04	75.76
17	70	0	0.24	66.07
18	81	0	-0.37	87.80

19	76	0	-0.10	77.73
20	79	0	-0.26	83.74
21	75	0	-0.04	75.76
22	76	0	-0.10	77.74
23	58	1	0.96	44.31

$SSE = 1508.63$

Exercise 2.2

Equations 2.11 and 2.12 are used to estimate the parameters of the linear regression model in equation 2.1:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.1}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \tag{2.11}$$

$$\hat{\beta}_0 = \frac{1}{n} (\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i) = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.12}$$

For this data set, we define:

$x_i = \sqrt{\text{Launch Temperature}}$

$y_i = \text{Number of O – rings with stress.}$

Using the data set in Table 2.1, we have:

Instance	$\sqrt{\text{Launch Temperature}}$	Number of	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
		O-rings				

1	8.12	0	-0.21	-0.30	0.06	0.04
2	8.37	1	0.04	0.70	0.03	0.00
3	8.31	0	-0.02	-0.30	0.01	0.00
4	8.25	0	-0.08	-0.30	0.03	0.01
5	8.19	0	-0.14	-0.30	0.04	0.02
6	8.49	0	0.16	-0.30	-0.05	0.02
7	8.54	0	0.21	-0.30	-0.06	0.05
8	8.37	0	0.04	-0.30	-0.01	0.00
9	7.55	1	-0.78	0.70	-0.55	0.61
10	7.94	1	-0.39	0.70	-0.27	0.15
11	8.37	1	0.04	0.70	0.03	0.00
12	8.83	0	0.51	-0.30	-0.15	0.25
13	8.19	0	-0.14	-0.30	0.04	0.02
14	7.28	2	-1.05	1.70	-1.78	1.10
15	8.18	0	-0.14	-0.30	0.04	0.02
16	8.66	0	0.33	-0.30	-0.10	0.11
17	8.37	0	0.04	-0.30	-0.01	0.00

18	9.00	0	0.67	-0.30	-0.20	0.45
19	8.72	0	0.39	-0.30	-0.12	0.15
20	8.89	0	0.56	-0.30	-0.17	0.31
21	8.66	0	0.33	-0.30	-0.10	0.11
22	8.72	0	0.39	-0.30	-0.11	0.15
23	7.62	1	-0.71	0.70	-0.50	0.51
Sum	191.59	7				-3.91
Average	$\bar{x} = 8.33$	$\bar{y} = 0.30$				

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{-3.91}{4.09} = -0.96$$
$$\hat{\beta}_0 = \frac{1}{n} (\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i) = \bar{y} - \hat{\beta}_1 \bar{x} = 0.30 - (-0.96)(8.33) = 8.30.$$

$y_i = \beta_0 + \beta_1 \sqrt{x_i} + \varepsilon_i = 8.30 - 0.96 \sqrt{x_i}$, which is similar to one in Exercise 2.1. Hence, SSE values are also similar to those in Exercise 2.1.

Chapter 3

Exercise 3.1

Given the balloon data set in Table 1.1 for classifying the target variable, y (Inflated = T or F), from the attribute variables, x_1 (Color = Yellow or Purple), x_2 (Size = Small or Large), x_3 (Act = Stretch or Dip), x_4 (Age = Adult or Child),

TABLE 1.1
The Balloon Data Set

Instance	Attribute Variables				Target Variable
	Color	Size	Act	Age	Inflated
1	Yello	Small	Stretch	Adult	T
	w				
2	Yello	Small	Stretch	Child	T
	w				
3	Yello	Small	Dip	Adult	T
	w				
4	Yello	Small	Dip	Child	T
	w				
5	Yello	Large	Stretch	Adult	T
	w				
6	Yello	Large	Stretch	Child	F
	w				
7	Yello	Large	Dip	Adult	F
	w				

8	Yello	Large	Dip	Child	F
	w				
9	Purple	Small	Stretch	Adult	T
10	Purple	Small	Stretch	Child	F
11	Purple	Small	Dip	Adult	F
12	Purple	Small	Dip	Child	F
13	Purple	Large	Stretch	Adult	T
14	Purple	Large	Stretch	Child	F
15	Purple	Large	Dip	Adult	F
16	Purple	Large	Dip	Child	F

we train a naïve Bayes classifier as follows.

$$n = 16$$

$$n_{y=T} = 7 \quad n_{y=F} = 9$$

$$n_{y=T \& x_1=Yellow} = 5 \quad n_{y=T \& x_1=Purple} = 2 \quad n_{y=F \& x_1=Yellow} = 3 \quad n_{y=F \& x_1=Purple} = 6$$

$$n_{y=T \& x_2=Small} = 5 \quad n_{y=T \& x_2=Large} = 2 \quad n_{y=F \& x_2=Small} = 3 \quad n_{y=F \& x_2=Large} = 6$$

$$n_{y=T \& x_3=Stretch} = 5 \quad n_{y=T \& x_3=Dip} = 2 \quad n_{y=F \& x_3=Stretch} = 3 \quad n_{y=F \& x_3=Dip} = 6$$

$$n_{y=T \& x_4=Adult} = 5 \quad n_{y=T \& x_4=Child} = 2 \quad n_{y=F \& x_4=Adult} = 3 \quad n_{y=F \& x_4=Child} = 6$$

Instance #1 in Table 1.1 with $\mathbf{x} = (\text{Yellow}, \text{Small}, \text{Stretch}, \text{Adult})$ is classified as follows:

$$\begin{aligned} p(y=T) \prod_{i=1}^4 P(x_i|y=T) &= \frac{n_{y=T}}{n} \prod_{i=1}^4 \frac{n_{y=T \& x_i}}{n_{y=T}} \\ &= \frac{n_{y=T}}{n} \frac{n_{y=T \& x_1=Yellow}}{n_{y=T}} \frac{n_{y=T \& x_2=Small}}{n_{y=T}} \frac{n_{y=T \& x_3=Stretch}}{n_{y=T}} \frac{n_{y=T \& x_4=Adult}}{n_{y=T}} \\ &= \frac{7}{16} \left(\frac{5}{7} \times \frac{5}{7} \times \frac{5}{7} \times \frac{5}{7} \right) = 0.1139 \end{aligned}$$

$$\begin{aligned} p(y=F) \prod_{i=1}^4 P(x_i|y=F) &= \frac{n_{y=F}}{n} \prod_{i=1}^4 \frac{n_{y=F \& x_i}}{n_{y=T}} \\ &= \frac{n_{y=F}}{n} \frac{n_{y=F \& x_1=Yellow}}{n_{y=F}} \frac{n_{y=F \& x_2=Small}}{n_{y=F}} \frac{n_{y=F \& x_3=Stretch}}{n_{y=F}} \frac{n_{y=F \& x_4=Adult}}{n_{y=F}} \\ &= \frac{9}{16} \left(\frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \right) = 0.0069 \end{aligned}$$

$y_{MAP} \approx \arg \max_{y \in Y} p(y) \prod_{i=1}^4 P(x_i|y) = T$, which is the correct classification.

Instance #2 in Table 1.1 with $\mathbf{x} = (\text{Yellow}, \text{Small}, \text{Stretch}, \text{Child})$ is classified as follows:

$$\begin{aligned} p(y=T) \prod_{i=1}^4 P(x_i|y=T) &= \frac{n_{y=T}}{n} \prod_{i=1}^4 \frac{n_{y=T \& x_i}}{n_{y=T}} \\ &= \frac{n_{y=T}}{n} \frac{n_{y=T \& x_1=Yellow}}{n_{y=T}} \frac{n_{y=T \& x_2=Small}}{n_{y=T}} \frac{n_{y=T \& x_3=Stretch}}{n_{y=T}} \frac{n_{y=T \& x_4=Child}}{n_{y=T}} \\ &= \frac{7}{16} \left(\frac{5}{7} \times \frac{5}{7} \times \frac{5}{7} \times \frac{2}{7} \right) = 0.0456 \end{aligned}$$

$$\begin{aligned} p(y = F) \prod_{i=1}^4 P(x_i|y = F) &= \frac{n_{y=F}}{n} \prod_{i=1}^4 \frac{n_{y=F \& x_i}}{n_{y=F}} \\ &= \frac{n_{y=F}}{n} \frac{n_{y=F \& x_1=Yellow}}{n_{y=F}} \frac{n_{y=F \& x_2=Small}}{n_{y=F}} \frac{n_{y=F \& x_3=Stretch}}{n_{y=F}} \frac{n_{y=F \& x_4=Child}}{n_{y=F}} \\ &= \frac{9}{16} \left(\frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{6}{9} \right) = 0.0138 \end{aligned}$$

$y_{MAP} \approx \arg \max_{y \in Y} p(y) \prod_{i=1}^4 P(x_i|y) = T$, which is the correct classification.

The computation for other instances is shown in the following table.

Attribute Variables					Target Variable (Inflated)	
Instance	Color	Size	Act	Age	True Value	Classified Value
1	Yello	Small	Stretch	Adult	T	$p(y = T) \prod_{i=1}^4 P(x_i y = T)$ $= \frac{7}{16} \left(\frac{5}{7} \times \frac{5}{7} \times \frac{5}{7} \times \frac{5}{7} \right) = 0.1139$ $p(y = F) \prod_{i=1}^4 P(x_i y = F)$ $= \frac{9}{16} \left(\frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \right) = 0.0069$ $y_{MAP} = T$
2	Yello	Small	Stretch	Child	T	$p(y = T) \prod_{i=1}^4 P(x_i y = T)$