1

(a) Develop a linear classification function for the data in Example 11.1 using (11-19).

(b) Using the function in (a) and (11-20), construct the "confusion matrix" by classifying the given observations. Compare your classification results with those of Figure 11.1, where the classification regions were determined "by eye." (See Example 11.6.)

(c) Given the results in (b), calculate the apparent error rate (APER).

(d) State any assumptions you make to justify the use of the method in Parts a and b.

(a) 根據計算可得

$$\overline{x}_1 = \begin{bmatrix} 109.48 \\ 20.27 \end{bmatrix} \quad \overline{x}_2 = \begin{bmatrix} 87.43 \\ 17.63 \end{bmatrix}$$

$$S_{Pooled} = \begin{bmatrix} 276.994 & -7.190 \\ -7.190 & 4.273 \end{bmatrix}$$

$$S_{Pooled}^{-1} = \begin{bmatrix} 0.00377504 & 0.00635134 \\ 0.00635134 & 0.24469522 \end{bmatrix}$$

根據 Fisher 的定義可得

$$\begin{aligned} \hat{y} = (\overline{x}_1 - \overline{x}_2)^T S_{Pooled}^{-1} x &= \begin{bmatrix} 22.05 & 2.64 \end{bmatrix} \begin{bmatrix} 0.00377504 & 0.00635134 \\ 0.00635134 & 0.24469522 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 0.1x_1 + 0.786x_2 \end{aligned}$$

$$\begin{aligned} \hat{m} = \frac{1}{2}(\overline{x}_1 - \overline{x}_2)^T S_{Pooled}^{-1} (\overline{x}_1 + \overline{x}_2) &= \frac{1}{2} \begin{bmatrix} 22.05 & 2.64 \end{bmatrix} \begin{bmatrix} 0.00377504 & 0.00635134 \\ 0.00635134 & 0.24469522 \end{bmatrix} \begin{bmatrix} 196.91 \\ 37.9 \end{bmatrix} \\ &= 24.719 \end{aligned}$$

因此可得 linear classification function 為

$$\hat{w} = \hat{y} - \hat{w} = 0.1x_1 + 0.786x_2 - 24.719$$

(b) 透過 Fisher's method 可以建立以下分類規則

$$\begin{cases} \hat{w}_0 \geq 0, \ x_0 \in \pi_1 \\ \hat{w}_0 < 0, \ x_0 \in \pi_2 \end{cases}$$

透過計算 24 個 observations 可得其 confusion matrix 為 跟 Figure 11.1 比

Actual \ Predicted	π_1	π_2
π_1	11	1
π_2	2	10

較可以發現分類結果大致上一樣，只有一個樣本與我們建立的分類規則所得出的結果有差異。

(c) 根據定義可計算 apparent error rate 為

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \times 100\% = \frac{1+2}{12+12} \times 100\% = 12.5\%$$

(d) 該分類法必須假設 π_1, π_2 來自 Multivariate Normal Distribution，且有相同的 Covariance matrix，意即 $\Sigma_1 = \Sigma_2 = \Sigma$

2

A researcher wants to determine a procedure for discriminating between two multivariate populations. The researcher has enough data available to estimate the density functions $f_1(\mathrm{x})$ and $f_2(\mathrm{x})$ associated with populations π_1 and π_2, respectively. Let $c(2|1) = 50$ (this is the cost of assigning items as π_2, given that π_1 is true) and $c(1|2) = 100$.
In addition, it is known that about 20% of all possible items (for which the

measurements x can be recorded) belong to π_2.

(a) Give the minimum ECM rule (in general form) for assigning a new item to one of the two populations.

(b) Measurements recorded on a new item yield the density values $f_1(\mathrm{x}) = .3$ and $f_2(\mathrm{x}) = .5$. Given the preceding information, assign this item to population π_1 or population π_2.

根據題意可得以下資訊

- $c(2|1) = 50, \quad c(1|2) = 100$
- $p_1 = 0.8, \quad p_2 = 0.2$

(a) 根據 minimum ECM，我們將 observation x 分類到 π_1，如果滿足

$$\frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} = \frac{100}{50} \cdot \frac{0.2}{0.8} = 0.5$$

反之，分類到 π_2，若滿足

$$\frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} = 0.5$$

(b) 將 $f_1(\mathrm{x}) = .3$ 和 $f_2(\mathrm{x}) = .5$ 代入可得

$$\frac{f_1(x)}{f_2(x)} = 0.6 > 0.5$$

因此將該 observation 分類到 π_1

3

Suppose that $n_1 = 11$ and $n_2 = 12$ observations are made on two random variables X_1 and X_2, where X_1 and X_2 are assumed to have a bivariate normal distribution with a common covariance matrix Σ, but possibly different mean vectors μ_1 and μ_2 for the two samples. The sample mean vectors and pooled covariance matrix are

$$\overline{x}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \overline{x}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$S_{\text{pooled}} = \begin{bmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{bmatrix}$$

(a) Construct Fisher's (sample) linear discriminant function. [See (11-19) and (11-25).]

(b) Assign the observation $x_0' = \begin{bmatrix} 0 & 1 \end{bmatrix}$ to either population π_1 or π_2. Assume equal costs and equal prior probabilities.

透過計算可得

$$S_{Pooled}^{-1} = \begin{bmatrix} 0.142 & 0.033 \\ 0.033 & 0.216 \end{bmatrix}$$

(a) 根據 Fisher 的定義可以建構

$$\begin{aligned}\hat{y} = (\overline{x}_1 - \overline{x}_2)^T S_{Pooled}^{-1} x &= \begin{bmatrix} -3 & -2 \end{bmatrix} \begin{bmatrix} 0.142 & 0.033 \\ 0.033 & 0.216 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= -0.491 x_1 - 0.529 x_2\end{aligned}$$

$$\begin{aligned}\hat{m} = \frac{1}{2}(\overline{x}_1 - \overline{x}_2)^T S_{Pooled}^{-1} (\overline{x}_1 + \overline{x}_2) &= \frac{1}{2} \begin{bmatrix} -3 & -2 \end{bmatrix} \begin{bmatrix} 0.142 & 0.033 \\ 0.033 & 0.216 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= -0.246\end{aligned}$$

可得 linear discriminant function 爲

$$\hat{w} = \hat{y} - \hat{m} = -0.491x_1 - 0.529x_2 + 0.246$$

(b) 在相同誤判成本及事前機率的情況下等同於使用 Fisher method 進行分類，將 x_0 代入

$$\hat{w}_0 = -0.529 + 0.246 = -0.283$$

由於 $\hat{w}_0 < 0$，因此將 x_0 分類至 π_2

4

The data in Table 11.5 contain observations on $X_2 =$ sepal width and $X_4 =$ petal width for samples from three species of irises. There are $n_1 = n_2 = n_3 = 50$ observations in each sample.

(a) Plot the data in the (x_2, x_4) variable space. Do the observations for the three groups appear normal?

(b) Assuming that the populations are bivariate normal, construct the quadratic discriminate scores $\hat{d}_i^Q(x)$ given by (11-47) with $p_1 = p_2 = p_3 = \frac{1}{3}$. Using Rule (11-48), classify the new observation $x_0' = [3.5 \quad 1.75]$ into population π_1, π_2, or π_3.

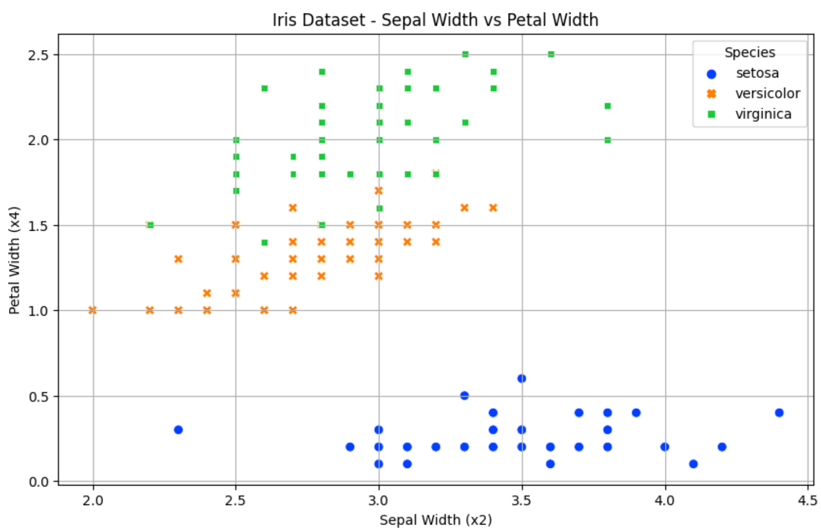
(c) Assume that the covariance matrices Σ_i are the same for all three bivariate normal populations. Construct the linear discriminate score $\hat{d}_i(x)$ given by (11-51), and use it to assign $x_0 = [3.5 \quad 1.75]$ to one of the populations π_i, $i = 1, 2, 3$ according to (11-52). Take $p_1 = p_2 = p_3 = \frac{1}{3}$. Compare the results in Parts b and c. Which approach do you prefer? Explain.

(d) Assuming equal covariance matrices and bivariate normal populations, and supposing that $p_1 = p_2 = p_3 = \frac{1}{3}$, allocate $x_0' = [3.5 \quad 1.75]$ to π_1, π_2, or π_3 using Rule (11-56). Compare the result with that in Part c. Delineate the classification regions $\hat{R}_1$, $\hat{R}_2$, and

$\hat{R}_3$ on your graph from Part a determined by the linear functions $\hat{d}_i(x_0)$ in (11-56).

(e) Using the linear discriminant scores from Part c, classify the sample observations. Calculate the APER and $\hat{E}(AER)$. (To calculate the latter, you should use Lachenbruch's holdout procedure. [See (11-57).])

(a) 透過 python 可以繪圖得到



可發現三種鳶尾花在圖中大致都是呈現橢圓形分佈，因此合理推斷其皆符合常態分佈，然而 setosa 的分佈方向與其他兩種略微不同，可能隱含其 Covariance matrix 與其他兩者不一樣。

(b) 根據上述資訊，我們可建構 $\hat{d}_i^Q(x)$ 如下:

$$\hat{d}_i^Q(x) = -\frac{1}{2}\ln|S_i| - \frac{1}{2}(x-\bar{x}_i)'\mathrm{S}_i^{-1}(x-\bar{x}_i) + \ln p_i \quad \text{for i = 1, 2 and 3}$$

其中

$$S_1 = \begin{bmatrix} 0.1436 & 0.0092 \\ 0.0092 & 0.0111 \end{bmatrix}, S_2 = \begin{bmatrix} 0.0984 & 0.0412 \\ 0.0412 & 0.0391 \end{bmatrix}, S_3 = \begin{bmatrix} 0.1040 & 0.0476 \\ 0.0476 & 0.0754 \end{bmatrix}$$

以及

$$\bar{x}_1' = \begin{bmatrix} 3.428 & 0.246 \end{bmatrix}, \bar{x}_2' = \begin{bmatrix} 2.77 & 1.326 \end{bmatrix}, \bar{x}_3' = \begin{bmatrix} 2.974 & 2.0026 \end{bmatrix}$$

$$p_1 = p_2 = p_3 = \frac{1}{3}$$

由此我們可計算出

$$\hat{d}_2^Q(x_0) = 1.1418 > \hat{d}_3^Q(x_0) = -0.1281 > \hat{d}_1^Q(x_0) = -102.6746$$

故我們可推論 $x_0 \in \pi_2$，即其爲 vesicolor。

(c) 根據上述定義與假設，我們可建構 $\hat{d}_i(x)$ 如下:

$$\hat{d}_i(x) = \bar{x}_i' \mathrm{S}_{pooled}^{-1} x - \frac{1}{2} \bar{x}_i' \mathrm{S}_{pooled}^{-1} \bar{x}_i + \ln p_i \quad \text{for i = 1, 2 and 3}$$

其中

$$\mathrm{S}_{pooled} = \begin{bmatrix} 0.1153 & 0.0327 \\ 0.00327 & 0.0418 \end{bmatrix}$$

由此我們可計算出

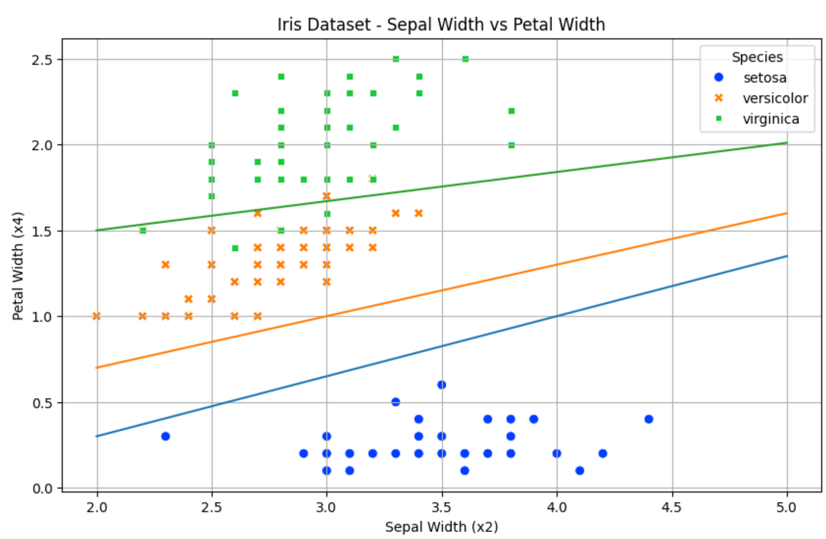
$$\hat{d}_2(x_0) = 57.7573 > \hat{d}_3(x_0) = 56.8190 > \hat{d}_1(x_0) = 27.0174$$

故我們可推論 $x_0 \in \pi_2$，即其爲 vesicolor。透過計算可發現 (b)(c) 的結論一樣，因此使用哪個方法並沒有差別。依我之見，根據本題的樣本，quadratic discriminate scores 應是較好的判別方法。因爲根據 (a) 劃出的散佈圖我們可以發現 setosa 的分佈方向與 versicolor 以及 virginica 略有不同，這表示 setosa 母體的 covariance matrix 應該不會與後兩者一致，此時使用假定三者有相同 covariance matrix 的 linear discriminate score 顯得有些武斷。

(d) 給定 $\hat{d}_{ij}(x) = \hat{d}_i(x) - \hat{d}_j(x)$ for i, j = 1, 2 and 3，我們可計算出如下表格:

$j \backslash i$	1	2	3
1	0	30.7399	29.8016
2	−30.7399	0	−0.9383
3	−29.8016	0.9383	0

透過觀察上方表格我們可以發現 $\hat{d}_{2j}(x) \geq 0$ for all j，由此我們可知 $x_0 \in \pi_2$，即其爲 vesicolor。
另外透過計算 $\hat{d}_i(x)$ for i, j = 1, 2 and 3，我們可以在 iris 的散佈圖上區分出 $\hat{R}_1, \hat{R}_2$ and $\hat{R}_3$。



(e) 透過觀察以上的 scatter plot 我們可以發現有 4 個 virginica 樣本被分到了 versicolor 中；有 1 個 versicolor 樣本被分到了 virginica 中，因此 APER $= \frac{4+1}{150} = 0.0333$。
透過執行 Lachenbruch's holdout procedure，我們可以得到 $(n_{1M}^{(H)}, n_{2M}^{(H)}, n_{3M}^{(H)}) = (0, 2, 4)$，由此我們可知 $\hat{E}(AER) = \frac{2+4}{150} = 0.04$。