Simple Discriminant Analysis

Rahul Shandilya

Introduction

- In discriminant analysis , the researcher is interested in the prediction and explanation of the relationships that affect the category in which an object is located, such as why a person is or is not a customer, or if a firm will succeed or fail.
- *The objective of discriminant analysis is to construct an effective rule for classifying previously unassigned individuals to two or more predetermined classes or groups based on several measurements.*
- Multiple discriminant analysis has widespread application in situations in which the primary objective is to identify the group to which an object belongs.
- Potential applications include predicting the success or failure of a new product, deciding whether a student should be admitted to graduate school, classifying students as to vocational interests, determining the category of credit risk for a person, or predicting whether a firm will be successful.

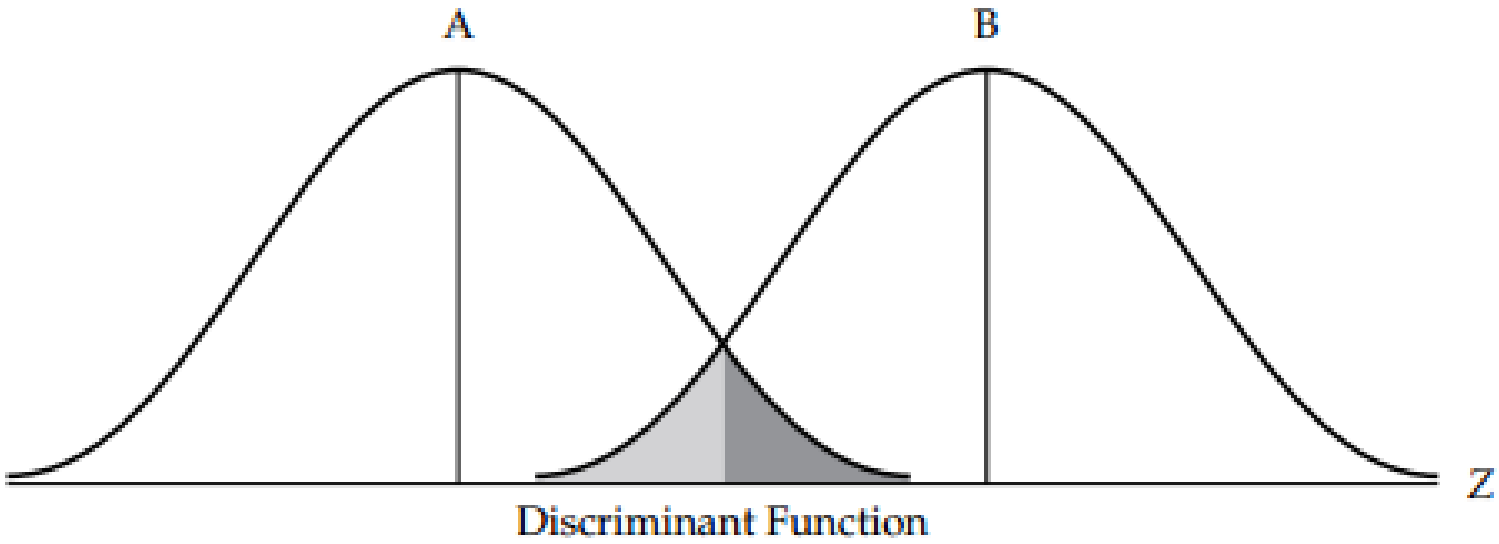
An example

Let us assume that certain evergreen trees can be broadly divided into two varieties on the basis of their leaf shape, and consider the use of width and length as two properties (variables) in constructing a formula for classification of newly obtained data known to be from one of the two varieties A and B but not from which one.

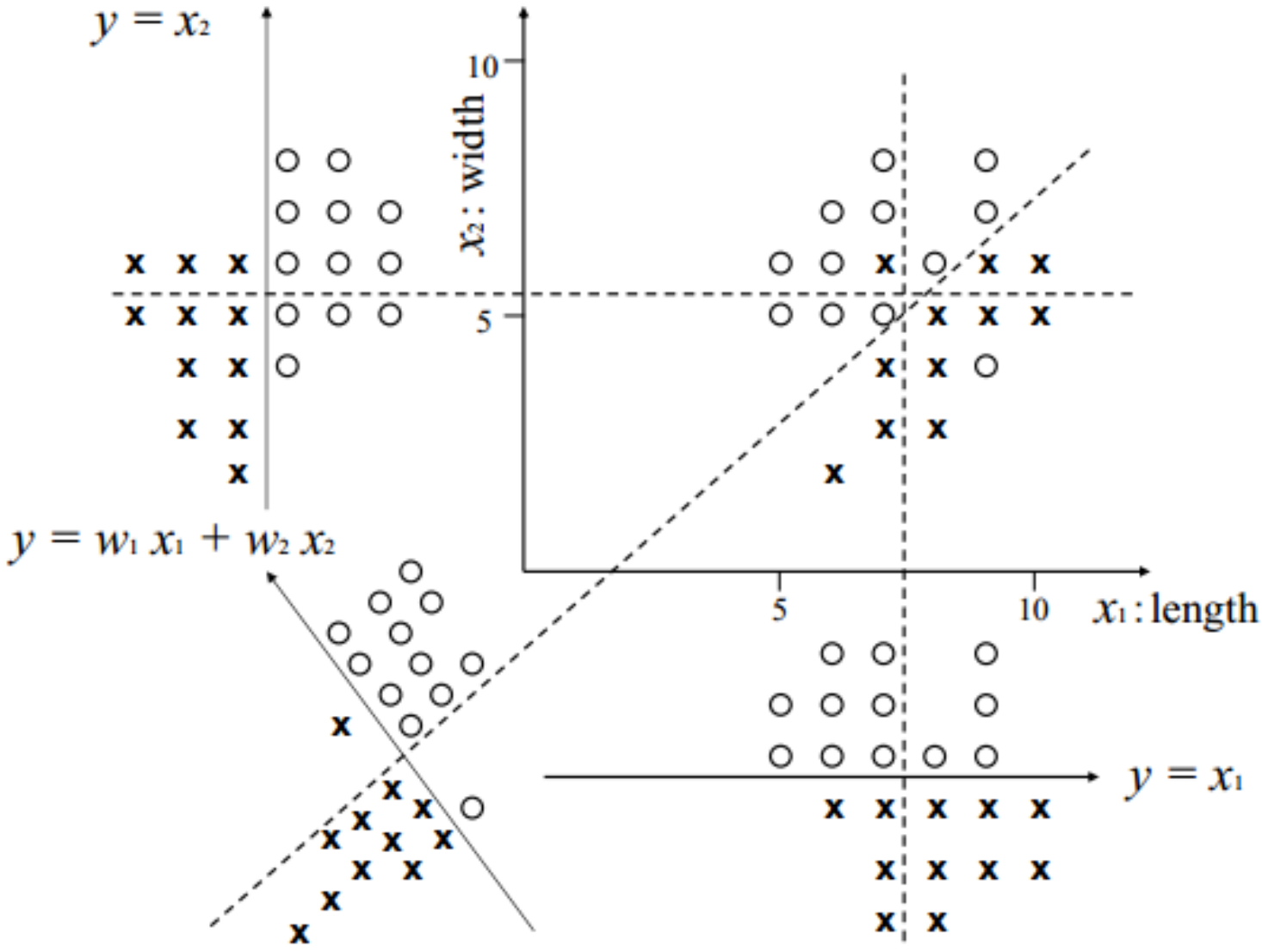
		1	2	3	4	5	6	7	8	9	10	11	12
A:	$L(x_1)$	5	7	6	8	5	9	6	9	7	6	7	9
	$W(x_2)$	5	7	7	6	6	8	6	7	5	5	8	4
B:	$L(x_1)$	6	8	7	9	7	10	8	10	9	8	7	
	$W(x_2)$	2	4	4	5	3	5	5	6	6	3	6	

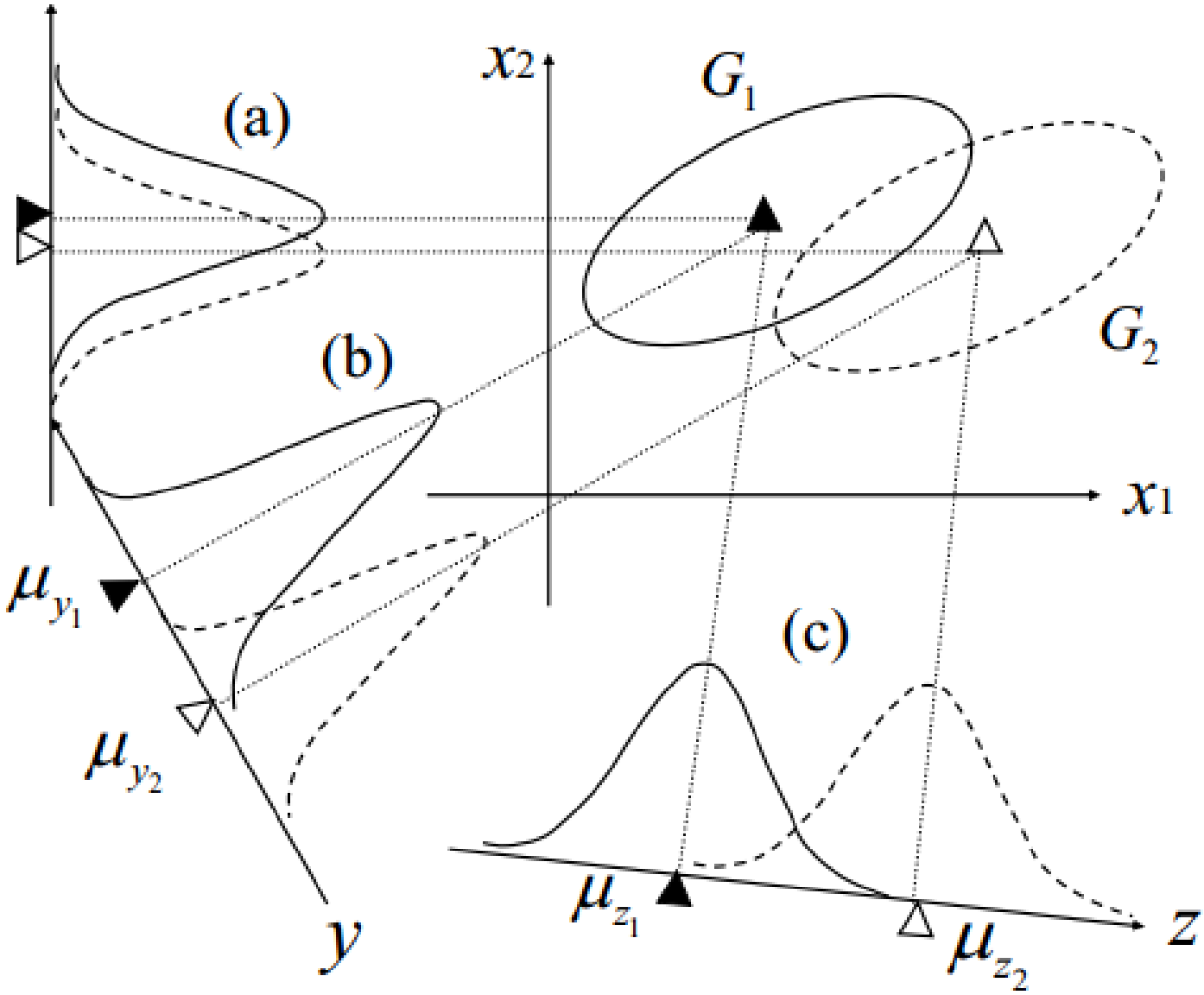
Figure: The 23 two-dimensional observed data from the varieties A and B.

Basis for classification



- To perform the classification based on the information presented by both variables, the question is then what kind of axis to use for the projection.
- This question can be reduced to variable weighting based on a criterion, by projecting the twodimensional data onto $y = w_1 x_1 + w_2 x_2$ in the figure and selecting the weighting that yields the best separation of the two classes.


$y = x_2$
10
x_2 : width
5
5
10
x_1 : length
$y = w_1 x_1 + w_2 x_2$
$y = x_1$


(a)
(b)
(c)
x_2
G_1
G_2
x_1
μ_{y_1}
μ_{y_2}
y
μ_{z_1}
μ_{z_2}
z

- A high degree of separation between the two class means generally facilitates classification.
- If it is attributable to a large dispersion, however, the region of overlap between the two classes will also tend to be large, with an adverse effect on the performance of the classification.
- to determine the optimum axis with respect to these advantageous and adverse effects one approach is to set coefficients w_1 and w_2 so as to obtain a high ratio of between-class variance to within-class variance in the projection onto axis $y = w_1 x_1 + w_2 x_2$ as follows:

$$\lambda = \frac{\text{between-class variance}}{\text{within-class variance}}$$

- the projection axis is selected to obtain the largest possible between-class variance in the numerator together with the smallest possible within-class variance in the denominator

- In the projection of the data onto axis y,

$$(\mu_{y1} - \mu_{y2})^2$$

can be regarded as a measure of the degree of separation of the two classes, referred to as the *between-class variance* or *between-class variation* on axis y.

- To determine the degree of data dispersion within each class with projection onto axis y, we consider the sum weighted for the number of data

$$\frac{(n_1 - 1)(\text{var. of } G_1 \text{ on } y) + (n_2 - 1)(\text{var. of } G_2 \text{ on } y)}{n_1 + n_2 - 2}$$

referred to as the *within-class variance* or *within-class variation* on axis y

- By projecting the i-th two-dimensional data $\mathbf{x}_i^{(1)} = (x_{i1}^{(1)}, x_{i2}^{(1)})^T$ of class G1 onto $y = w_1 x_1 + w_2 x_2$, we have

$$y_i^{(1)} = w_1 x_{i1}^{(1)} + w_2 x_{i2}^{(1)} \qquad i = 1, 2, \ldots, n_1$$

  Similary for G2

$$y_i^{(2)} = w_1 x_{i1}^{(2)} + w_2 x_{i2}^{(2)} \qquad i = 1, 2, \ldots, n_1$$

- The sample means of classes G1 and G2, as obtained from the onedimensional data on y, may be given by

$$\overline{y}^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} (w_1 x_{i1}^{(1)} + w_2 x_{i2}^{(1)}) = w_1 \overline{x}_1^{(1)} + w_2 \overline{x}_2^{(1)},$$

$$\overline{y}^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} (w_1 x_{i1}^{(2)} + w_2 x_{i2}^{(2)}) = w_1 \overline{x}_1^{(2)} + w_2 \overline{x}_2^{(2)}.$$

The *between-class variance* can be expressed as

$$(\bar{y}^{(1)} - \bar{y}^{(2)})^2 = \{w_1(\bar{x}_1^{(1)} - \bar{x}_1^{(2)}) + w_2(\bar{x}_2^{(1)} - \bar{x}_2^{(2)})\} \quad (1)$$

$$= \{\mathbf{w}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\}^2 \quad (2)$$

where

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \qquad \bar{\mathbf{x}}_1 = \begin{pmatrix} \bar{x}_1^{(1)} \\ \bar{x}_2^{(1)} \end{pmatrix}, \qquad \bar{\mathbf{x}}_2 = \begin{pmatrix} \bar{x}_1^{(2)} \\ \bar{x}_2^{(2)} \end{pmatrix}$$

when we project the data of class G_1 onto y, the sample variance on y is given by

$$\begin{aligned}
&\frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i^{(1)}-\bar{y}^{(1)})^2 \\
&=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\{w_1^2(x_{i1}^{(1)}-x_1^{(1)})^2+ \\
&\qquad 2w_1w_2(x_{i1}^{(1)}-x_1^{(1)})(x_{i2}^{(1)}-x_2^{(1)})+w_2^2(x_{i2}^{(1)}-x_2^{(1)})^2\} \\
&=w_1^2s_{11}^{(1)}+2w_1w_2s_{12}^{(1)}+w_2^2s_{22}^{(1)}=\mathbf{w}^TS_1\mathbf{w}
\end{aligned}$$

Similarly we have the sample variance on y for the data of class G_2

$$\frac{1}{n_1-1}\sum_{i=1}^{n_2}(y_i^{(2)}-\bar{y}^{(2)})^2=w_1^2s_{11}^{(2)}+2w_1w_2s_{12}^{(2)}+w_2^2s_{22}^{(2)}=\mathbf{w}^TS_2\mathbf{w}$$

where S_1 and S_2 are, respectively, the sample variance-covariance matrices of G_1 and G_2 given by

$$S_1 = \begin{pmatrix} s_{11}^{(1)} & s_{12}^{(1)} \\ s_{21}^{(1)} & s_{22}^{(1)} \end{pmatrix}, \qquad S_2 = \begin{pmatrix} s_{11}^{(2)} & s_{12}^{(2)} \\ s_{21}^{(2)} & s_{22}^{(2)} \end{pmatrix} \tag{6.11}$$

Hence the within-class variance defined by the formula (6.3) can be written as

$$\frac{1}{n_1 + n_2 - 2} \left\{ (n_1 - 1)\boldsymbol{w}^T S_1 \boldsymbol{w} + (n_2 - 1)\boldsymbol{w}^T S_2 \boldsymbol{w} \right\} = \boldsymbol{w}^T S \boldsymbol{w}, \tag{6.12}$$

where

$$S = \frac{1}{n_1 + n_2 - 2} \{ (n_1 - 1)S_1 + (n_2 - 1)S_2 \}. \tag{6.13}$$

The matrix S is called the *pooled sample variance-covariance matrix.*

- The ratio of betweenclass variance to within-class variance in the projection onto axis $y = w_1 x_1 + w_2 x_2$ can be expressed as

$$\lambda = \frac{\{\mathbf{w}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\}^2}{\mathbf{w}^T S \mathbf{w}}$$

- the coefficient vector $\mathbf{w}$ which maximizes the ratio λ is

$$\hat{\mathbf{w}} = S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

- The optimum projection axis is

$$y = \hat{w}_1 x_1 + \hat{w}_2 x_2 = \hat{\mathbf{w}}^T \mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S^{-1} \mathbf{x}$$

for maximum separation of the two classes. This linear function is called *Fisher's linear discriminant function*.

Comparing with the mid-point we have classification rule

$$h(\boldsymbol{x}) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T S^{-1} \boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T S^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \begin{cases} \geq 0 & \Rightarrow \quad G_1 \\ & \qquad (6.19) \\ < 0 & \Rightarrow \quad G_2. \end{cases}$$