Classification and Multi-group LDA

ST 558: Multivariate Analytics

Module 5

Lecture 2

Discriminant Analysis as a Classification Tool

We have seen Linear Discriminant Analysis (LDA) as a descriptive tool, used to identify a linear combination of variables that best separates groups.

Now we consider using LDA as a classification tool: can we assign a *new* observation to a group using the linear discriminant function(s)?

Classification/Prediction with Linear Discriminant Analysis

Suppose we have a new observation:

$$\mathbf{X}_0 = [X_{01}, X_{02}, \dots, X_{0p}]$$

We would like to decide whether this observation is more likely to belong to Group 1 or Group 2.

A classification rule based on the linear combination $y = \mathbf{a}^T\mathbf{x}$ naturally follows:

- Classify $\mathbf{X}_0$ as belonging to Group 1 if

$$Y_0 = \mathbf{a}^T\mathbf{X}_0 \geq \frac{\mathbf{a}^T\overline{\mathbf{X}}_1 + \mathbf{a}^T\overline{\mathbf{X}}_2}{2}$$

- Classify $\mathbf{X}_0$ as belonging to Group 2 if

$$Y_0 = \mathbf{a}^T\mathbf{X}_0 < \frac{\mathbf{a}^T\overline{\mathbf{X}}_1 + \mathbf{a}^T\overline{\mathbf{X}}_2}{2}$$

Classification/Prediction with Linear Discriminant Analysis

Suppose we have a new observation:

$$\mathbf{X}_0 = [X_{01}, X_{02}, \dots, X_{0p}]$$

We would like to decide whether this observation is more likely to belong to Group 1 or Group 2.

A classification rule based on the lin Average of the linear combinations of the two means. $\mathbf{a}^T\mathbf{x}$ naturally follows:

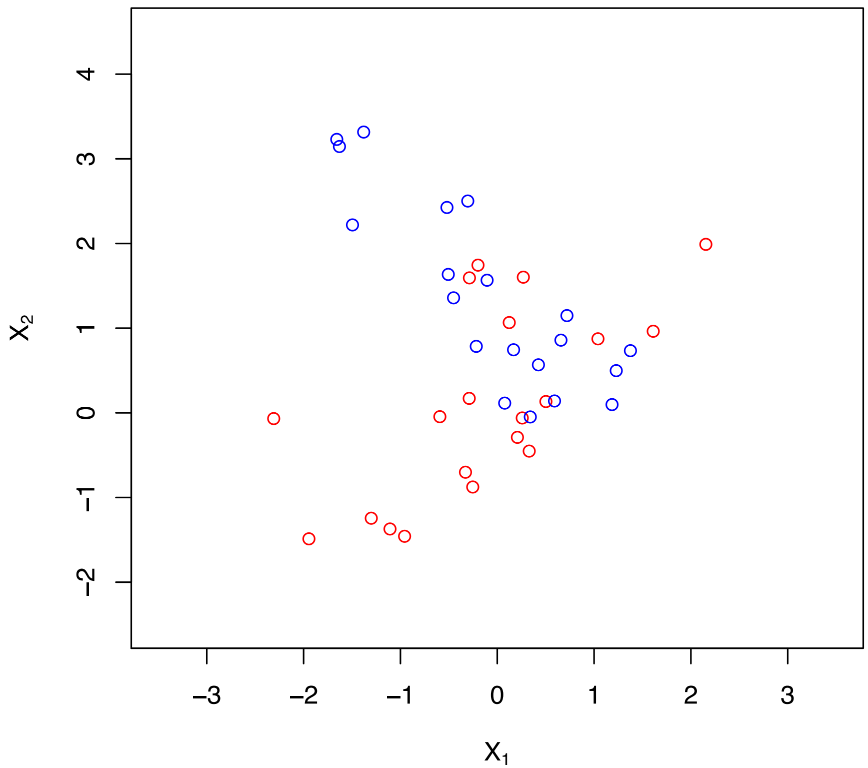
- Classify $\mathbf{X}_0$ as belonging to Group 1 if

$$Y_0 = \mathbf{a}^T\mathbf{X}_0 \geq \frac{\mathbf{a}^T\overline{\mathbf{X}}_1 + \mathbf{a}^T\overline{\mathbf{X}}_2}{2}$$

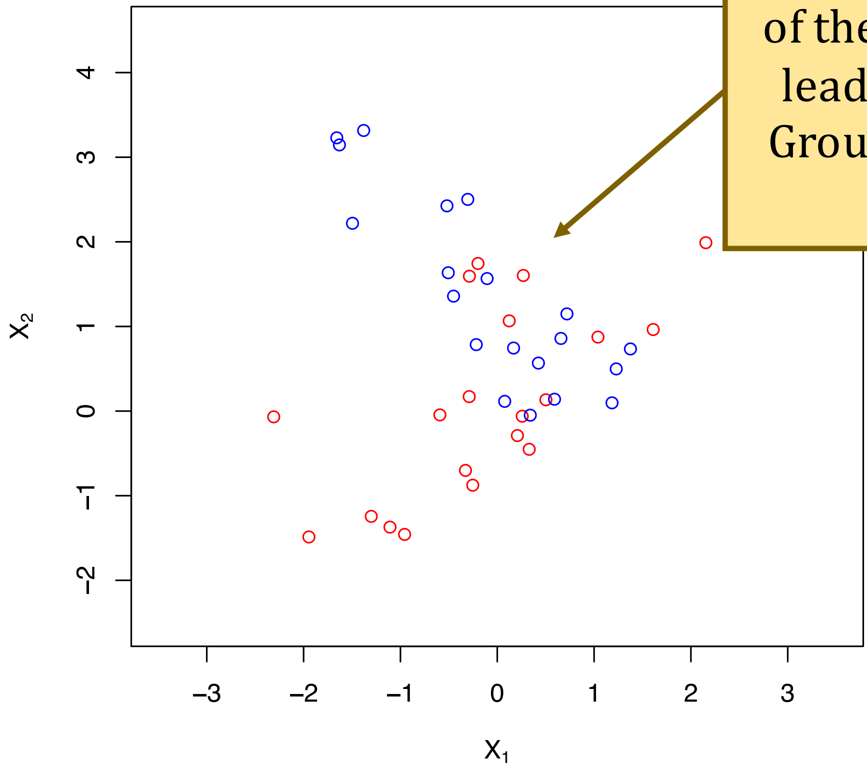
- Classify $\mathbf{X}_0$ as belonging to Group 2 if

$$Y_0 = \mathbf{a}^T\mathbf{X}_0 < \frac{\mathbf{a}^T\overline{\mathbf{X}}_1 + \mathbf{a}^T\overline{\mathbf{X}}_2}{2}$$

Classification/Prediction with Linear Discriminant Analysis

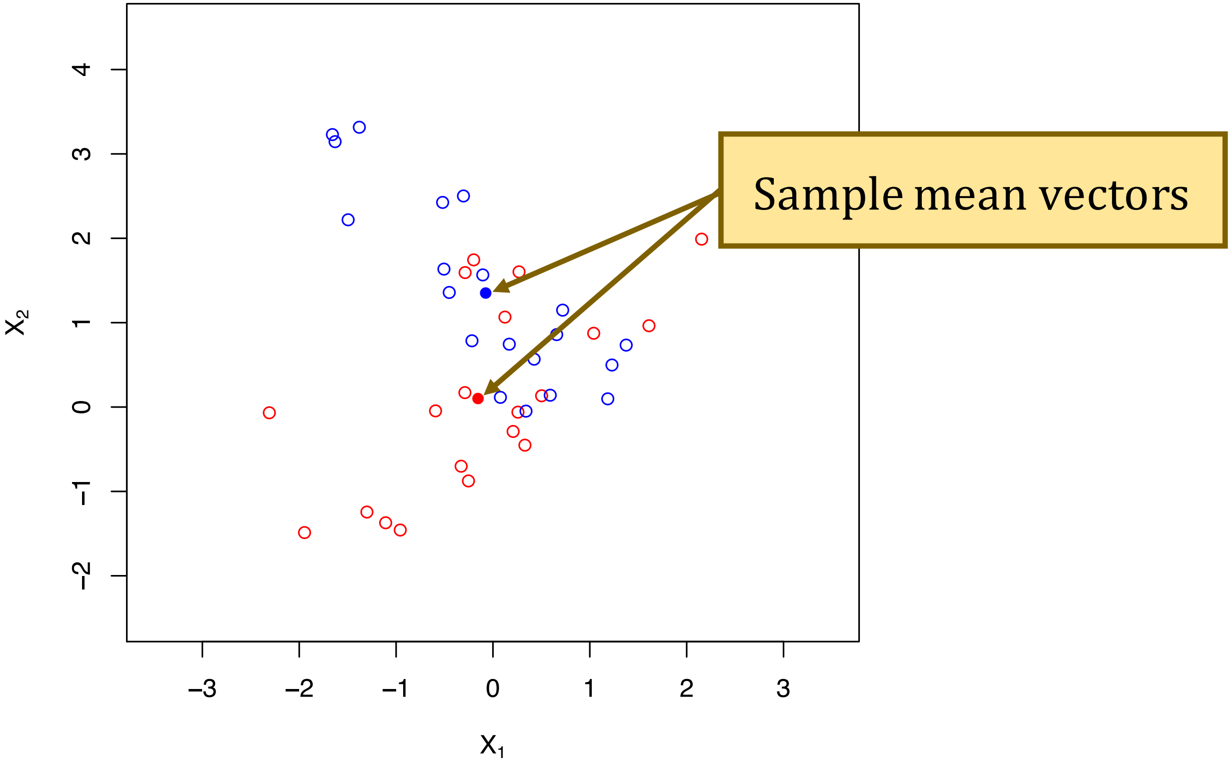


Classification/Prediction with Linear Discriminant Analysis

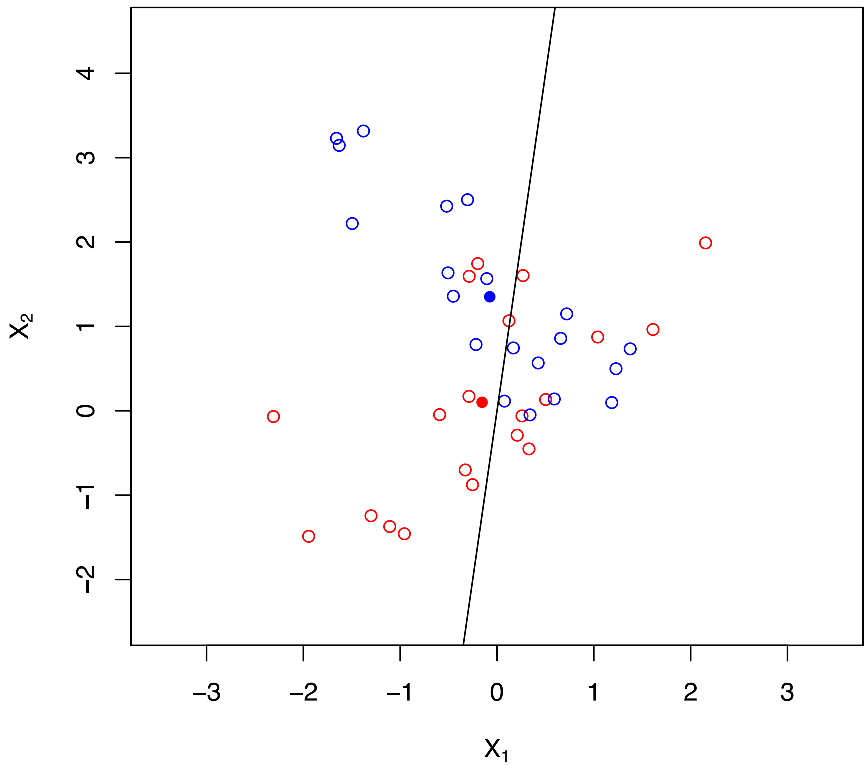


Goal: Decide which regions of the sample space would lead to a classification of Group 1 (Red) vs. Group 2 (Blue)

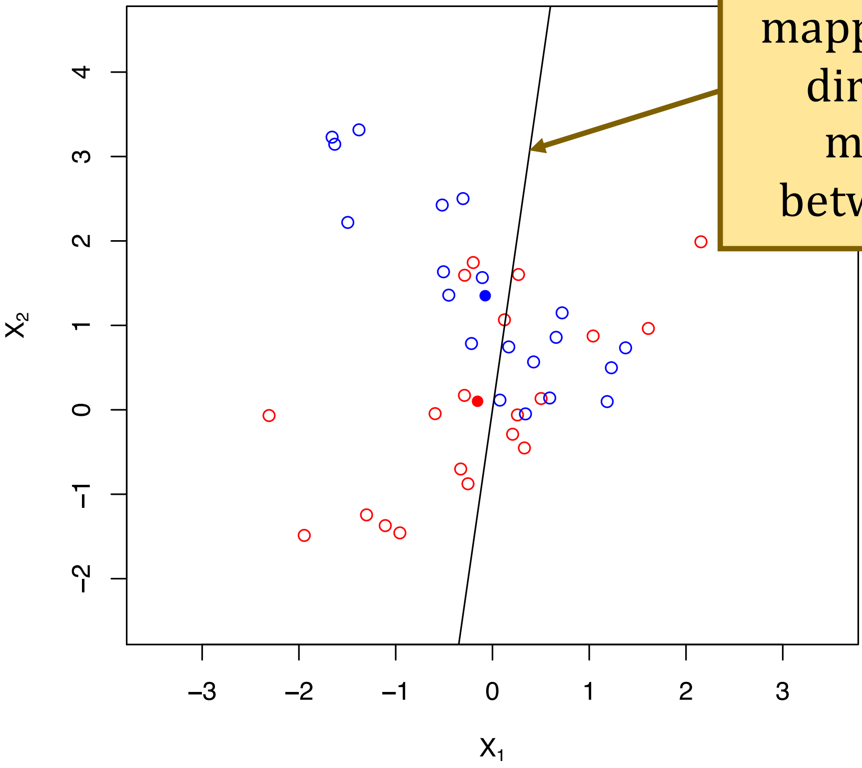
Classification/Prediction with Linear Discriminant Analysis



Classification/Prediction with Linear Discriminant Analysis

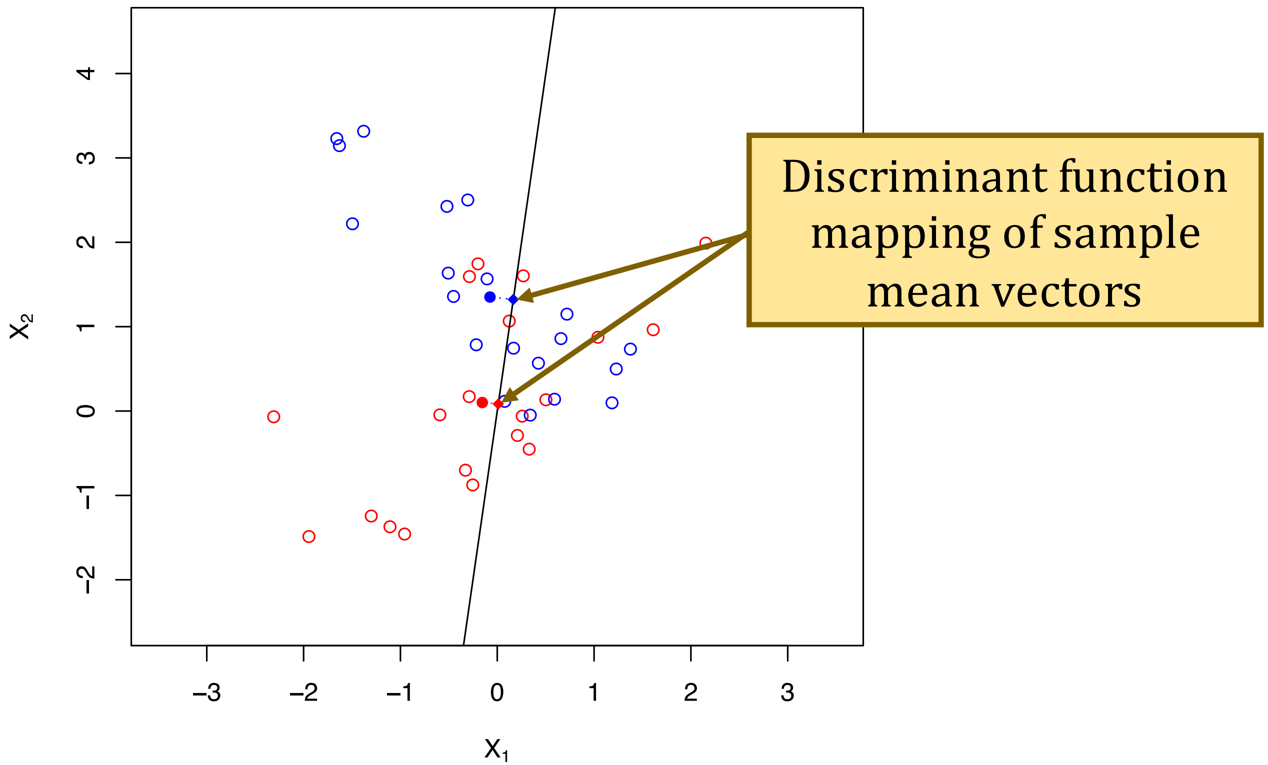


Classification/Prediction with Linear Discriminant Analysis

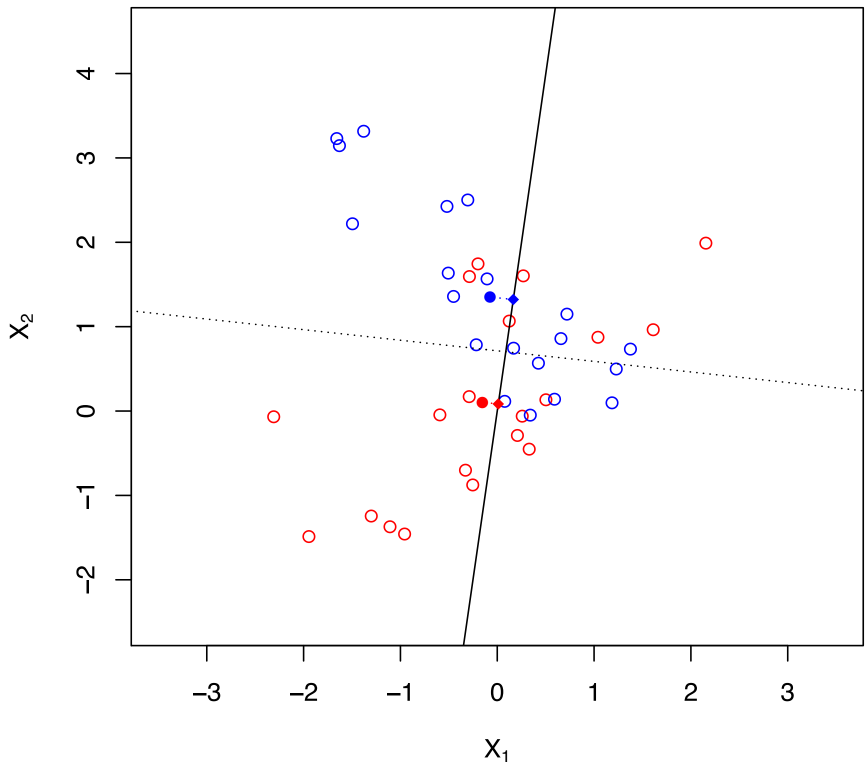


Discriminant direction: mapping the points to this direction will provide maximal separation between the two groups

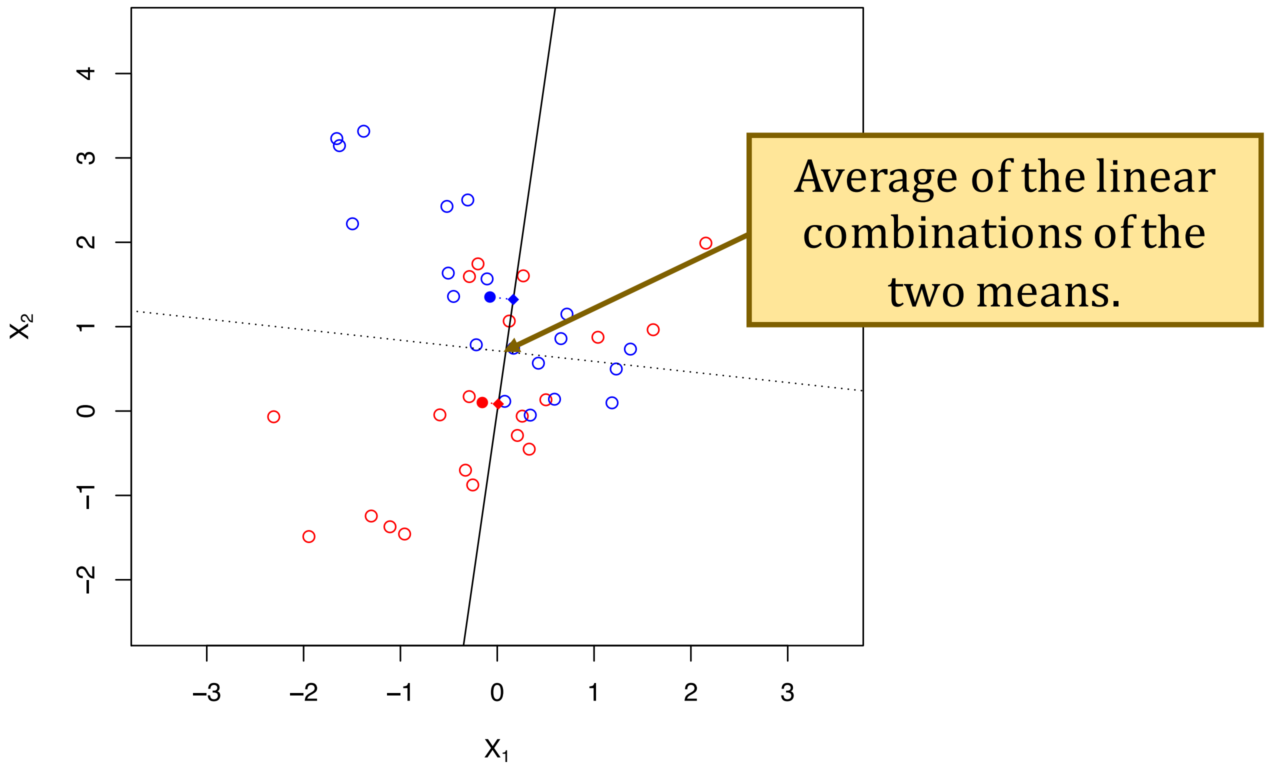
Classification/Prediction with Linear Discriminant Analysis



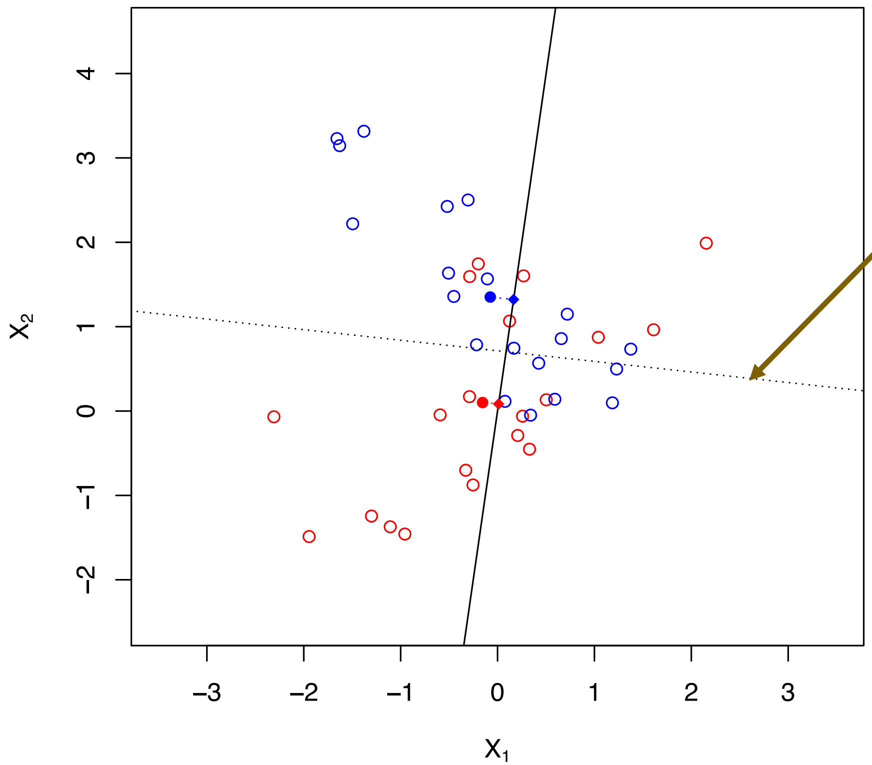
Classification/Prediction with Linear Discriminant Analysis



Classification/Prediction with Linear Discriminant Analysis

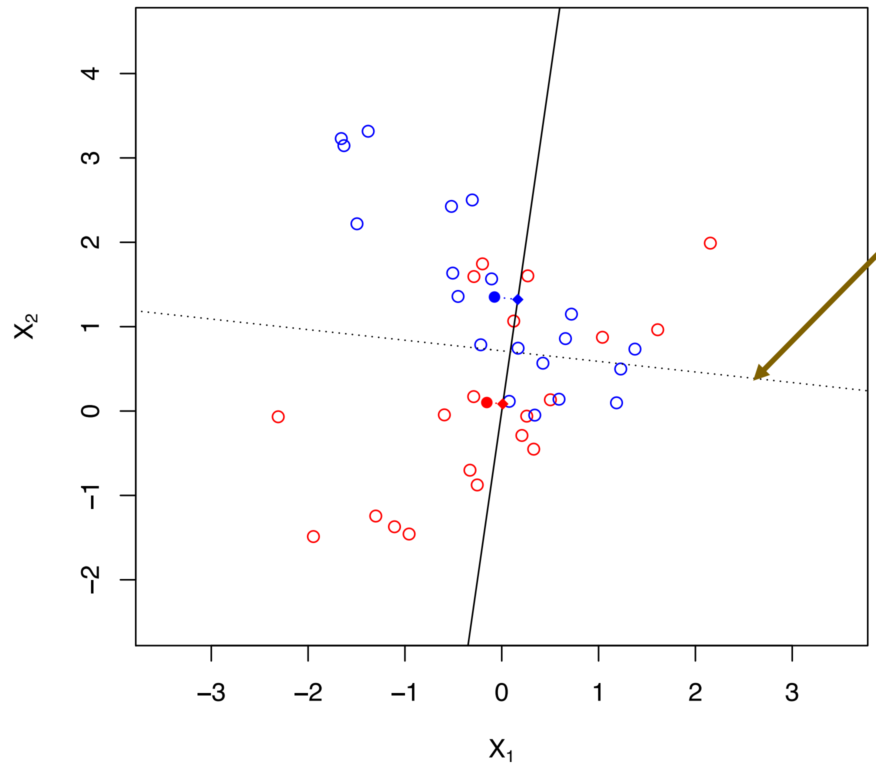


Classification/Prediction with Linear Discriminant Analysis



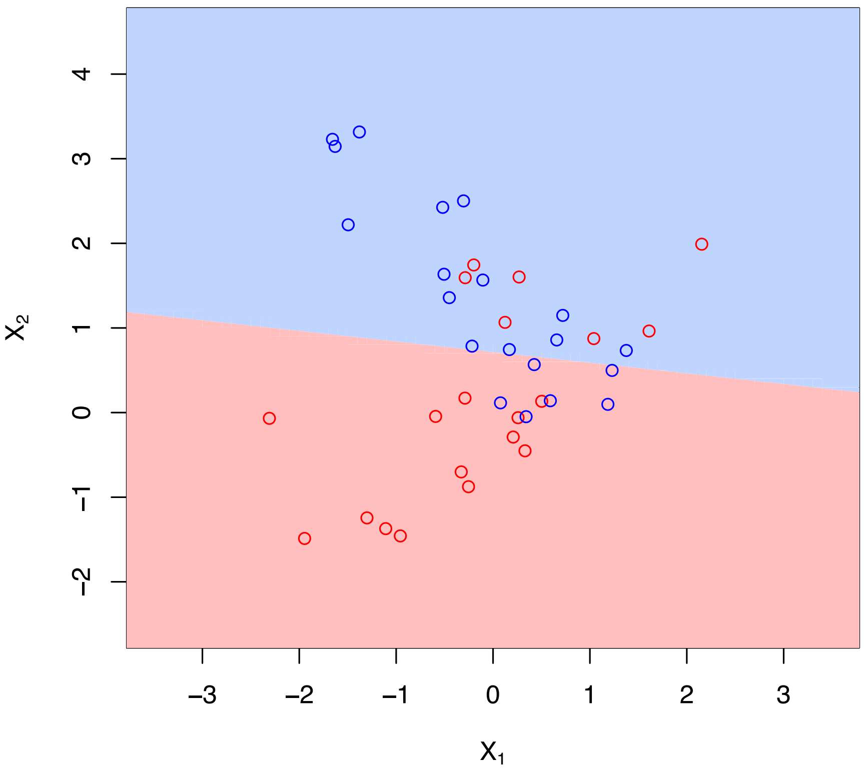
Any point ***below*** this line will map to a Y value that is closer to the mapped mean of **Group 1 (Red)**

Classification/Prediction with Linear Discriminant Analysis

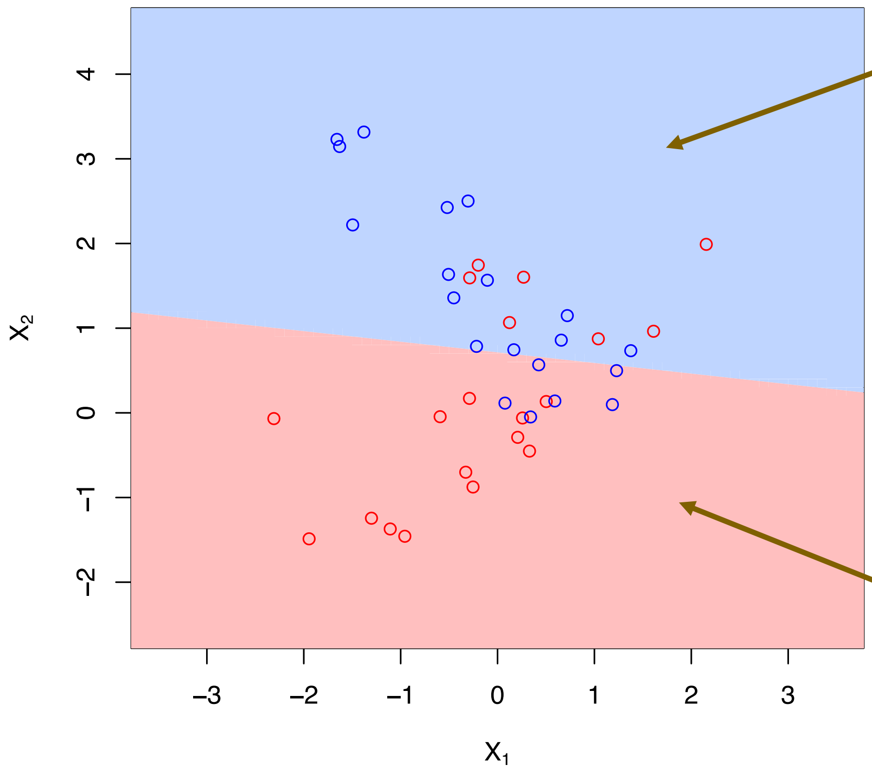


Any point ***above*** this line will map to a Y value that is closer to the mapped mean of **Group 2 (Blue)**

Classification/Prediction with Linear Discriminant Analysis



Classification/Prediction with Linear Discriminant Analysis



Any new observation in this region of the sample space will be classified as **Group 2 (Blue)**

Any new observation in this region of the sample space will be classified as **Group 1 (Red)**

Multi-group LDA: Discriminant Directions

We can also use Linear Discriminant Analysis with more than two groups.

- Discrimination Goal: Find linear combination(s)

$$y_1 = \mathbf{a}_1^T \mathbf{x}$$
$$\vdots$$
$$y_s = \mathbf{a}_s^T \mathbf{x}$$

  such that the groups are well-separated on these new variables y_1, …, y_s.

Multi-group LDA: Discriminant Directions

Setting:

- Variables $X_1, X_2, \ldots, X_p$ measured on subjects/observation units from $\boxed{k > 1}$ different populations/groups.

General Goal:

- Based on these variables, partition the *sample space* into regions $R_1, R_2, \ldots, R_k$ such that R_ℓ is the region of values $\mathbf{x} = \left[x_1, x_2, \ldots, x_p\right]$ for which an observation is more likely to belong to group ℓ

Multi-group LDA: Discriminant Directions

Again, we define separation by comparing the spread of the linear combination group means to the spread within groups.

Recall the two-group separation:

$$\text{Separation} = \frac{|\bar{Y}_1 - \bar{Y}_2|}{s_Y} = \frac{|\mathbf{a}^T \bar{\mathbf{X}}_1 - \mathbf{a}^T \bar{\mathbf{X}}_2|}{\mathbf{a}^T \mathbf{S}_P \mathbf{a}}$$

$$\text{Separation}^2 = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{s_Y^2} = \frac{\mathbf{a}^T (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\,(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \mathbf{a}}{\mathbf{a}^T \mathbf{S}_P \mathbf{a}}$$

Multi-group LDA: Discriminant Directions

Two-group squared separation:

$$\text{Separation}^2 = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{s_Y} = \frac{\mathbf{a}^T(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\,(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T\mathbf{a}}{\mathbf{a}^T\mathbf{S}_P\mathbf{a}}$$

Extension to multiple groups:

$$\text{Separation}^2 = \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}$$

Here **B** is the ‘Between Sum of Squares’ matrix that we first encountered in MANOVA.

Multi-group LDA: Discriminant Directions

Two-group squared separation:

$$\text{Separation}^2 = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{s_Y} = \frac{\mathbf{a}^T \boxed{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\,(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T} \mathbf{a}}{\mathbf{a}^T \mathbf{S}_P \mathbf{a}}$$

Measure of the spread between different group means.

Extension to multiple groups:

$$\text{Separation}^2 = \frac{\mathbf{a}^T \boxed{\mathbf{B}} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}}$$

Multi-group LDA: Discriminant Directions

Two-group squared separation:

$$\text{Separation}^2 = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{s_Y} = \frac{\mathbf{a}^T(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\,(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T\mathbf{a}}{\mathbf{a}^T\boxed{\mathbf{S}_P}\mathbf{a}}$$

Extension to multiple groups:

$$\text{Separation}^2 = \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\boxed{\mathbf{W}}\mathbf{a}}$$

Measure of the spread within individual groups

Multi-group LDA: Discriminant Directions

Consider the matrix $\mathbf{W}^{-1}\mathbf{B}$.

This is not a symmetric matrix, but it does still have:

- Eigenvalues: $\lambda_1 > \lambda_2 > \ \ldots > \lambda_s$
- Corresponding eigenvectors: $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{v}_j = \lambda_j \mathbf{v}_j$$

Multi-group LDA: Discriminant Directions

Consider the matrix $\mathbf{W}^{-1}\mathbf{B}$.

This is not a symmetric matrix, but it c

- Eigenvalues: $\lambda_1 > \lambda_2 > \ \dots > \lambda_s$
- Corresponding eigenvectors: $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_s$

s = Number of non-zero eigenvalues
= Rank($\mathbf{B}$)
= min(k – 1 , p)

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{v}_j = \lambda_j \mathbf{v}_j$$

Multi-group LDA: Discriminant Directions

Consider the matrix $\mathbf{W}^{-1}\mathbf{B}$.

This is not a symmetric matrix, but it does still have:

- Eigenvalues: $\lambda_1 > \lambda_2 > \ \ldots > \lambda_s$
- Corresponding eigenvectors: $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{v}_j = \lambda_j \mathbf{v}_j$$

The linear combination (direction) that produces maximal separation is given by the eigenvector $\mathbf{v}_1$ corresponding to the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$.

Multi-group LDA: Discriminant Directions

Consider the matrix $\mathbf{W}^{-1}\mathbf{B}$.

This is not a symmetric matrix, but it does still have:

- Eigenvalues: $\lambda_1 > \lambda_2 > \ \dots > \lambda_s$
- Corresponding eigenvectors: $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_s$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{v}_j = \lambda_j \mathbf{v}_j$$

That is, $Y_1 = \mathbf{v}_1\mathbf{X}$ is the linear combination that produces maximal separation between the groups.

Multi-group LDA: Discriminant Directions

Consider the matrix $\mathbf{W}^{-1}\mathbf{B}$.

This is not a symmetric matrix, but it does still have:

- Eigenvalues: $\lambda_1 > \lambda_2 > \ \ldots > \lambda_s$
- Corresponding eigenvectors: $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{v}_j = \lambda_j \mathbf{v}_j$$

We can construct a linear combination corresponding to each eigenvector: $Y_j = \mathbf{v}_j\mathbf{X}$

Multi-group LDA: Discriminant Directions

Consider the matrix $\mathbf{W}^{-1}\mathbf{B}$.

This is not a symmetric matrix, but it does still have:

- Eigenvalues: $\lambda_1 > \lambda_2 > \ \dots > \lambda_s$
- Corresponding eigenvectors: $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_s$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{v}_j = \lambda_j \mathbf{v}_j$$

These different linear combinations are in sorted order based on the separation they achieve:

$$\text{Sep}(Y_1) > \text{Sep}(Y_2) > \ \dots > \text{Sep}(Y_s)$$

Multi-group LDA: Discriminant Directions

Often we would like to use just a few discriminant directions (*discriminant functions*) to describe the data and represent the separation between the groups.

In many cases, just a few of these discriminant functions Y_1, Y_2, …, Y_s, will be sufficient to capture the structure of the data.

The relative importance of a given discriminant function Y_j can be calculated as

$$\frac{\lambda_j}{\sum_{\ell=1}^{s} \lambda_\ell}$$

λ_j = jth largest eigenvalue
= eigenvalue corresponding to eigenvector $\mathbf{v}_j$

which tells how much of the total separation between groups is given by the jth discriminant function.

Discrimination and Classification: Example

<u>Example:</u> Recall the Iris Dataset, which measures the following four variables for 50 samples from each of three types of Iris:

- X_1 = Sepal Length
- X_2 = Sepal Width
- X_3 = Petal Length
- X_4 = Petal Width

Now we will consider all 3 Types of Iris, and we will calculate the discriminant functions in order of the separation they provide between the three groups.

Discrimination and Classification: Example

<u>Example:</u> We calculate the matrices $\mathbf{W}$ and $\mathbf{B}$, and then compute the eigenvalues and eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$. First, note that s = min(k-1, p) = min(3-1, 4) = 2. Therefore, we report two eigenvalue-eigenvector pairs:

Eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$	Corresponding Eigenvector of $\mathbf{W}^{-1}\mathbf{B}$
32.192	$[\ 0.21 \quad 0.39 \quad -0.55 \quad -0.71]^T$
0.285	$[-0.01 \quad -0.59 \quad 0.25 \quad -0.77]^T$

Discrimination and Classification: Example

Example: We calculate the matrices **W** and **B**, and then compute the eigenvalues and eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$. First, note that s = min(k-1, p) = min(3-1, 4) = 2. Therefore, we report two eigenvalue-eigenvector pairs:

Eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$	Corresponding Eigenvector of $\mathbf{W}^{-1}\mathbf{B}$
32.192	$[0.21 \quad 0.39 \quad -0.55 \quad -0.71]^T$
0.285	$[-0.01 \quad -0.59 \quad 0.25 \quad -0.77]^T$

The first discriminant function is therefore

$$Y_1 = 0.21X_1 + 0.39X_2 - 0.55X_3 - 0.71X_4$$

This discriminant explains $\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{32.192}{32.192+0.285} = 99.12\%$ of the separation between the three groups.

Discrimination and Classification: Example

<u>Example:</u> We calculate the matrices $\mathbf{W}$ and $\mathbf{B}$, and then compute the eigenvalues and eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$. First, note that s = min(k-1, p) = min(3-1, 4) = 2. Therefore, we report two eigenvalue-eigenvector pairs:

Eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$	Corresponding Eigenvector of $\mathbf{W}^{-1}\mathbf{B}$
32.192	$[\,0.21 \quad 0.39 \quad -0.55 \quad -0.71]^T$
0.285	$[-0.01 \quad -0.59 \quad 0.25 \quad -0.77]^T$

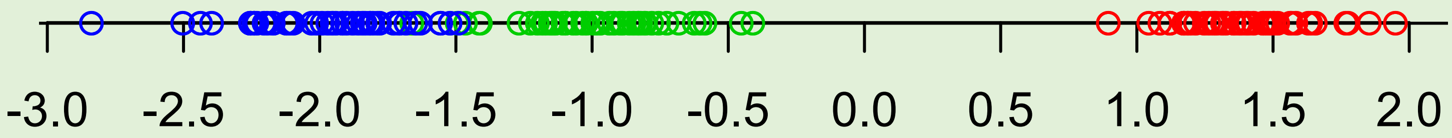
The second discriminant function is

$$Y_2 = -0.01X_1 - 0.59X_2 + 0.25X_3 - 0.77X_4$$

This discriminant explains $\frac{\lambda_2}{\lambda_1+\lambda_2} = \frac{0.285}{32.192+0.285} = 0.88\%$ of the separation between the three groups.

Discrimination and Classification: Example

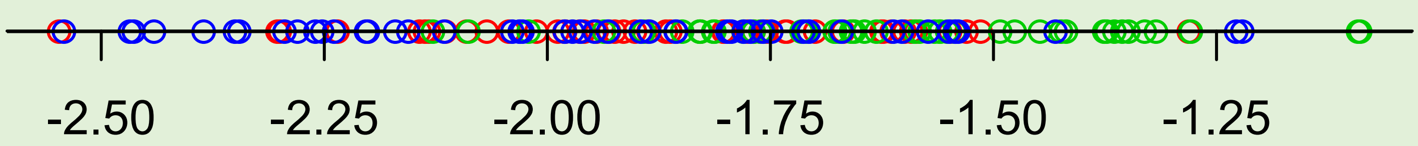
Example: Values of the first discriminant function Y_1:



Note that the k = 3 groups are quite well separated in this linear combination.

Discrimination and Classification: Example

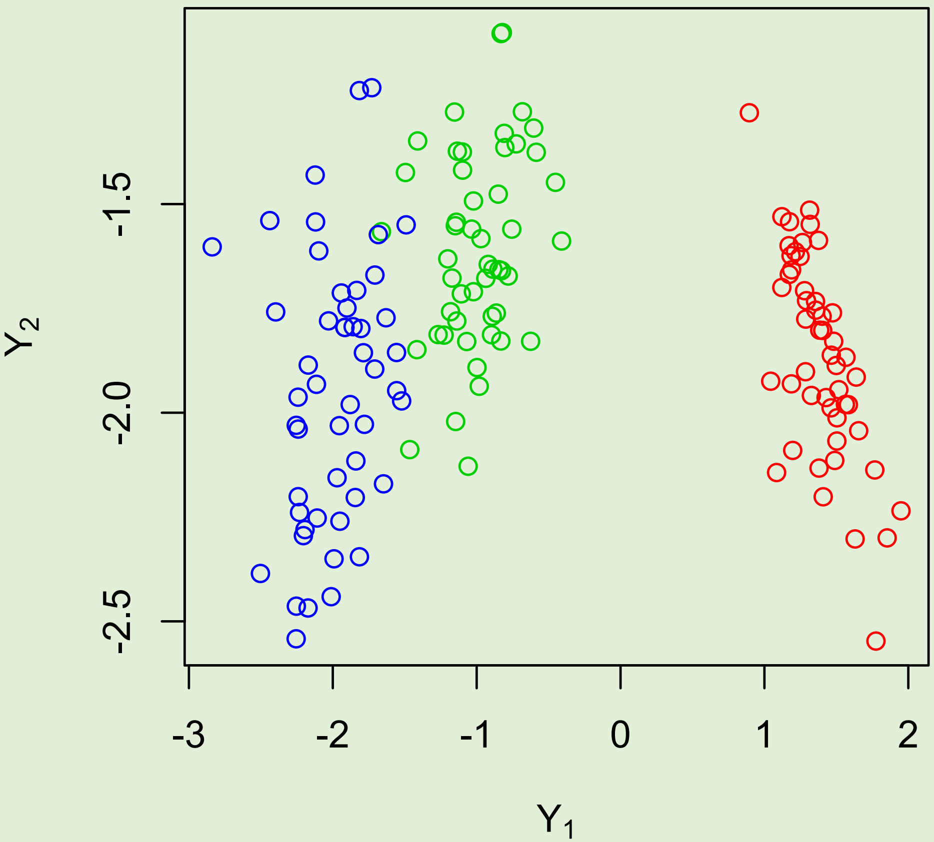
Example: Values of the second discriminant function Y_2:



Note that the $k = 3$ groups are not at all well separated in this second linear combination.

Discrimination and Classification: Example

Example: Iris data plotted in terms of the two linear discriminant functions. We see that these two discriminant functions do quite a good job at separating the three groups, with the majority of the work done by the first discriminant function Y_1.



Multi-group LDA: Class Predictions

Multi-group LDA can also be used to make class predictions (classifications) for new observations.

The basic idea of the classification rule is to decide which sample mean a new observation is closest to, where *closest* is defined in terms of distance relative to the pooled covariance estimate.

Multi-group LDA: Class Predictions

For a new observation

$$\mathbf{X}_0 = [X_{01}, X_{02}, \dots, X_{0p}]$$

we would like to decide whether this observation should be classified as Group 1, Group 2, …, or Group k.

To do this, we calculate a *distance* between the new observation $\mathbf{X}_0$ and the mean vector $\bar{\mathbf{X}}_\ell$ for each group $\ell = 1, 2, \dots, k$:

$$D_\ell(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_\ell)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_\ell)$$

Multi-group LDA: Class Predictions

For a new observation

$$\mathbf{X}_0 = [X_{01}, X_{02}, \dots, X_{0p}]$$

we would like to decide whether this observation should be classified as Group 1, Group 2, …, or Group k.

To do this, we calculate a *distance* between the new observation $\mathbf{X}_0$ and the mean of each group $\ell = 1, 2, \dots, k$:

Recall the pooled covariance matrix:

$$\mathbf{S}_P = \frac{\sum_{\ell=1}^{k}(n_\ell - 1)\mathbf{S}_\ell}{N-k} = \frac{\mathbf{W}}{N-k}$$

$$D_\ell(\mathbf{X}_0) = (\mathbf{X}_0 - \overline{\mathbf{X}}_\ell)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \overline{\mathbf{X}}_\ell)$$

Multi-group LDA: Class Predictions

For a new observation

$$\mathbf{X}_0 = [X_{01}, X_{02}, \dots, X_{0p}]$$

we would like to decide whether this observation should be classified as Group 1, Group 2, …, or Group k.

We will see in an upcoming lecture that this distance is called the *Mahalanobis distance* or *statistical distance* between $\mathbf{X}_0$ and $\bar{\mathbf{X}}_\ell$

To do this, we calculate a *distanc* new observation $\mathbf{X}_0$ and the mea each group $\ell = 1, 2, \dots, k$:

$$D_\ell(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_\ell)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_\ell)$$

Multi-group LDA: Class Predictions

We get k distances for the new point:

$$
\begin{gathered}
D_1(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_1)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_1) \\
D_2(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_2)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_2) \\
\vdots \\
D_k(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_k)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_k)
\end{gathered}
$$

Then we classify $\mathbf{X}_0$ to the group ℓ for which $D_\ell(\mathbf{X}_0)$ is *smallest*.

Discrimination and Classification: Example

<u>Example:</u> Recall the Iris Dataset, which measures the following four variables for 50 samples from each of three types of Iris:

- X_1 = Sepal Length
- X_2 = Sepal Width
- X_3 = Petal Length
- X_4 = Petal Width

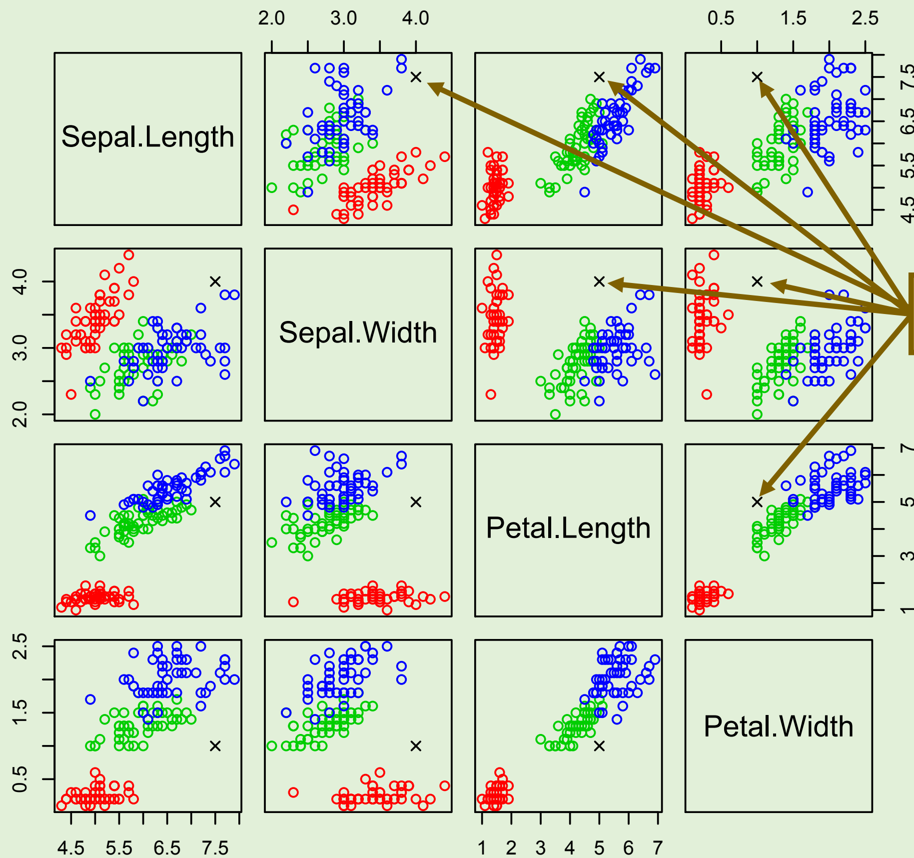
Suppose we have observed a new flower with measurements:

- $X_{0,1}$ = Sepal Length = 7.5
- $X_{0,2}$ = Sepal Width = 4.0
- $X_{0,3}$ = Petal Length = 5.0
- $X_{0,4}$ = Petal Width = 1.0

Which group should we assign this new flower to, based on these measurements?

Discrimination and Classification: Example

<u>Example:</u>



$\mathbf{X}_0 = [7.5 \;\; 4.0 \;\; 5.0 \;\; 1.0]^T$

Discrimination and Classification: Example

Example:

$$\bar{\mathbf{X}}_1 = [5.006 \;\; 3.428 \;\; 1.462 \;\; 0.246]^T$$
$$\bar{\mathbf{X}}_2 = [5.936 \;\; 2.770 \;\; 4.260 \;\; 1.326]^T$$
$$\bar{\mathbf{X}}_3 = [6.588 \;\; 2.974 \;\; 5.552 \;\; 2.026]^T$$

$$\mathbf{S}_P = \begin{bmatrix} 0.265 & 0.093 & 0.168 & 0.038 \\ 0.093 & 0.115 & 0.055 & 0.033 \\ 0.168 & 0.055 & 0.185 & 0.043 \\ 0.038 & 0.033 & 0.043 & 0.042 \end{bmatrix}$$

$$D_1(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_1)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_1) = 72.564$$
$$D_2(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_2)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_2) = 31.378$$
$$D_3(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_3)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_3) = 65.452$$

Discrimination and Classification: Example

Example:

$$\bar{\mathbf{X}}_1 = [5.006 \;\; 3.428 \;\; 1.462 \;\; 0.246]^T$$
$$\bar{\mathbf{X}}_2 = [5.936 \;\; 2.770 \;\; 4.260 \;\; 1.326]^T$$
$$\bar{\mathbf{X}}_3 = [6.588 \;\; 2.974 \;\; 5.552 \;\; 2.026]^T$$

$$\mathbf{S}_P = \begin{bmatrix} 0.265 & 0.093 & 0.168 & 0.038 \\ 0.093 & 0.115 & 0.055 & 0.033 \\ 0.168 & 0.055 & 0.185 & 0.043 \\ 0.038 & 0.033 & 0.043 & 0.042 \end{bmatrix}$$

$$D_1(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_1)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_1) = 72.564$$
$$\boxed{D_2(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_2)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_2) = 31.378}$$
$$D_3(\mathbf{X}_0) = (\mathbf{X}_0 - \bar{\mathbf{X}}_3)^T \mathbf{S}_P^{-1} (\mathbf{X}_0 - \bar{\mathbf{X}}_3) = 65.452$$

Smallest distance corresponds to **Group 2**, so we would classify this new flower as belonging to Group 2.