Discriminant Analysis

Data Mining

Prof. Sujee Lee

Department of Systems Management Engineering

Sungkyunkwan University

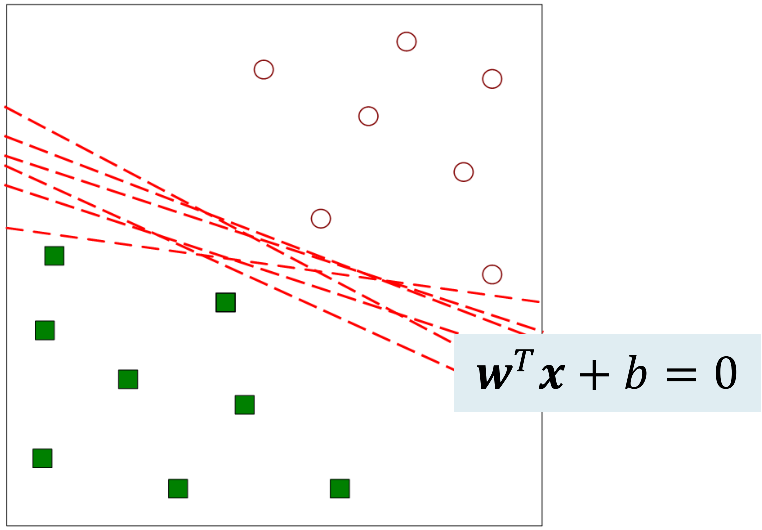
(Recap.) Logistic Regression

- **(Recap.) Logistic Regression**
  - Use linear model $\mathbf{w}^T \boldsymbol{x} + b$, but $f(\boldsymbol{x})$ to be a probability

$$\hat{y} = f(\boldsymbol{x}) = \sigma(\mathbf{w}^T \boldsymbol{x} + b) = \frac{1}{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x} - b)}$$

$$\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d, \qquad y \in \mathbb{B}, \qquad 0 \le \hat{y} \le 1$$

  - This algorithm is a linear classifier



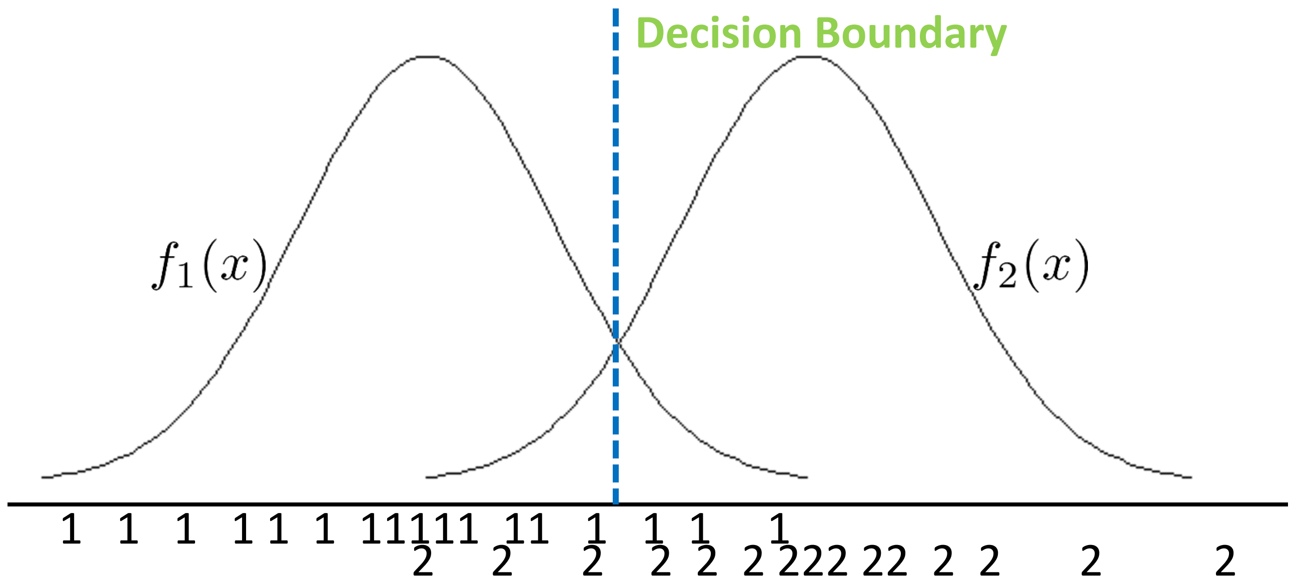
Linear Discriminant Analysis (LDA)

- **Univariate & Two-Class Case**
  - Normal Distribution Assumption

$$X \sim N(\mu, \sigma^2),\ P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

  - Two Classes

$$f_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}} \quad , \quad f_2(x) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{\frac{-(x-\mu_2)^2}{2\sigma_2^2}}$$



LDA in Univariate & Two-Class Case

- **Classification Rule**
  - Assign an observation to class 1, if $$\frac{f_1(x)}{f_2(x)} \geq 1$$
  - Assign an observation to class 2, otherwise.

- **A Further Assumption: Equal variances**: $\sigma_1 = \sigma_2 = \sigma$

$$\frac{f_1(x)}{f_2(x)} = \frac{e^{\frac{-(x-\mu_1)^2}{2\sigma^2}}}{e^{\frac{-(x-\mu_2)^2}{2\sigma^2}}} = e^{-\frac{(x-\mu_1)^2-(x-\mu_2)^2}{2\sigma^2}}$$

$$\ln \frac{f_1(x)}{f_2(x)} = -\frac{(x-\mu_1)^2-(x-\mu_2)^2}{2\sigma^2} \geq 0$$

- **Again, Classification Rule is**
  - Assign an observation to class 1, if $(x-\mu_1)^2 \leq (x-\mu_2)^2$

LDA in Multivariate & Two-Class Case

- **Multivariate & Two-Class Case**
  - Mean vector & var-covariance matrix

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & Cov(X_2, X_p) & \cdots & Var(X_p) \end{bmatrix}$$

  - Multivariate Normal Distribution

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)} \;, \quad i = 1, 2$$

LDA in Multivariate & Two-Class Case

- **Classification Rule**
  - Assign an observation to class 1, if $$\frac{f_1(x)}{f_2(x)} \geq 1$$
  - Assign an observation to class 2, otherwise.

- **A Further Assumption: Equal var-covariance matrices:** $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

$$\frac{f_1(x)}{f_2(x)} = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}}{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}} = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}$$

$$\ln \frac{f_1(x)}{f_2(x)} = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \geq 0$$

- **Again, the classification rule is**
  - Assign an observation to class 1, if

$$(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \leq (\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$$

Linear Discriminant Analysis

- **Why Is This Called LDA (Linear Discriminant Analysis)?**
  - **Univariate Case**

$$(x-\mu_1)^2 \leq (x-\mu_2)^2 \iff (\mu_1-\mu_2)x \geq \frac{1}{2}(\mu_1^2-\mu_2^2)$$

  - **Multivariate Case**

$$(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \leq (\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$$

$$\iff (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \geq \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$$

$$\iff (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \geq \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$$

Fisher's Method

- **Fisher's Linear Discriminant**
  - To find a linear combination of the independent variables $Z = w^T X$ such the between-class variance is maximized relative to the within-class variance.

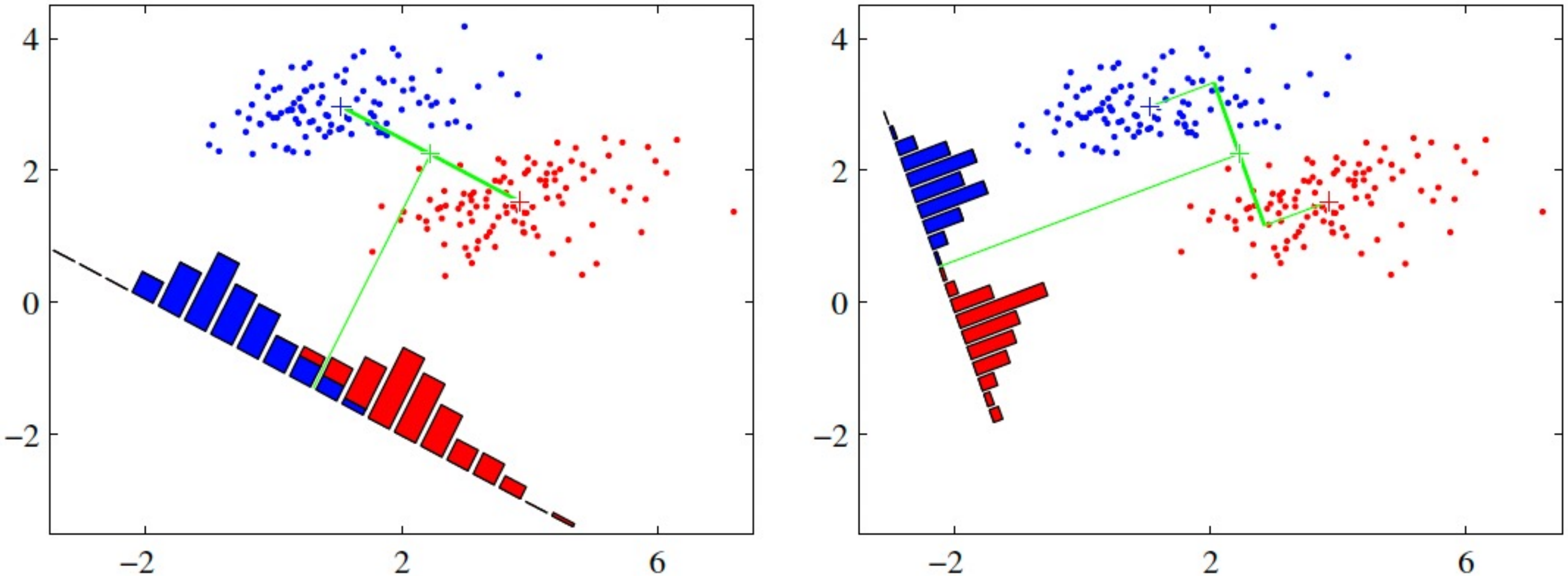


Figure 4.6 The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

Fisher's Method

Fisher's Discriminant Function

- Discriminant function: $Z = \mathbf{w}^\top \mathbf{x} = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in C_k} \mathbf{x}_n, \qquad m_k = \mathbf{w}^{\mathrm{T}} \mathbf{m}_k, \qquad s_k^2 = \sum_{n \in C_k} (z_n - m_k)^2$$

- Find $\mathbf{w}$ so as to maximize

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}.$$ which is equivalent to $$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{B}} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{W}} \mathbf{w}}$$

where

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}.$$

- Solution is :

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1} (\mathbf{m}_2 - \mathbf{m}_1)$$

$$\rightarrow \mathbf{w} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \text{or} \quad \mathbf{w}^\top = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}$$

Fisher's Method

- **Classification Rule**

$$Z_1 = \mathbf{w}^\top \boldsymbol{\mu}_1 \ , \ Z_2 = \mathbf{w}^\top \boldsymbol{\mu}_2$$

  - Assign an observation into class 1, if

$$\begin{aligned} & (\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \boldsymbol{\mu}_1)^2 \leq (\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \boldsymbol{\mu}_2)^2 \\ \Longleftrightarrow \ & (Z - Z_1)^2 \leq (Z - Z_2)^2 \\ \Longleftrightarrow \ & Z \geq \frac{Z_1 + Z_2}{2} \\ \Longleftrightarrow \ & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} \geq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \end{aligned}$$

  - Assign an observation to class 2, otherwise.

Computations for LDA

- **Estimating sample mean & var-covariance matrix**
  - Class 1: $\mathbf{x}_i^{(1)}\ , \ i = 1, 2, \ldots, n_1$
  - Class 2: $\mathbf{x}_i^{(2)}\ , \ i = 1, 2, \ldots, n_2$
  - Sample mean for each group

$$\boldsymbol{\mu}_i = \bar{\mathbf{x}}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}_i^{(j)}\ , \ j = 1, 2$$

  - Pooled variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

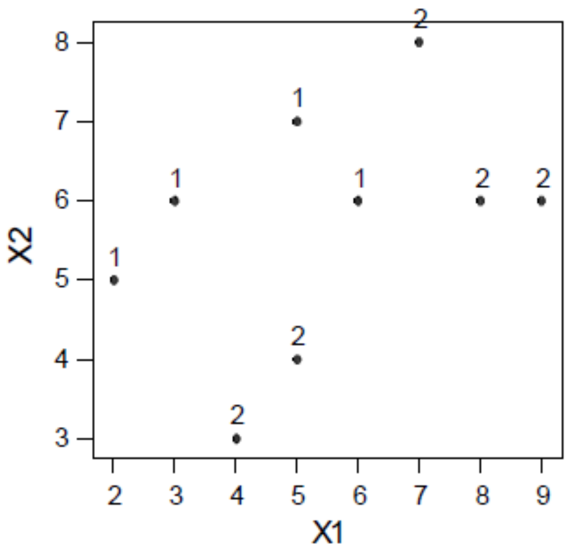
where

$$\mathbf{S}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)})(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)})^\top\ , \quad i = 1, 2$$

Fisher's Method – Example

Data

object	X1	X2	Class
1	5	7	1
2	4	3	2
3	7	8	2
4	8	6	2
5	3	6	1
6	2	5	1
7	6	6	1
8	9	6	2
9	5	4	2



$$\bar{\mathbf{x}}^{(1)} = \begin{pmatrix} 4.0 \\ 6.0 \end{pmatrix}, \ \bar{\mathbf{x}}^{(2)} = \begin{pmatrix} 6.6 \\ 5.4 \end{pmatrix}$$

$$\mathbf{S}_1 = \begin{pmatrix} 3.33 & 1.00 \\ 1.00 & 0.66 \end{pmatrix}$$

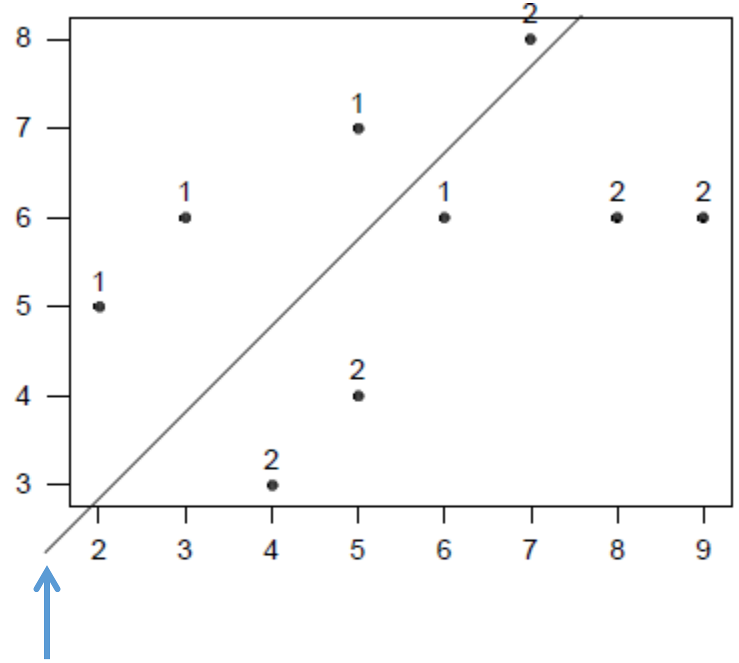
$$\mathbf{S}_2 = \begin{pmatrix} 4.30 & 2.95 \\ 2.95 & 3.80 \end{pmatrix}$$

$$\mathbf{S}_p = \begin{pmatrix} 3.8857 & 2.1143 \\ 2.1143 & 2.4571 \end{pmatrix}$$

$$\mathbf{w} = \mathbf{S}_p^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) = \begin{pmatrix} 0.4839 & -0.4164 \\ -0.4164 & 0.7653 \end{pmatrix} \begin{pmatrix} -2.6 \\ 0.6 \end{pmatrix} = \begin{pmatrix} -1.5080 \\ 1.5418 \end{pmatrix}$$

Fisher's Method – Example

object	X1	X2	Z	True Class	Predicted Class
1	5	7	3.2526	1	1
2	3	6	4.7268	1	1
3	2	5	4.6929	1	1
4	6	6	0.2026	1	2
5	4	3	-1.4068	2	2
6	7	8	1.7781	2	1
7	8	6	-2.8135	2	2
8	9	6	-4.3215	2	2
9	5	4	-1.3730	2	2



$$-1.5080x_1 + 1.5418x_2 = 0.7958$$

Quadratic Discriminant Analysis (QDA)

- **Unequal var-covariance matrices:** $\Sigma_1 \neq \Sigma_2$

$$\frac{f_1(x)}{f_2(x)} = \frac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}} \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}}{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}} = \frac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}$$

$$\ln \frac{f_1(x)}{f_2(x)} = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \geq 0$$

- **Classification Rule**
  - Assign an observation into class 1, if

$$-\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) - \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \geq 0$$

  - Assign an observation into class 2, otherwise.

- **Estimating sample mean vectors and var-cov matrices**

$$\hat{\boldsymbol{\mu}}_1 = \overline{\mathbf{x}}^{(1)}, \quad \hat{\boldsymbol{\mu}}_2 = \overline{\mathbf{x}}^{(2)}$$

$$\hat{\boldsymbol{\Sigma}}_1 = \mathbf{S}_1, \quad \hat{\boldsymbol{\Sigma}}_2 = \mathbf{S}_2$$

QDA – Example

- **The previous example revisited**

$$\bar{\mathbf{x}}^{(1)} = \begin{pmatrix} 4.0 \\ 6.0 \end{pmatrix}, \bar{\mathbf{x}}^{(2)} = \begin{pmatrix} 6.6 \\ 5.4 \end{pmatrix}$$

$$\mathbf{S}_1 = \begin{pmatrix} 3.33 & 1.00 \\ 1.00 & 0.66 \end{pmatrix} \quad \mathbf{S}_1^{-1} = \begin{pmatrix} 0.5455 & -0.8182 \\ -0.8182 & 2.7273 \end{pmatrix} \quad |\mathbf{S}_1| = 1.2222$$

$$\mathbf{S}_2 = \begin{pmatrix} 4.30 & 2.95 \\ 2.95 & 3.80 \end{pmatrix} \quad \mathbf{S}_2^{-1} = \begin{pmatrix} 0.4975 & -0.3863 \\ -0.3863 & 0.5630 \end{pmatrix} \quad |\mathbf{S}_2| = 7.6375$$

- **The classification rule rewritten**
  - Assign an observation into class 1, if $U_1 \geq U_2$

where
$$U_1 = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1|$$

$$U_2 = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_2|$$

QDA – Example

object	X1	X2	U1	U2	True Class	Predicted Class
1	5	7	-0.9156	-3.3617	1	1
2	3	6	-0.3713	-5.1755	1	1
3	2	5	-0.9174	-5.6144	1	1
4	6	6	-1.1885	-1.3457	1	1
5	4	3	-12.3722	-1.9094	2	2
6	7	8	-3.0959	-2.5561	2	2
7	8	6	-4.4603	-1.2805	2	2
8	9	6	-6.9143	-1.9943	2	2
9	5	4	-7.4625	-1.3397	2	2

$$U_1 = -0.2727x_1^2 + 0.8182x_1x_2 - 1.3636x_2^2 - 2.7273x_1 + 13.0909x_2 - 33.9185$$

$$U_2 = -0.2488x_1^2 + 0.3863x_1x_2 - 0.2815x_2^2 + 1.198x_1 + 0.491x_2 - 6.2957$$

Classification of Three or More Classes

- **Multi-class LDA**
  - Pooled variance-covariance matrix and mean vectors

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S}_p = \frac{\sum_{j=1}^{J}(n_j - 1)\mathbf{S}_j}{\sum_{j=1}^{J}(n_j - 1)}\,,\ \hat{\boldsymbol{\mu}}_j = \bar{\mathbf{x}}^{(j)}\,,\ j = 1, 2, \ldots, J$$

- **Multi-class QDA**
  - Var-covariance matrices and mean vectors

$$\hat{\boldsymbol{\Sigma}}_j = \mathbf{S}_j\,,\ \hat{\boldsymbol{\mu}}_j = \bar{\mathbf{x}}^{(j)}\,,\ j = 1, 2, \ldots, J$$

Under assumption that prior probabilities are same

- **Classification Rule**
  - Assign an object into class j if $\dfrac{f_j(\mathbf{x})}{\sum_{j=1}^{J} f_j(\mathbf{x})}$ is maximal.
  - For QDA, the above rule is equivalent to assigning an object into class j if U_j is maximal, where

$$U_j = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j|$$