Chapter 2. Classifiers Based on Bayes Decision Theory

Chang-Su Kim

Many slides are modified from Serigos Theodoridis's own notes.

Classification Problem

- There are M classes: $\omega_1, \ldots, \omega_M$
- Given a pattern with feature vector $\mathbf{x}$, classify it into one of the classes

BAYESIAN CLASSIFICATION

Bayesian Classification Rule

- Classify $\mathbf{x}$ into ω_i if (1)

$$P(\omega_i|\mathbf{x}) > P\left(\omega_j\middle|\mathbf{x}\right)$$

for all j

Bayesian Classification Rule

- Classify $\mathbf{x}$ into ω_{i^*} where (2)

$$i^* = \arg\max_i P(\omega_i|\mathbf{x})$$

  - $P(\omega_i)$: $a\ priori$ probability
  - $P(\omega_i|\mathbf{x})$: *a posteriori* probability
  - $P(\mathbf{x}|\omega_i)$: likelihood of ω_i with respect to $\mathbf{x}$
  - Bayesian decision is also called **maximum *a posteriori* (MAP) decision**

Bayesian Classification Rule

- Bayes rule

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_j P\left(\mathbf{x}\middle|\omega_j\right)P(\omega_j)}$$

- Classify $\mathbf{x}$ into ω_{i^*} where (3)

$$i^* = \arg\max_i P(\mathbf{x}|\omega_i)P(\omega_i)$$

- When all prior probabilities are identical, this becomes
  - Classify $\mathbf{x}$ into ω_{i^*} where

$$i^* = \arg\max_i P(\mathbf{x}|\omega_i)$$

  - This is the **maximum likelihood (ML) decision**

Bayesian Classification Rule

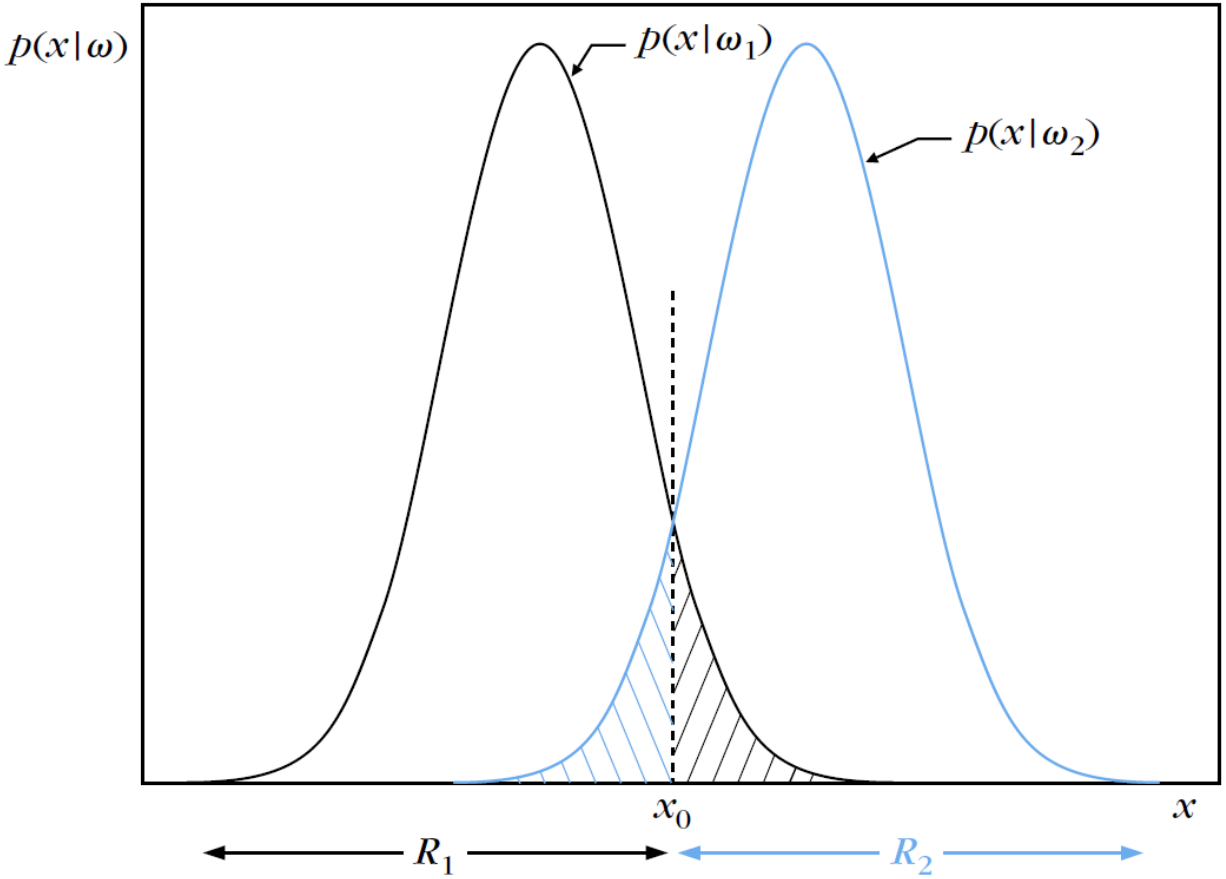


FIGURE 2.1

Example of the two regions R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes.

Bayesian classifier minimizes classification error probability

- Two-class problem
  - Classification error probability

$$P_e = P(\mathbf{x} \in R_2, \omega_1) + P(\mathbf{x} \in R_1, \omega_2)$$

  - To minimize P_e,

$$R_1 = \{\mathbf{x}: P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})\}$$
$$R_2 = \{\mathbf{x}: P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})\}$$

- The Bayesian classifier is optimal in that it minimizes P_e

Discriminant Functions and Decision Surfaces

- If R_i, R_j are contiguous, they are separated by a **decision surface**

$$P(\omega_i|\mathbf{x}) - P\left(\omega_j\middle|\mathbf{x}\right) = 0$$

- Equivalently, the decision surface is given by

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$

where $g_i(\mathbf{x}) \equiv f\left(P(\omega_i|\mathbf{x})\right)$ is a **discriminant function** and f is monotonically increasing

Bayesian Classification for Normal Distributions

- Multivariate Gaussian PDF

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} = E[\mathbf{x}]$ is the mean vector
$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$ is the covariance matrix

Bayesian Classification for Normal Distributions

- Multivariate Gaussian PDF

$$\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

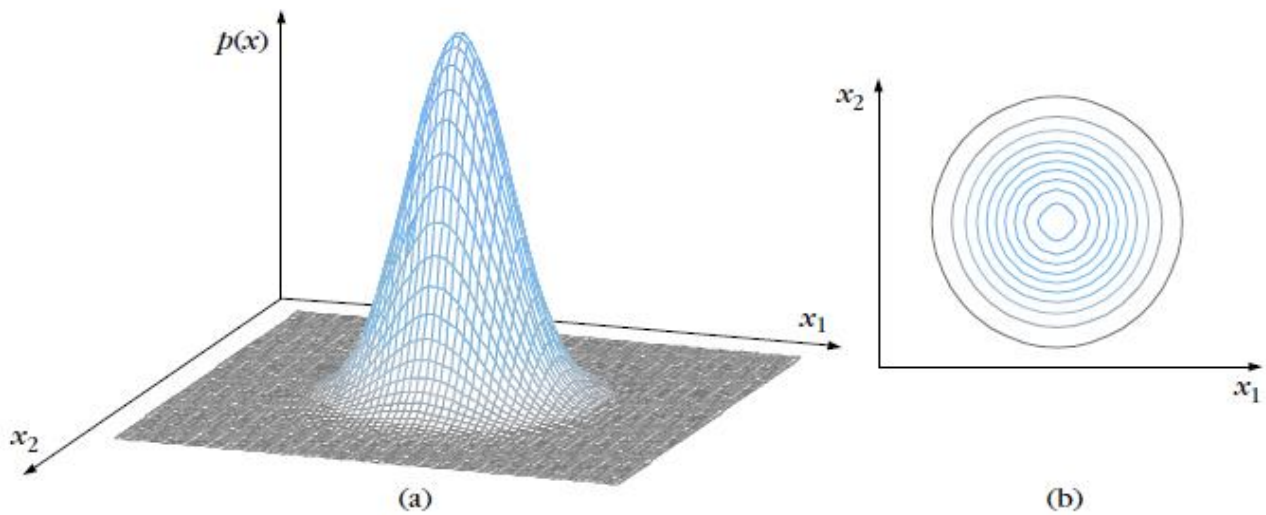


FIGURE 2.3

Bayesian Classification for Normal Distributions

- Multivariate Gaussian PDF

$$\Sigma = \begin{bmatrix} 15 & 0 \\ 0 & 3 \end{bmatrix}$$

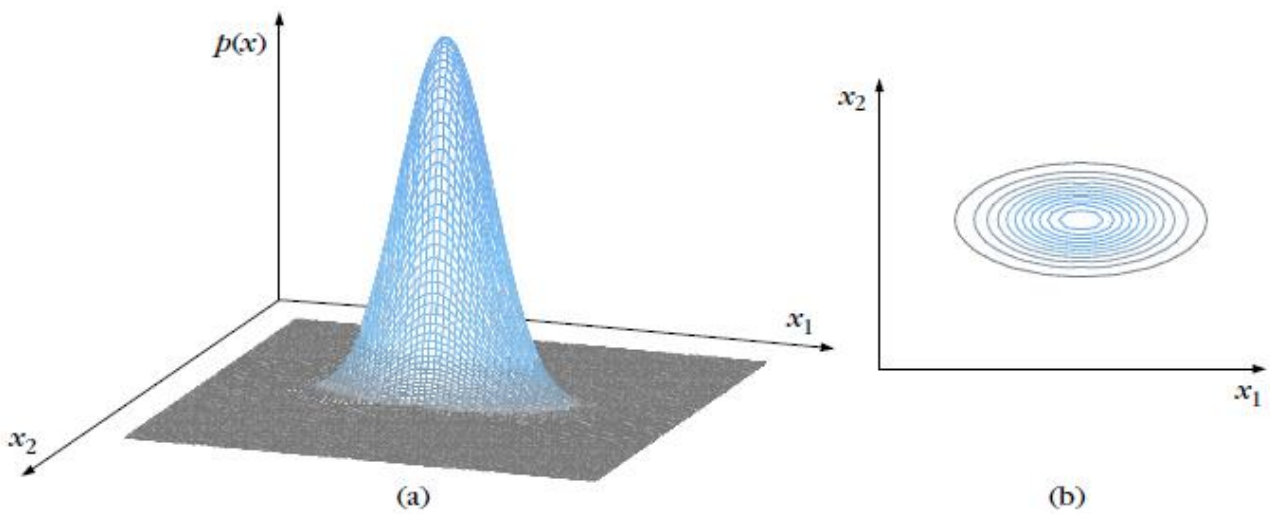


FIGURE 2.4

Bayesian Classification for Normal Distributions

- Multivariate Gaussian PDF

$$\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 15 \end{bmatrix}$$

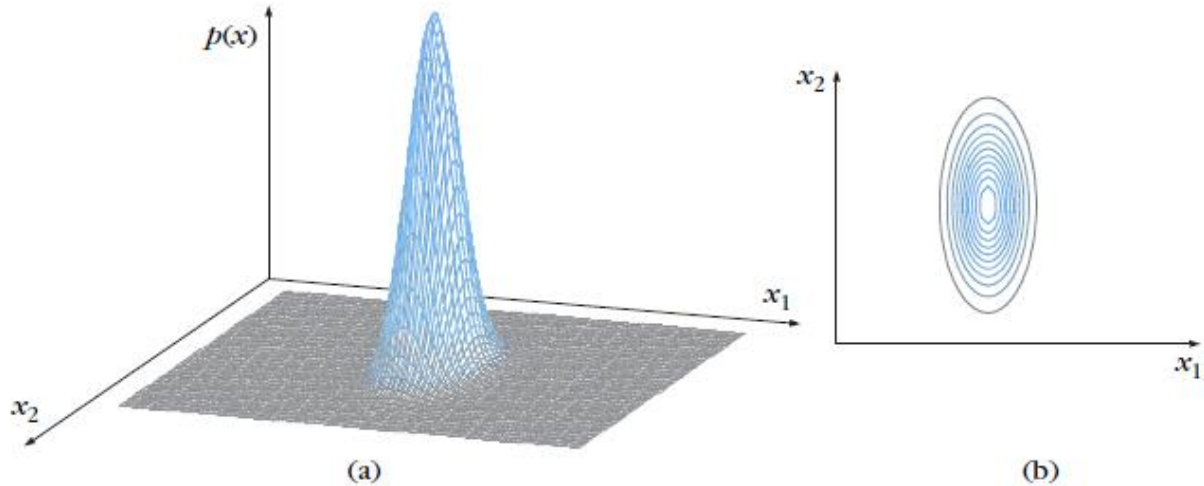


FIGURE 2.5

Bayesian Classification for Normal Distributions

- Multivariate Gaussian PDF

$$\Sigma = \begin{bmatrix} 15 & 6 \\ 6 & 3 \end{bmatrix}$$

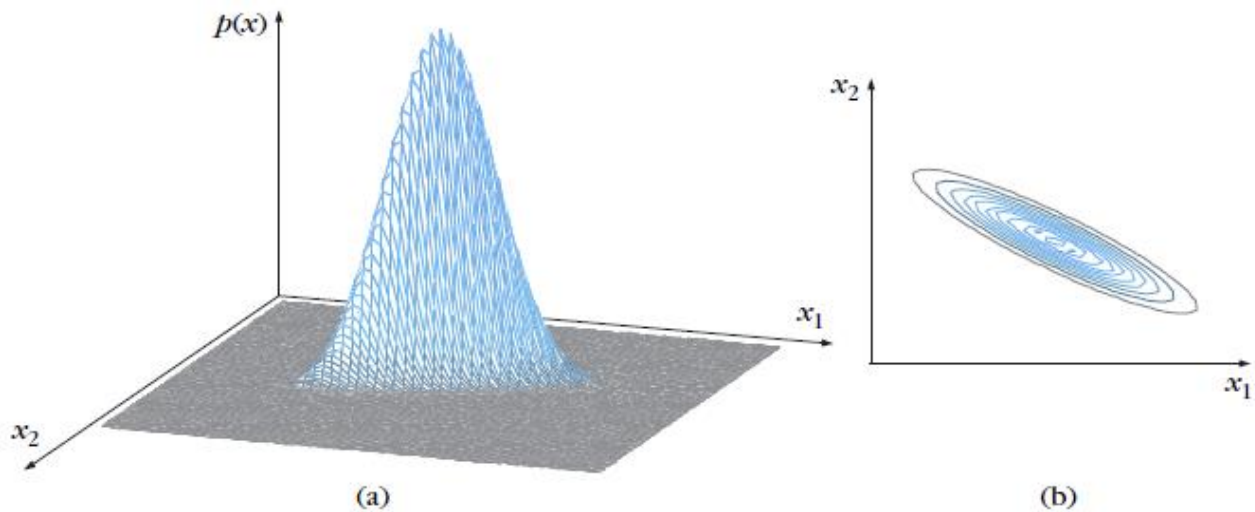


FIGURE 2.6

Normal Distributions

- $P(\mathbf{x}) \sim N(\boldsymbol{\mu}, \Sigma)$

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$$

  where $\boldsymbol{\mu} = E[\mathbf{x}]$ and $\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$

- Σ is symmetric and positive definite, and thus its eigenvalue decomposition

$$\Sigma = \mathrm{Q}\Lambda\mathrm{Q}^T$$

  is possible

- A contour line of equal probability density

$$(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = 1$$

  - It is a hyper-ellipsoid
  - Its principal axes are given by the eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_l$ of Σ
  - Its axes have lengths $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_l}$
  - The main axis with length $\sqrt{\lambda_1}$ is in the direction of $\mathbf{v}_1$

Bayesian Classification for Normal Distributions

- Discriminant function

$$g_i(\mathbf{x}) = \log P(\mathbf{x}|\omega_i)P(\omega_i)$$
$$= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + C_i$$

- Thus, decision surfaces are quadrics (ellipsoids, parabolas, hyperbolas, and pairs of lines)

Bayesian Classification for Normal Distributions

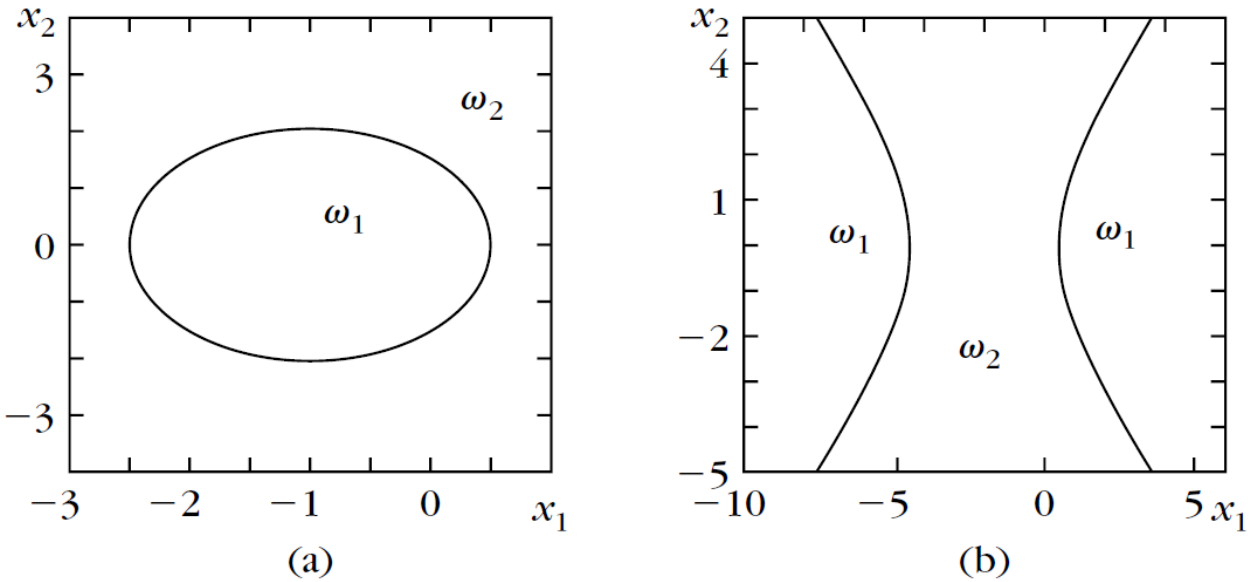


FIGURE 2.7

Examples of quadric decision curves. Playing with the covariance matrices of the Gaussian functions, different decision curves result, that is, ellipsoids, parabolas, hyperbolas, pairs of lines.

Bayesian Classification for Normal Distributions

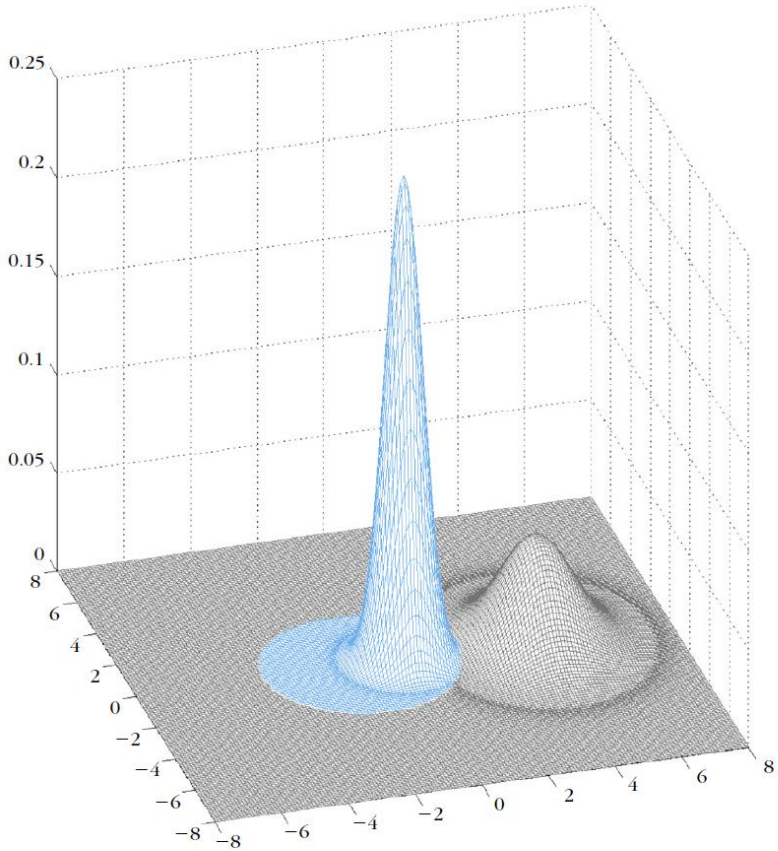
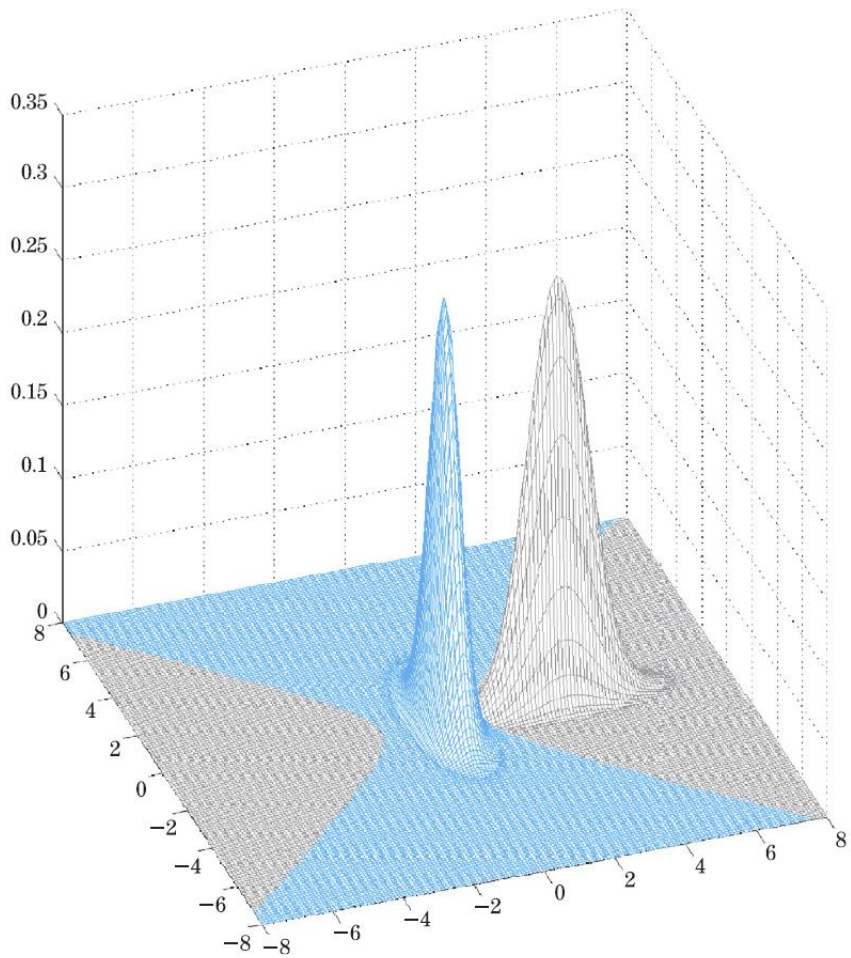


FIGURE 2.8



Special Case I: $\Sigma_i = \sigma^2 \mathrm{I}$

- Decision hyperplane

$$g_{ij}(\mathbf{x}) = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

  - $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$
  - $\mathbf{x}_0 = \frac{1}{2}\left(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j\right) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}$

Special Case I: $\Sigma_i = \sigma^2 \mathrm{I}$

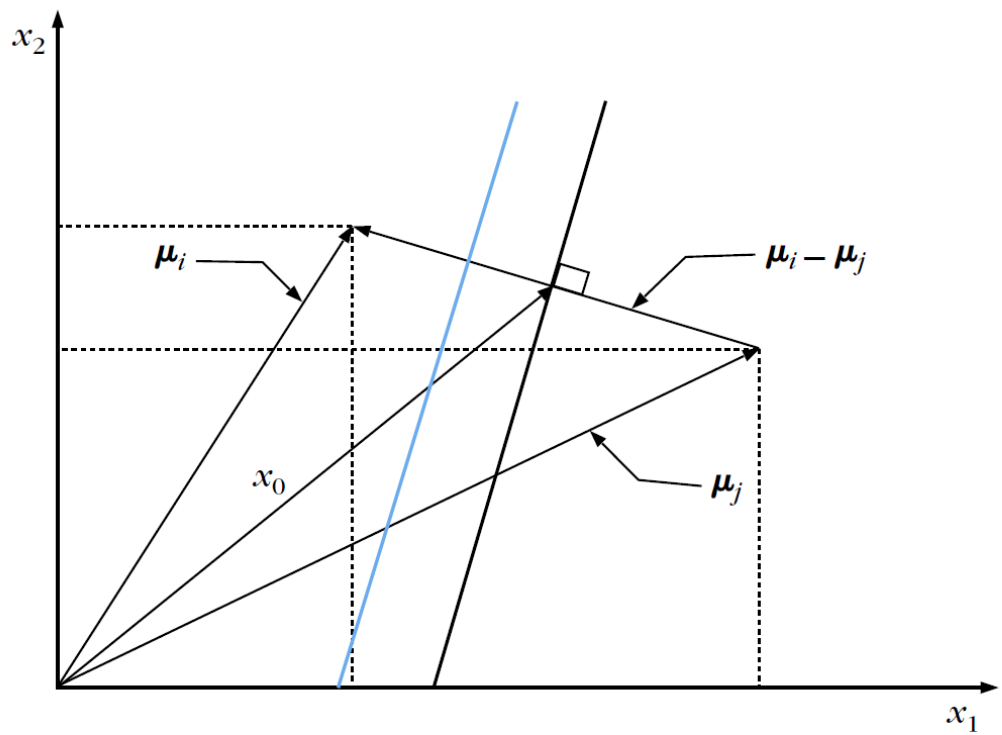


FIGURE 2.10

Decision lines for normally distributed vectors with $\Sigma = \sigma^2 I$. The black line corresponds to the case of $P(\omega_j) = P(\omega_i)$ and it passes through the middle point of the line segment joining the mean values of the two classes. The red line corresponds to the case of $P(\omega_j) > P(\omega_i)$ and it is closer to $\boldsymbol{\mu}_i$, leaving more "room" to the more probable of the two classes. If we had assumed $P(\omega_j) < P(\omega_i)$, the decision line would have moved closer to $\boldsymbol{\mu}_j$.