Chapter 2 (Part 2): Bayesian Decision Theory (Sections 2.3-2.5)

Classifiers, Discriminant Functions and Decision Surfaces

Error Bounds

Quiz

- Select the optimal decision where: $W = \{\omega_1, \omega_2\}$
- $P(x \mid \omega_1)$ → $N(2, 0.5)$ (Normal distribution)
- $P(x \mid \omega_2)$ → $N(1.5, 0.2)$
- $P(\omega_1) = 2/3$
- $P(\omega_2) = 1/3$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

The Normal Density

- Univariate density
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

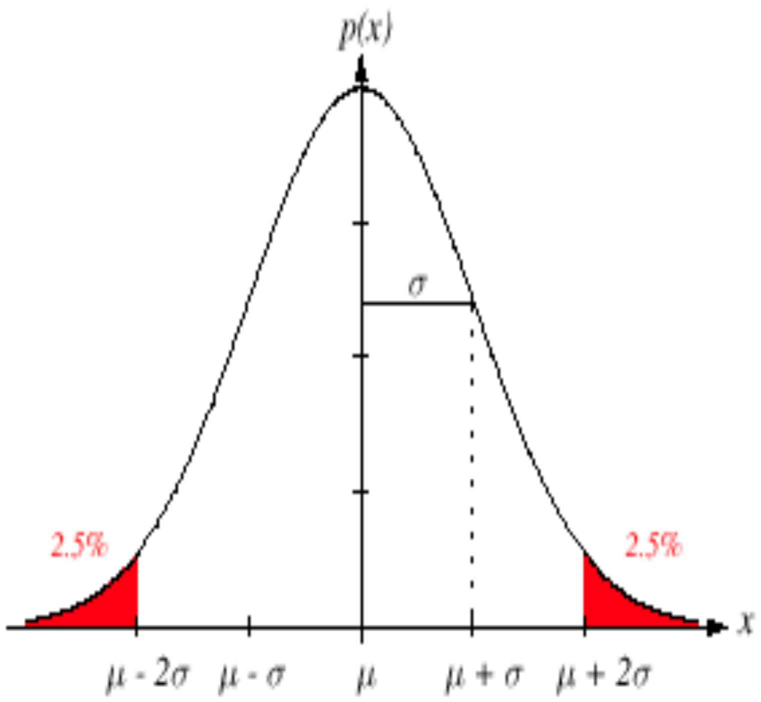


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Normal Density

- Multivariate density: Multivariate normal density in d dimensions is:

$$N(x;\mu,\sigma^2) = \frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right]$$

$x = (x_1, x_2, \ldots, x_d)^t$

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d{*}d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

- We saw that the minimum error-rate classification can be achieved by the discriminant function
- gi(x) = ln P(x | ωi) + ln P(ωi)
- Case of multivariate normal

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Discriminant Functions for the Normal Density…

- Case $\Sigma_i = \sigma^2.I$ (I stands for the identity matrix)
  - $\sigma_{ij} = 0$ (statistically independent) and σ_{ii} are same for all features

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\mu_i^t\mathbf{x} + \mu_i^t\mu_i] + \ln P(\omega_i)$$

Discriminant Functions for the Normal Density…

- Disregarding $x^t x$, we get a linear discriminant function

$$g_i(x) = w_i^t x + w_{i0}$$

where :

$$w_i = \frac{\mu_i}{\sigma^2}; \quad w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t \mu_i + \ln P(\omega_i)$$

(ω_{i0} is called the threshold for the ***i*** th category!)

Discriminant Functions for the Normal Density…

- A classifier that uses linear discriminant functions is called "a linear machine"
- The decision surfaces for a linear machine are hyperplanes defined by $g_i(x) = g_j(x)$

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

where $\mathbf{w} = \mu_i - \mu_j$, and $\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i}{P(\omega_j}(\mu_i - \mu_j)$

Discriminant Functions for the Normal Density…

- The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

is always orthogonal to the line linking the means!

$$\text{if } P(\omega_i) = P(\omega_j) \text{ then } x_0 = \frac{1}{2}(\mu_i + \mu_j)$$

$$g_i(x) = -\|x - \mu_i\|^2$$

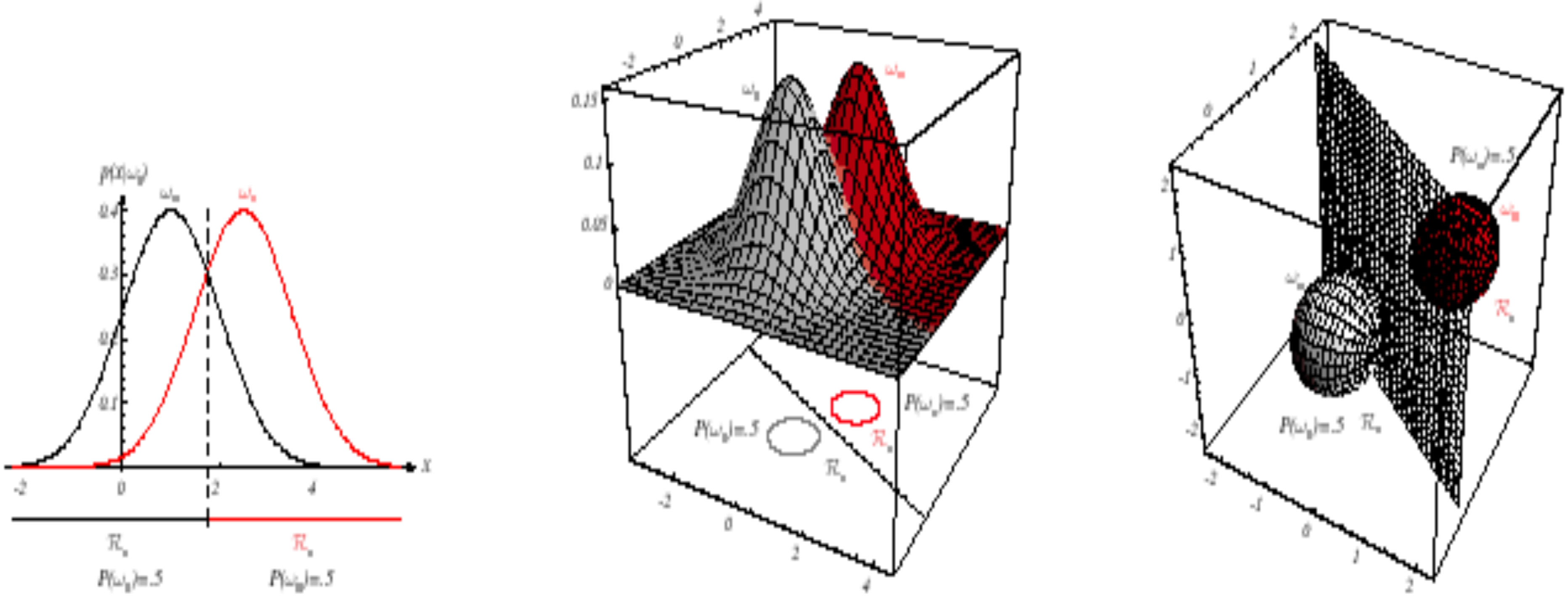
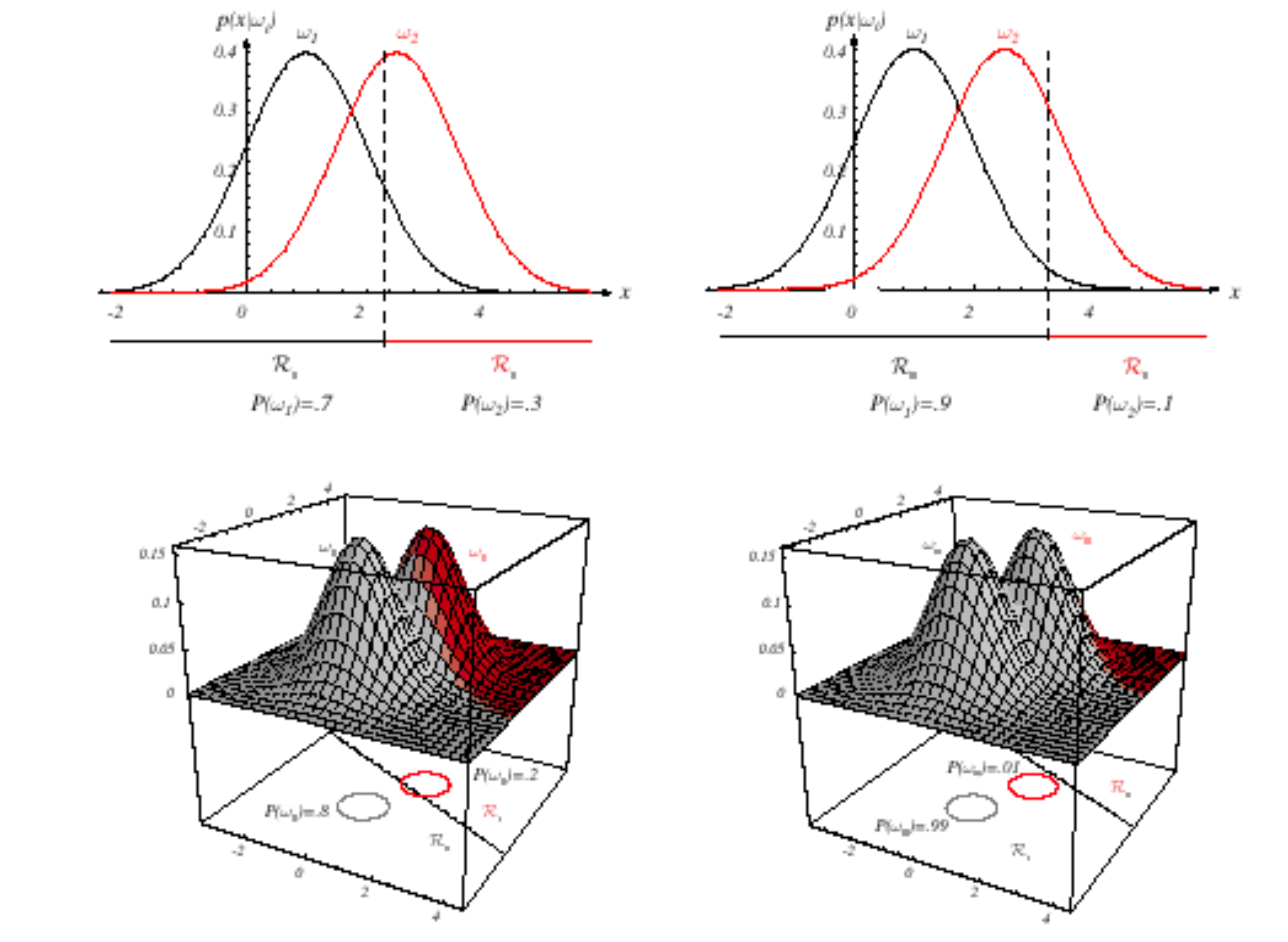


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



p(x|ω_i)
ω_1
ω_2
P(ω_1)=.7
P(ω_2)=.3
P(ω_1)=.9
P(ω_2)=.1
P(ω_2)=.2
P(ω_1)=.8
P(ω_2)=.01
P(ω_1)=.99

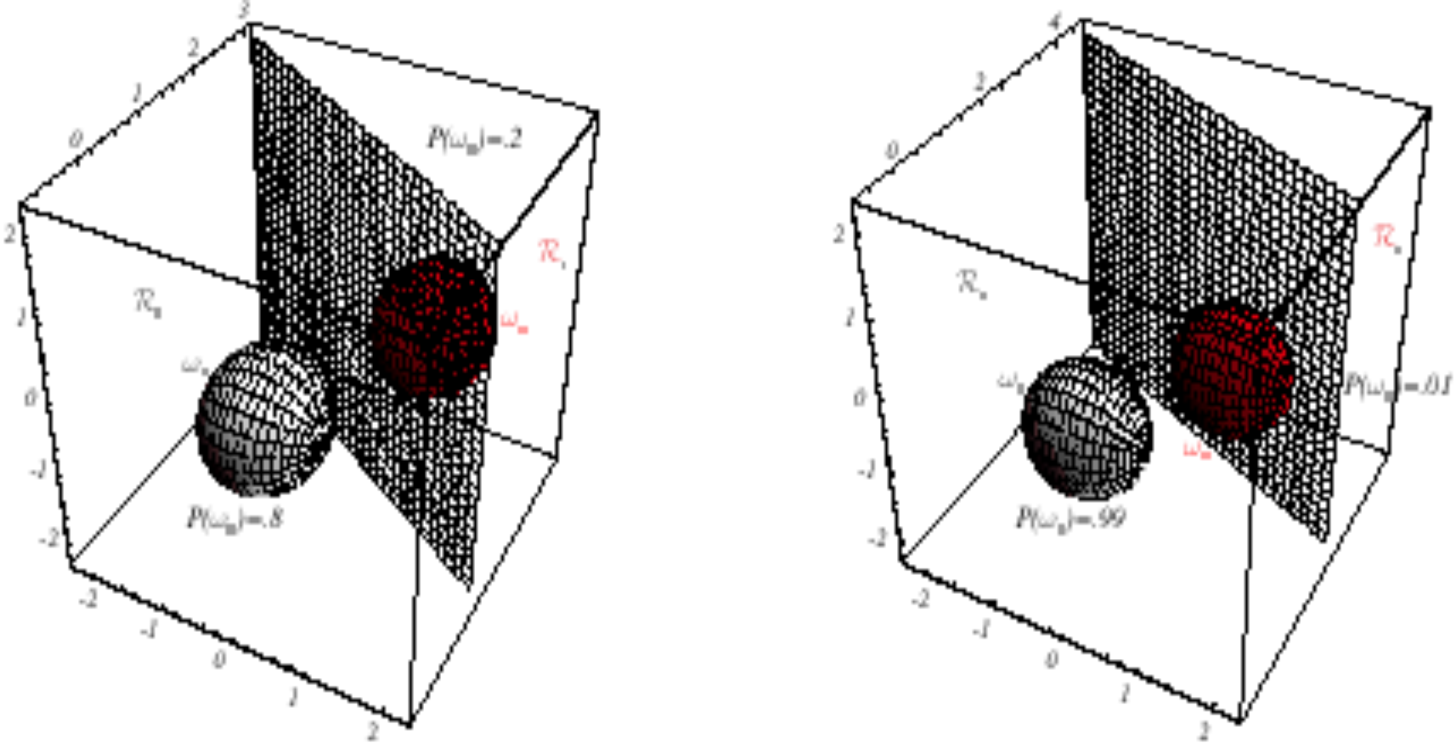


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions for the Normal Density…

- Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- Expand the term and disregard the quadratic expression

$$g_i(x) = w_i^t x + w_{i0}$$

where :

$$w_i = \sum^{-1}\mu; \quad w_{i0} = -\frac{1}{2}\mu_i^t \sum^{-1}\mu_i + \ln P(\omega_i)$$

Discriminant Functions for the Normal Density…

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}.(\mu_i - \mu_j)$$

- Comments about this hyperplane:
  - It passes through $\mathbf{x_0}$
  - It is NOT orthogonal to the line linking the means.
  - What happens when $P(\omega_i)= P(\omega_j)$?
  - If $P(\omega_i)= P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the more likely mean.

Discriminant functions for the normal density…

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

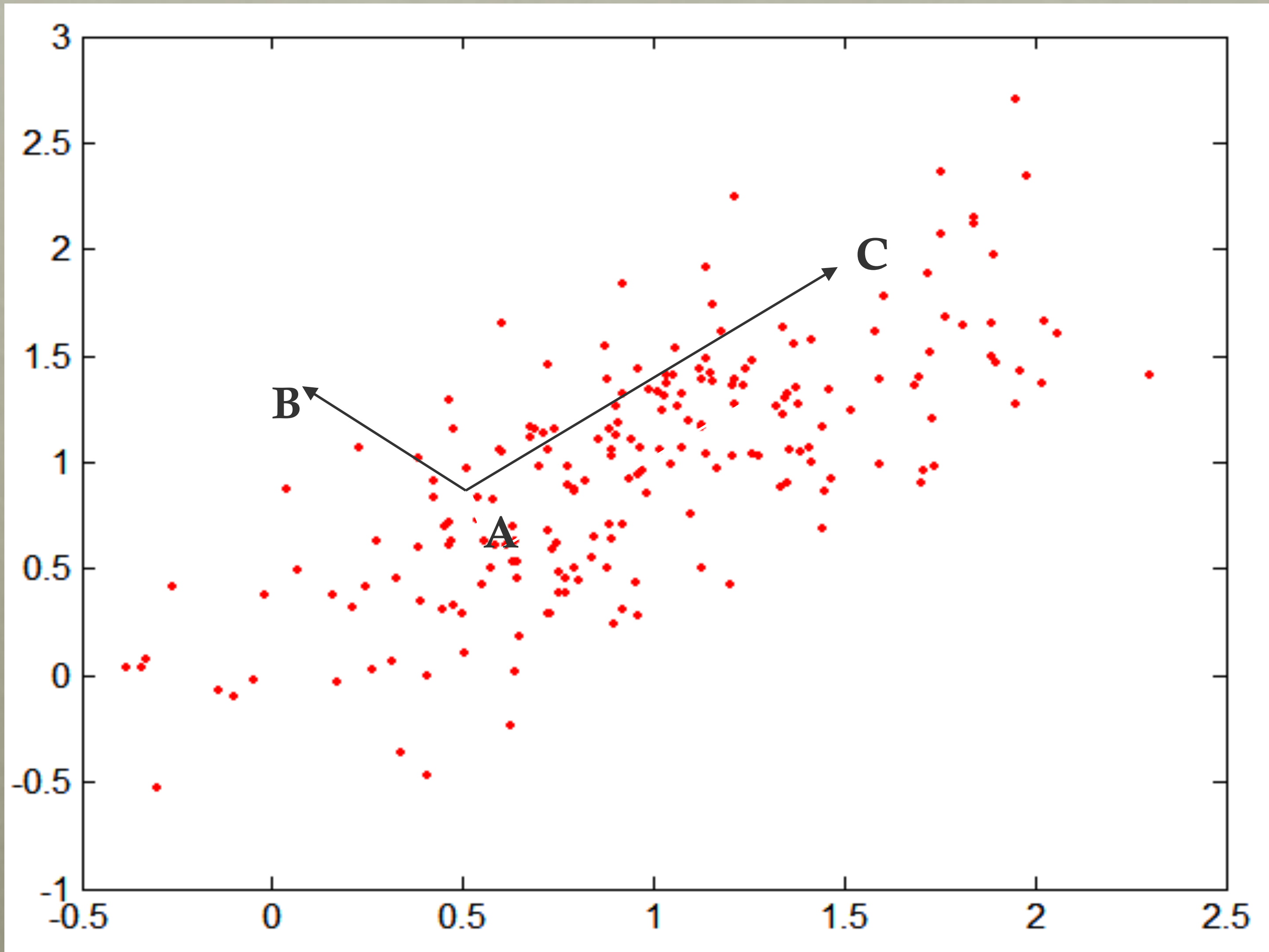
- When $P(\omega_i)$ is the same for each of the c classes
- Case I: Euclidean distance classifier

$$g_i(x) = -\|x-\mu_i\|^2$$

- Case II: Mahalanobis distance classifier

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)$$

MAHALANOBIS DISTANCE



Covariance Matrix:

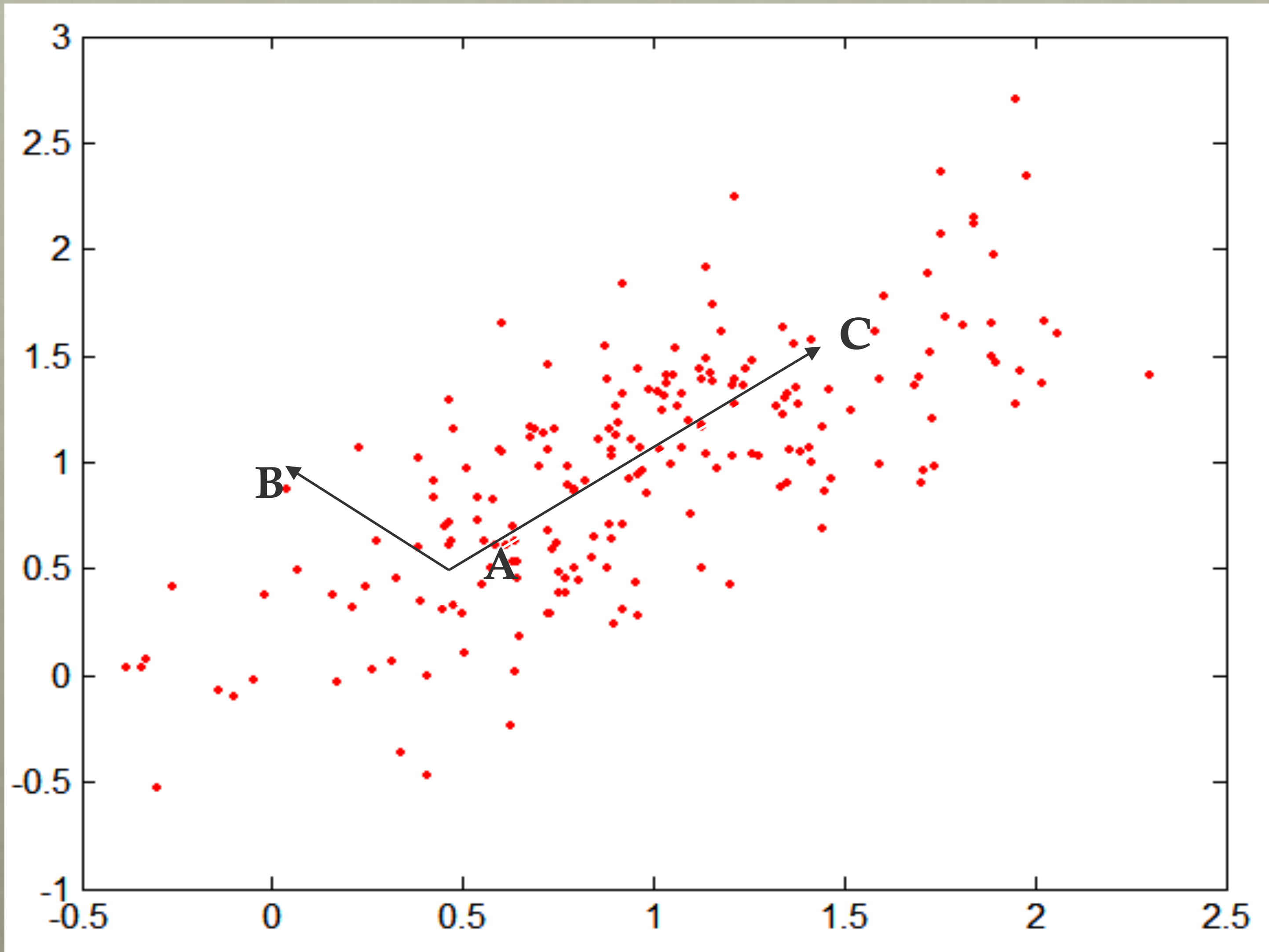
$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

MAHALANOBIS DISTANCE



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

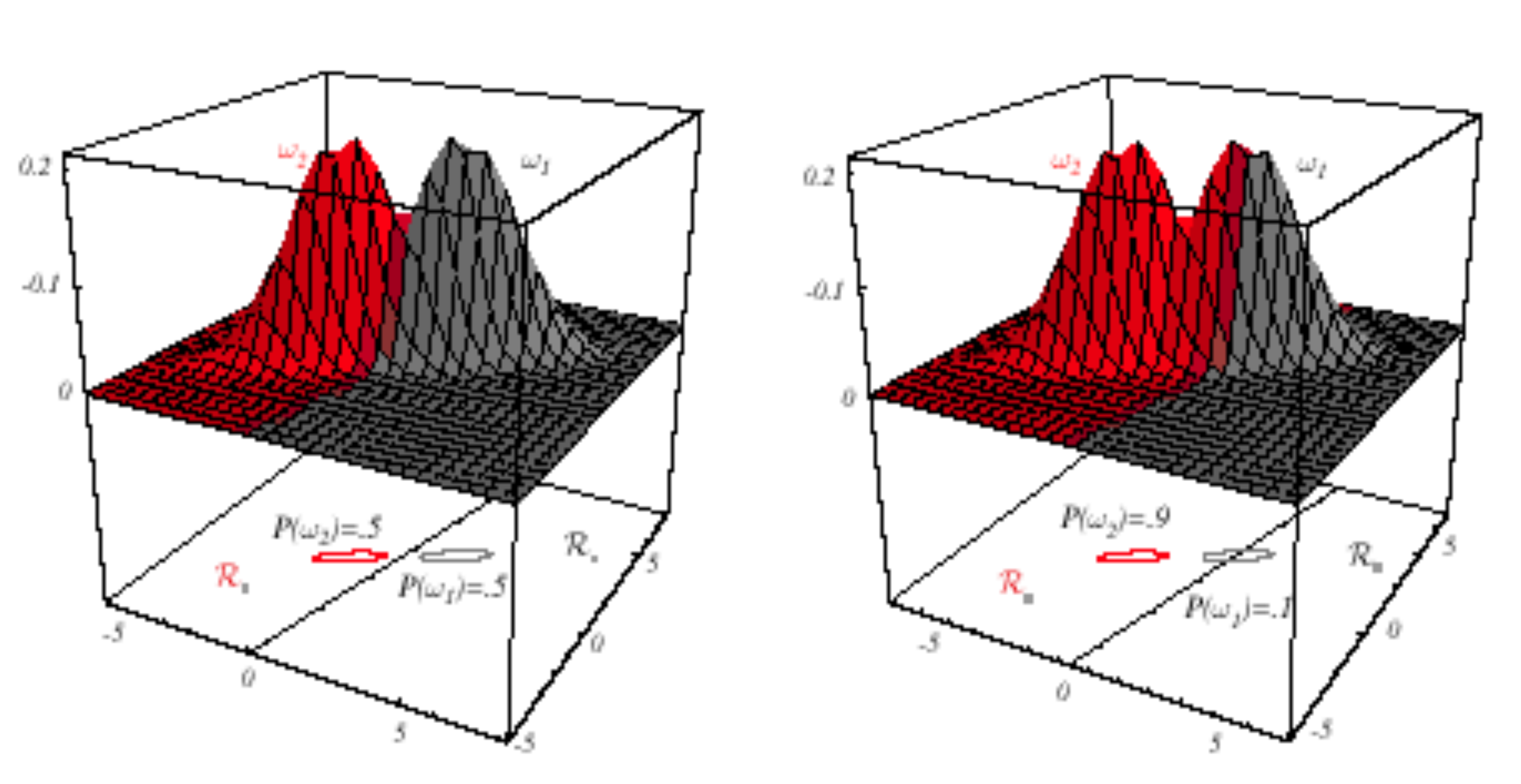
C: (1.5, 1.5)

Euclid(A,B) = 0.5

Euclid(A,B) = 2

Mahal(A,B) = 5

Mahal(A,C) = 4


0.2
0.1
0
ω_2
ω_1
$P(\omega_2)=.5$
$P(\omega_1)=.5$
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
0.2
0.1
0
ω_2
ω_1
$P(\omega_2)=.9$
$P(\omega_1)=.1$
-5
0
5

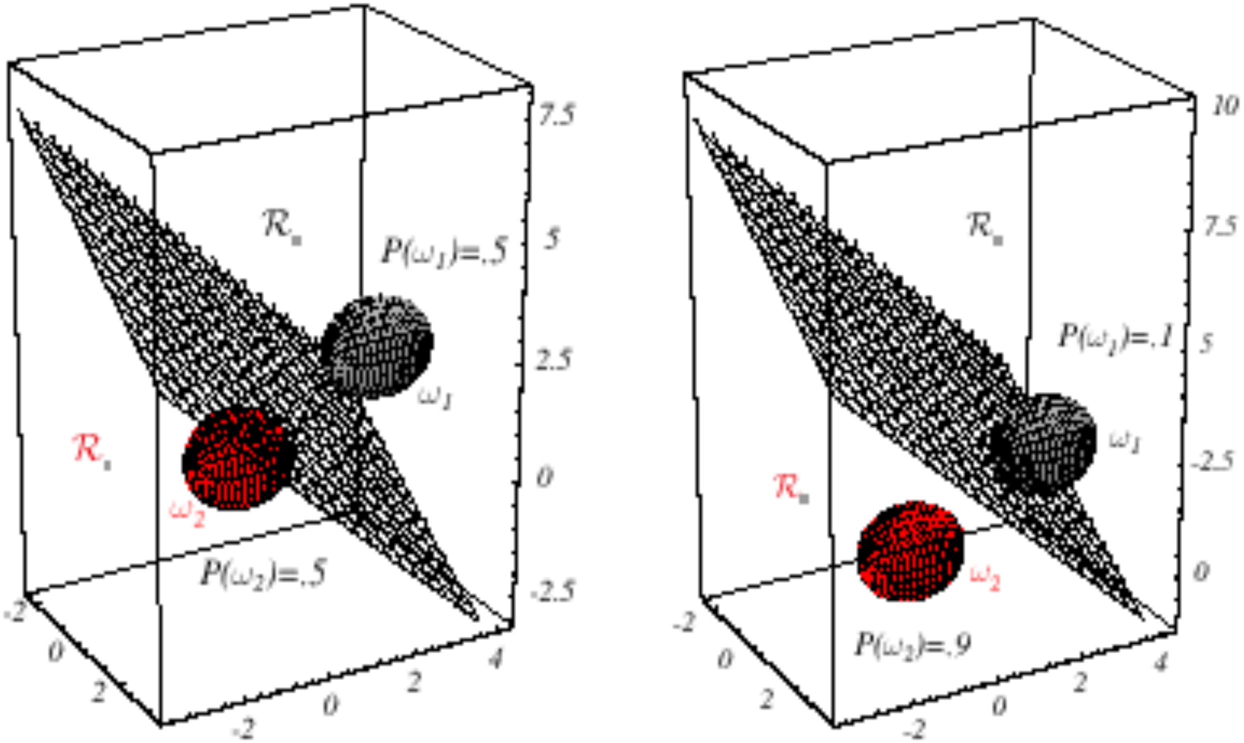


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions for the Normal Density…

- Case Σi = arbitrary
  - The covariance matrices are different for each category

$$g_i(x) = x^t W_i x + w_i^t x = w_{i0}$$

where :

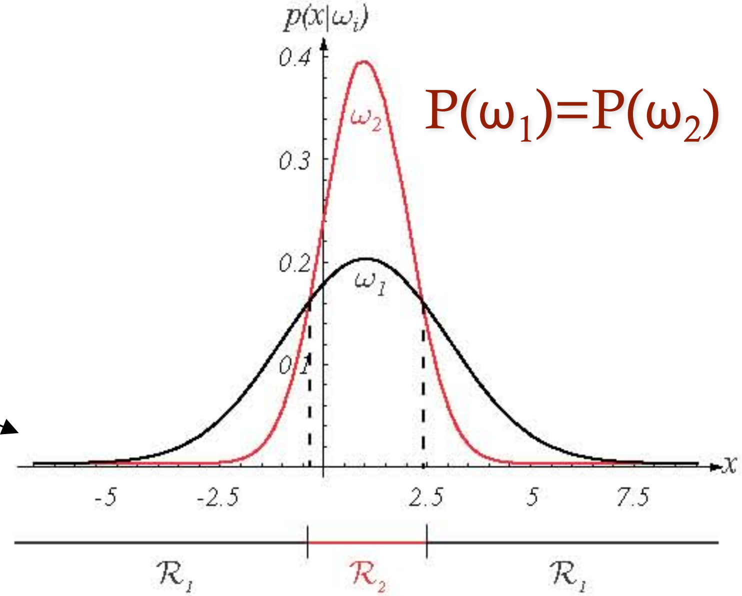
$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

$$w_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

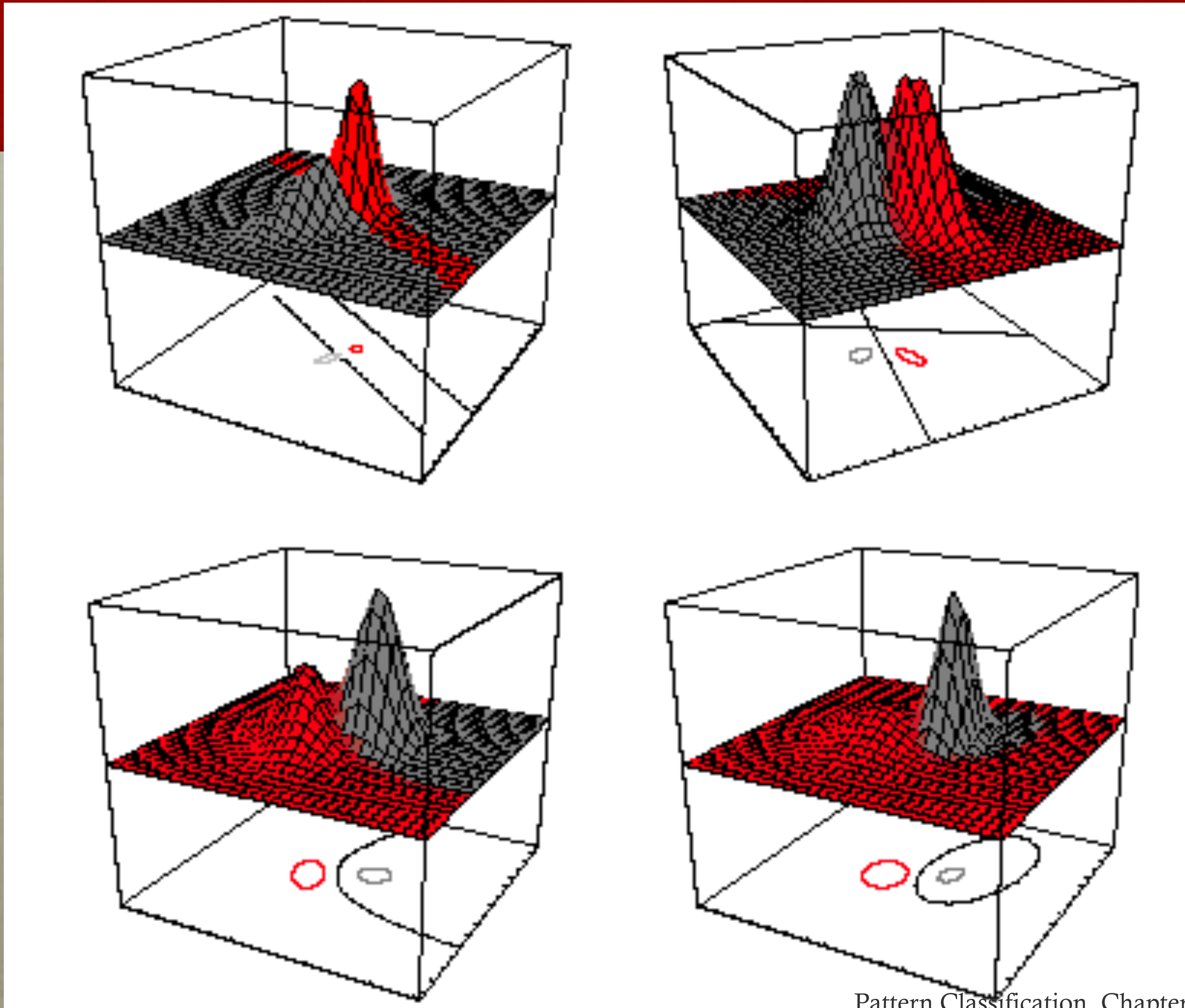
Discriminant Functions for the Normal Density…

Disconnected decision regions



$P(\omega_1)=P(\omega_2)$

FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



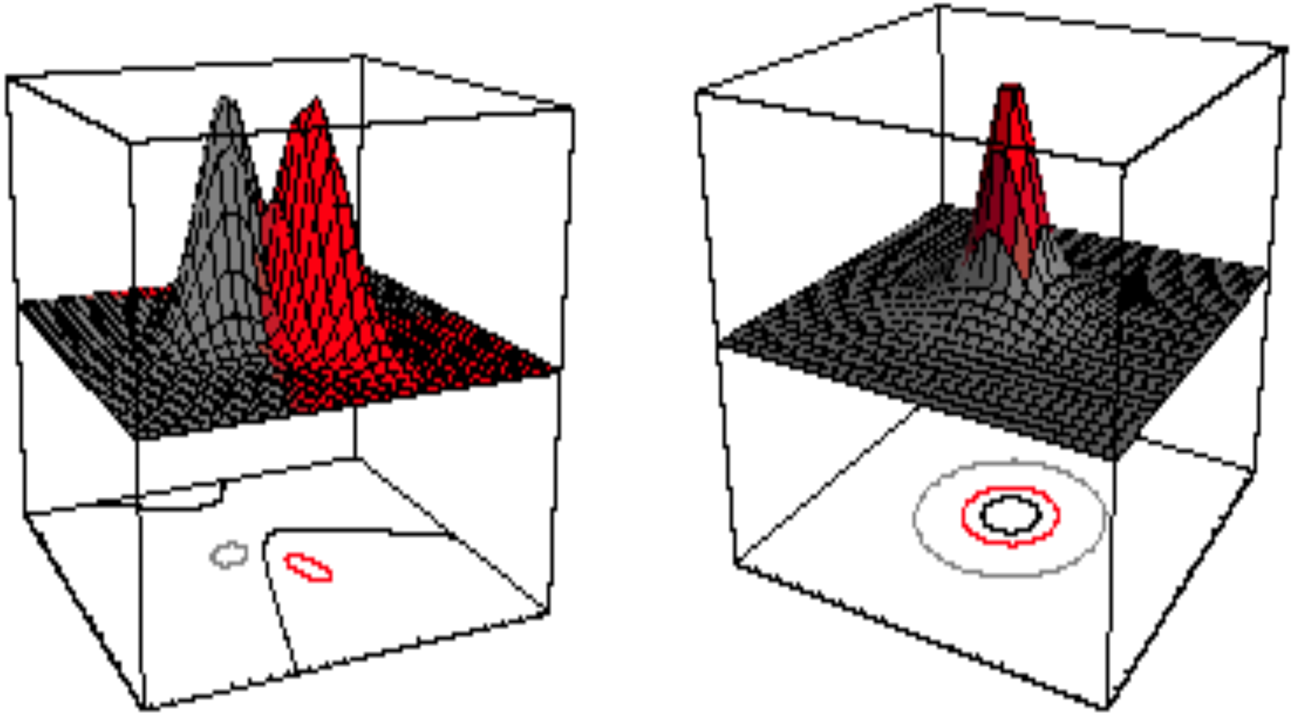


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions for the Normal Density

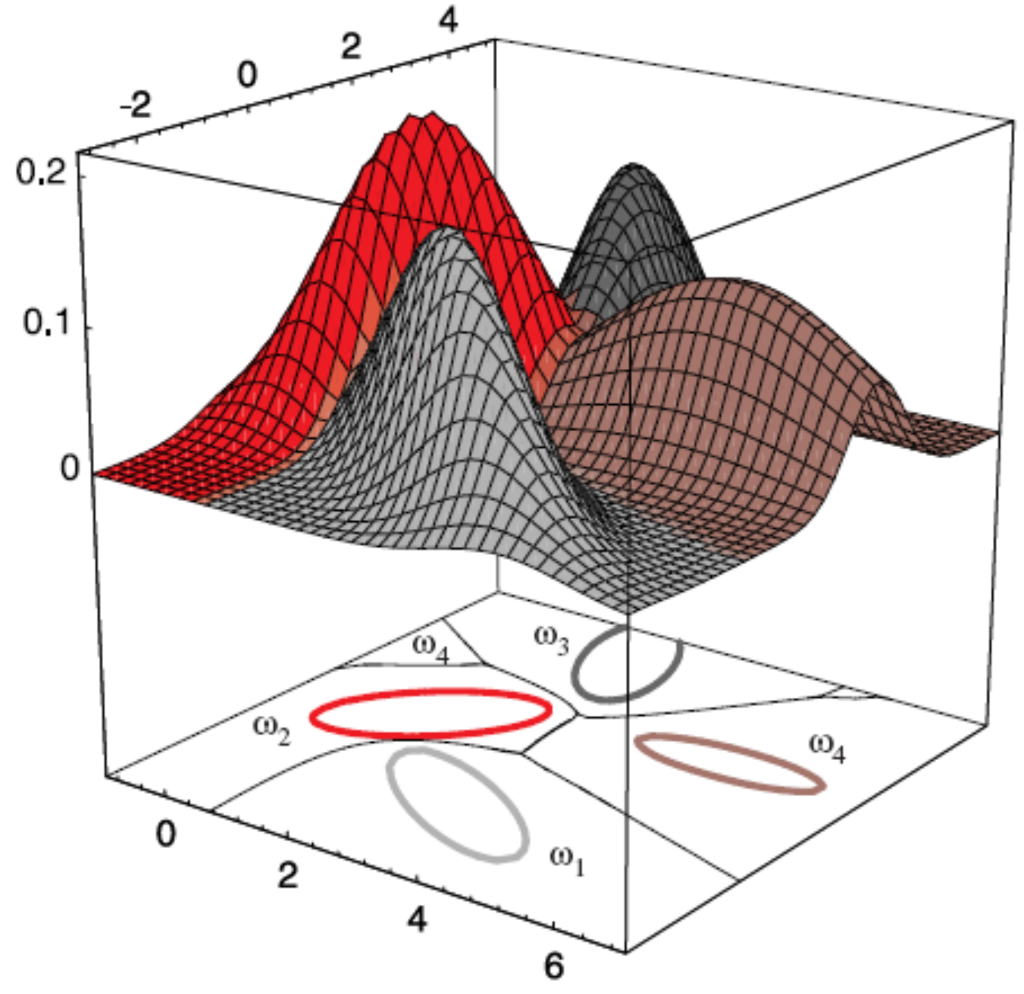
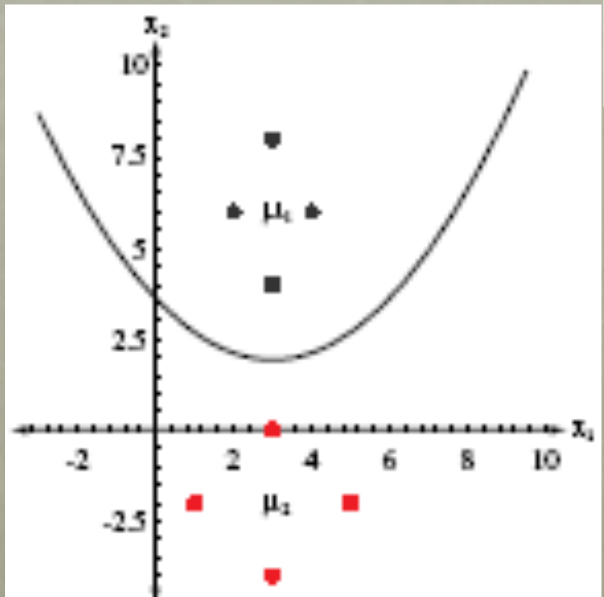


Figure 2.16: The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.

DECISION REGIONS FOR TWO-DIMENSIONAL GAUSSIAN DATA

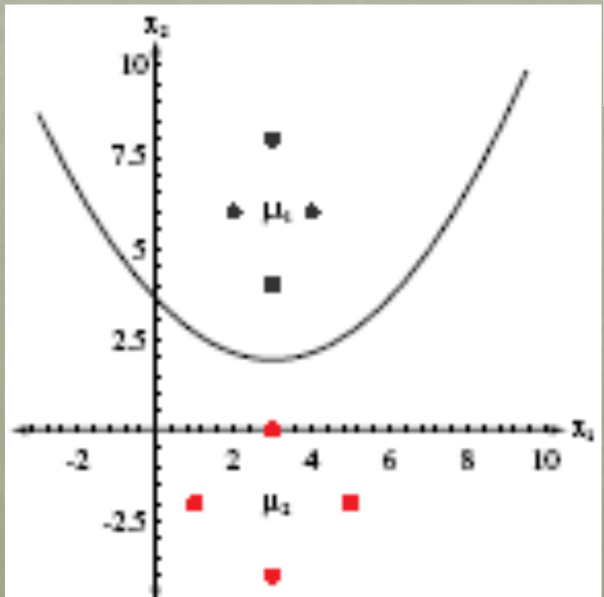


$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

Decision Regions for Two-Dimensional Gaussian Data

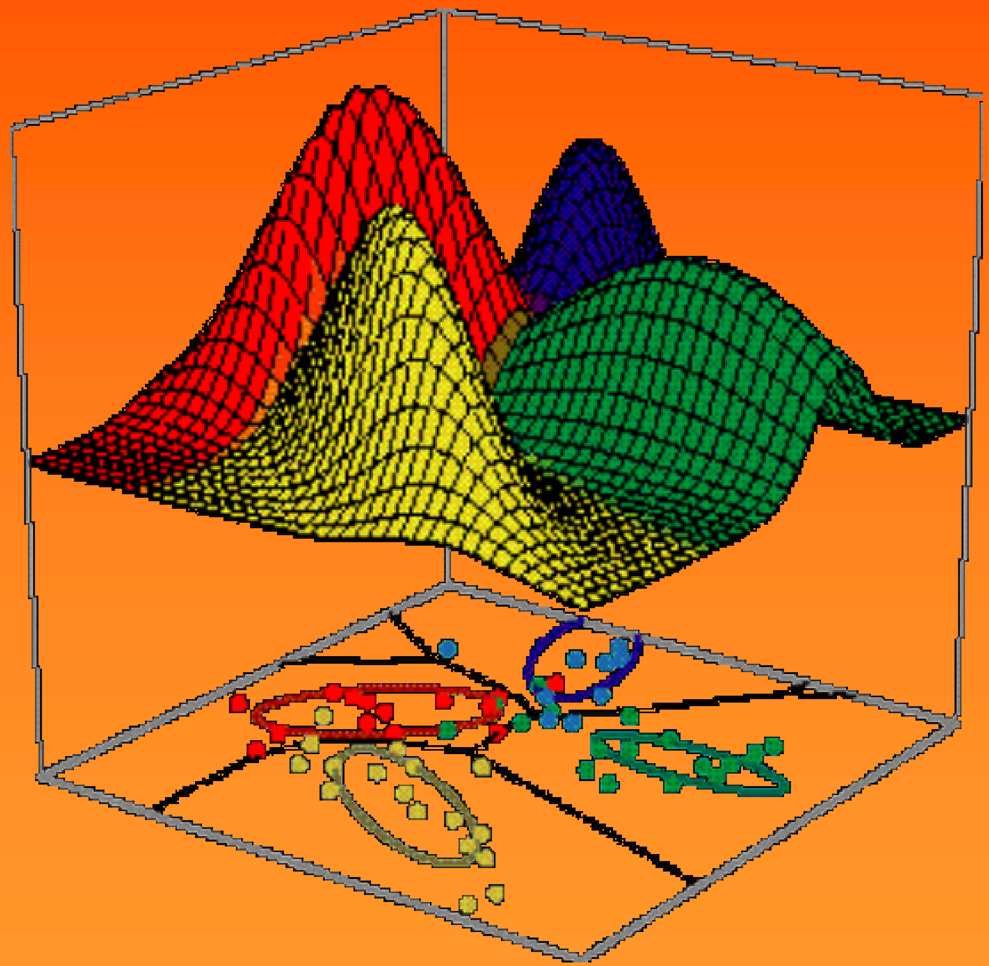


$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$



Pattern Classification

All materials in these slides were taken from
Pattern Classification (2nd ed) *by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*
with the permission of the authors and the publisher