

Chapter 2 :

Bayesian Decision Theory

(Sections 1-6)

1. Introduction – Bayesian Decision Theory
 - Pure statistics, probabilities known, optimal decision
2. Bayesian Decision Theory–Continuous Features
3. Minimum-Error-Rate Classification (omit)
4. Classifiers, Discriminant Functions, Decision Surfaces
5. The Normal Density
6. Discriminant Functions for the Normal Density

1. Introduction

- The sea bass/salmon example
 - State of nature, prior
 - State of nature is a random variable
 - The catch of salmon and sea bass is equiprobable
 - $P(\omega_1) = P(\omega_2)$ (uniform priors)
 - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

- Decision rule with only the prior information
 - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2
- Use of the class –conditional information
- $P(x | \omega_1)$ and $P(x | \omega_2)$ describe the difference in lightness between populations of sea and salmon

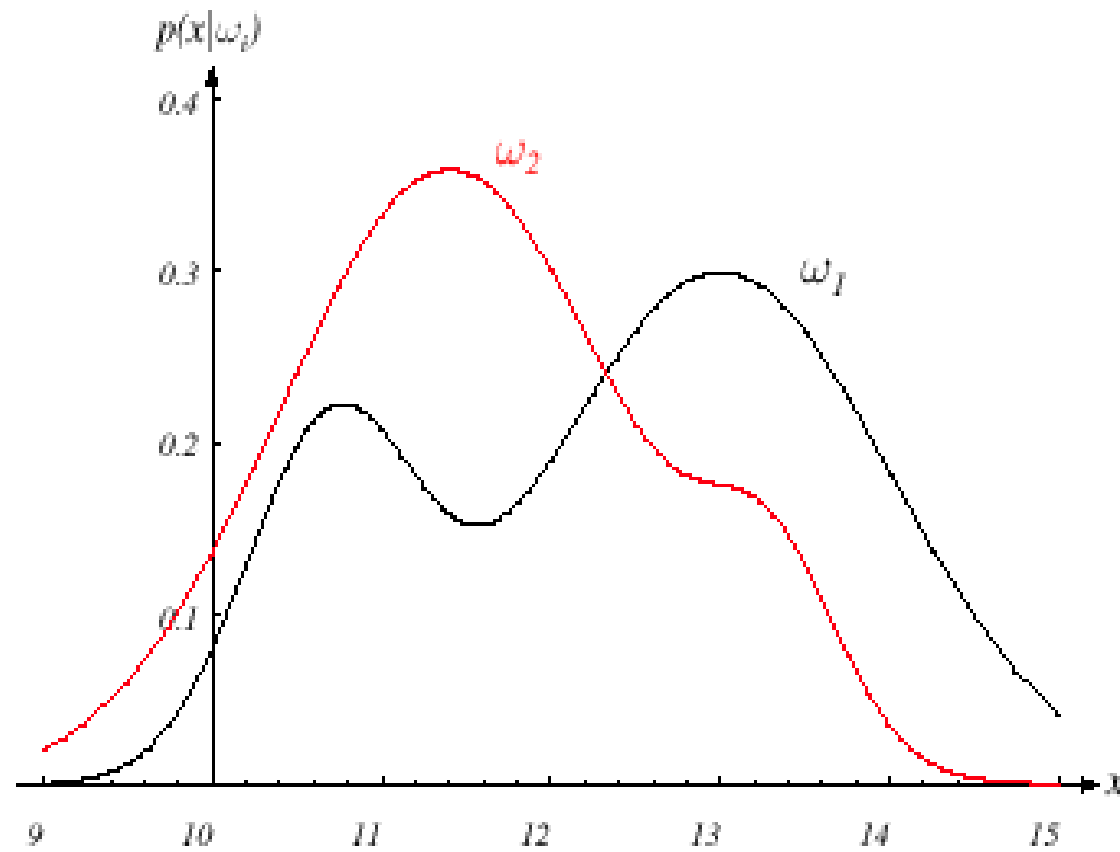


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i . If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Posterior, likelihood, evidence

- $P(\omega_j | x) = P(x | \omega_j) P(\omega_j) / P(x)$ (Bayes Rule)

- Where in case of two categories

$$P(x) = \sum_{j=1}^{j=2} P(x | \omega_j) P(\omega_j)$$

- Posterior = (Likelihood. Prior) / Evidence

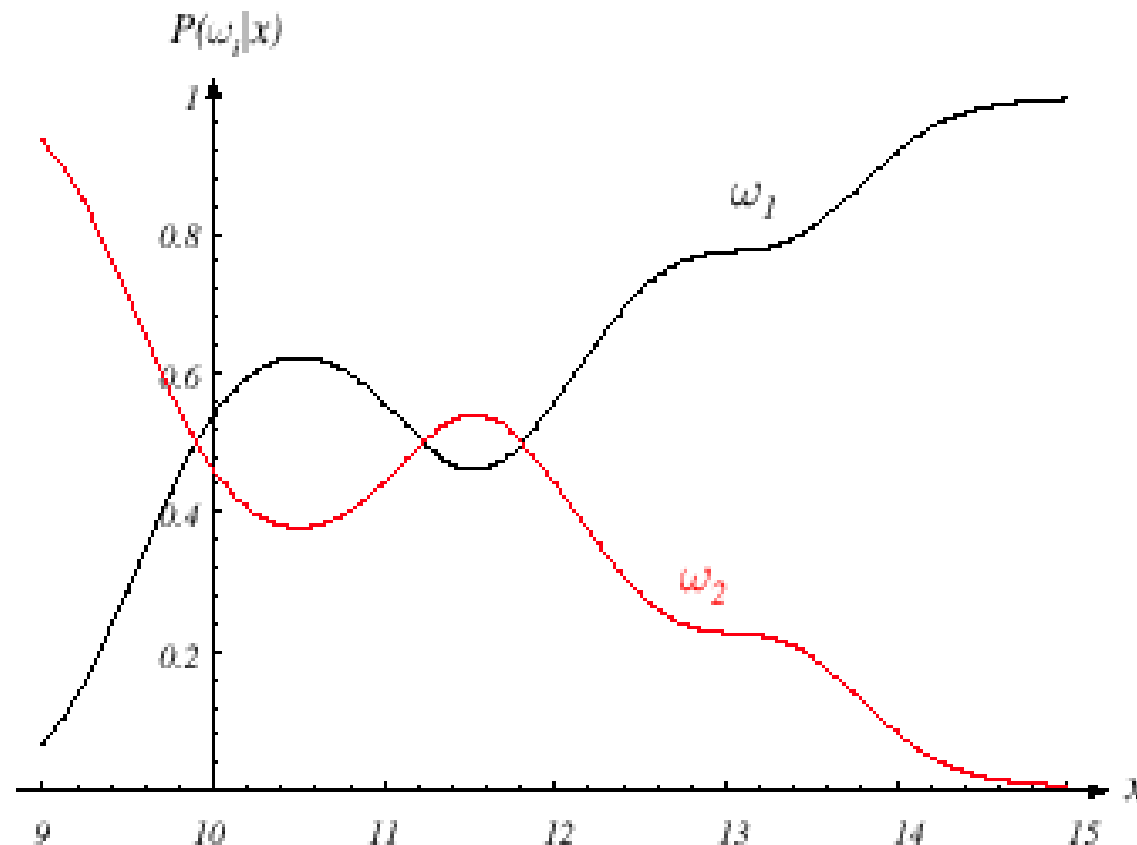




FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x , the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Decision given the posterior probabilities

X is an observation for which:

if $P(\omega_1 | x) > P(\omega_2 | x)$  True state of nature = ω_1
if $P(\omega_1 | x) < P(\omega_2 | x)$  True state of nature = ω_2

2. Bayesian Decision Theory – Continuous Features

- Generalization of the preceding ideas
 1. Use of more than one feature, **vector X**
 2. Use more than two states of nature, **c classes**
 3. Allowing actions other than deciding the state of nature (skip)
 4. Introduce a loss of function which is more general than the probability of error (skip)

4. Classifiers, Discriminant Functions, and Decision Surfaces

- The multi-category case
 - Set of discriminant functions $g_i(x)$, $i = 1, \dots, c$
 - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \quad \forall j \neq i$$

- Discriminant function

$$g_i(x) = P(\omega_i | x)$$

(max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x | \omega_i) P(\omega_i)$$

$$g_i(x) = \ln P(x | \omega_i) + \ln P(\omega_i)$$

(ln: natural logarithm!)

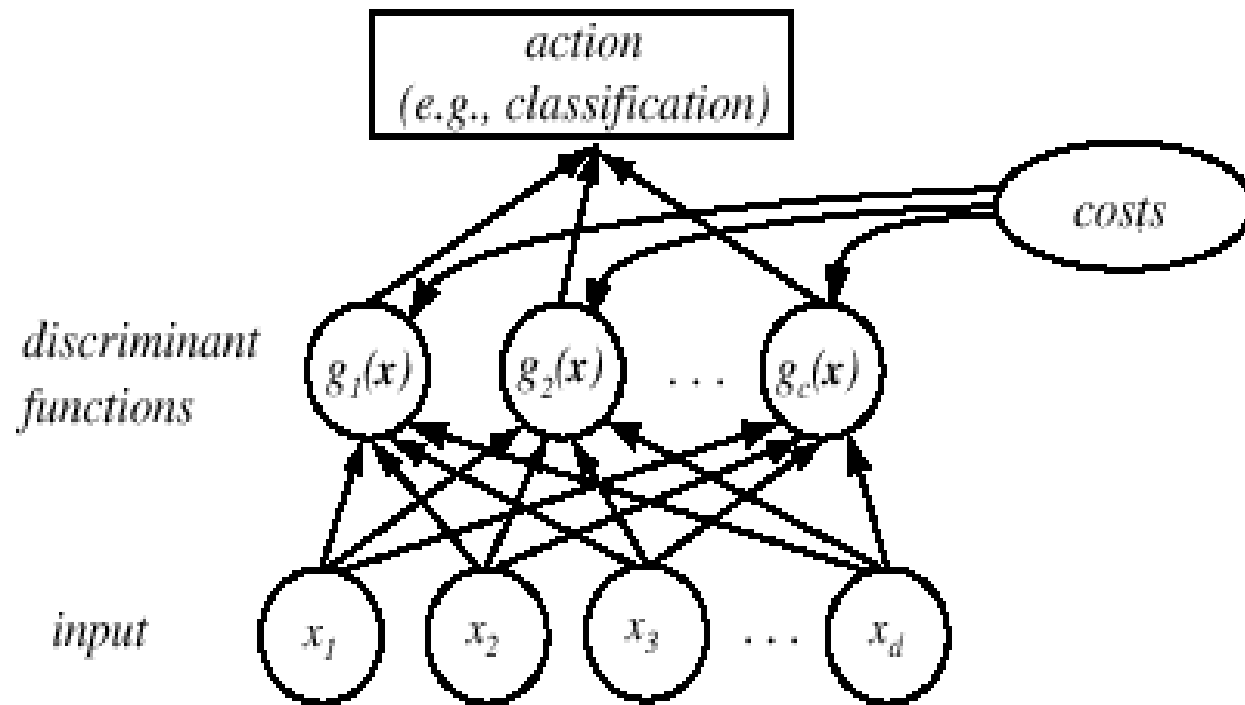


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Feature space divided into c decision regions

if $g_i(x) > g_j(x) \quad \forall j \neq i$ then x is in R_i

(R_i means assign x to ω_i)

- The two-category case
 - A classifier is a “dichotomizer” that has two discriminant functions g_1 and g_2

Let $g(x) \equiv g_1(x) - g_2(x)$

Decide ω_1 if $g(x) > 0$; Otherwise decide ω_2

- The computation of $g(x)$

$$\begin{aligned} g(x) &= P(\omega_1 | x) - P(\omega_2 | x) \\ &= \ln \frac{P(x | \omega_1)}{P(x | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \end{aligned}$$

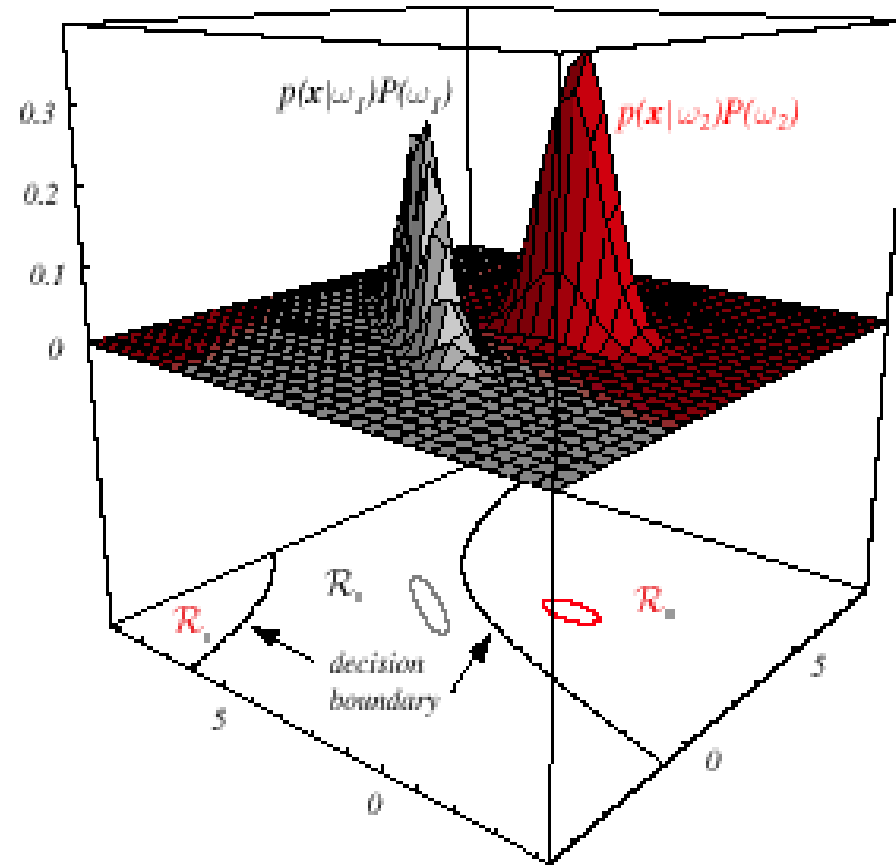


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region \mathcal{R}_2 is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

5. The Normal Density

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

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

Where:

μ = mean (or expected value) of x

σ^2 = expected squared standard deviation or variance

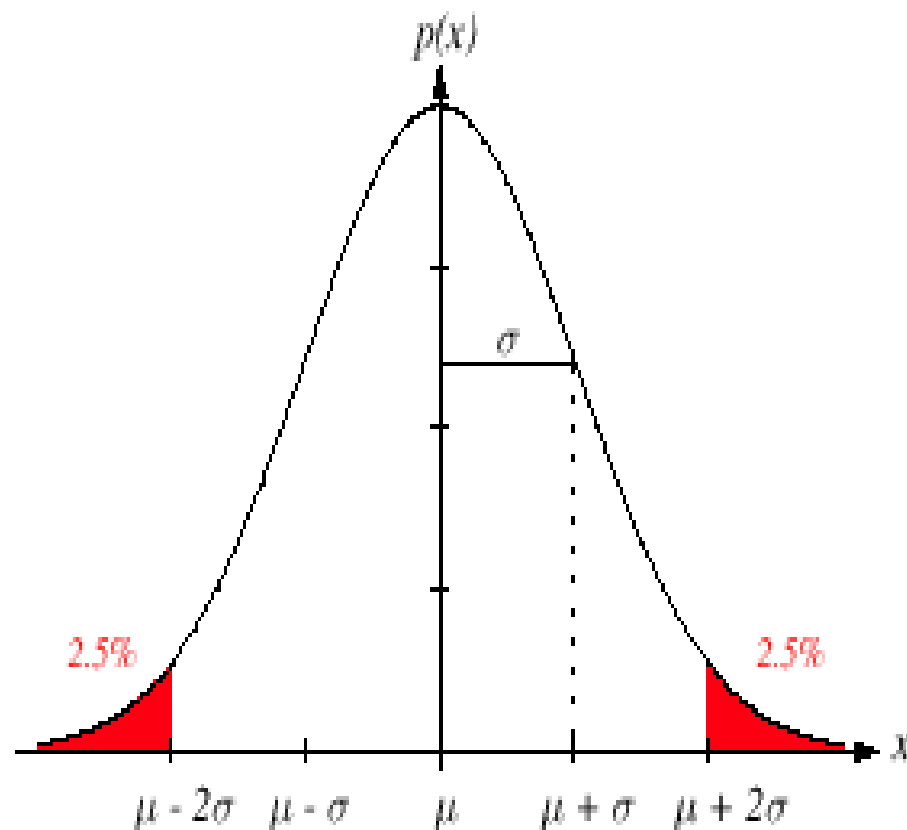


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

- Multivariate normal density $p(x) \sim N(\mu, \Sigma)$
 - Multivariate normal density in d dimensions is:

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

where:

$x = (x_1, x_2, \dots, x_d)^t$ (t stands for the transpose vector form)

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

$\Sigma = d \times d$ covariance matrix

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

Properties of the Normal Density Covariance Matrix Σ

- Always symmetric
- Always positive semi-definite, but for our purposes Σ is positive definite
 - determinant is positive
 - $\forall \Sigma$ invertible
- Eigenvalues real and positive, and the principal axes of the hyperellipsoidal loci of points of constant density are eigenvectors of Σ

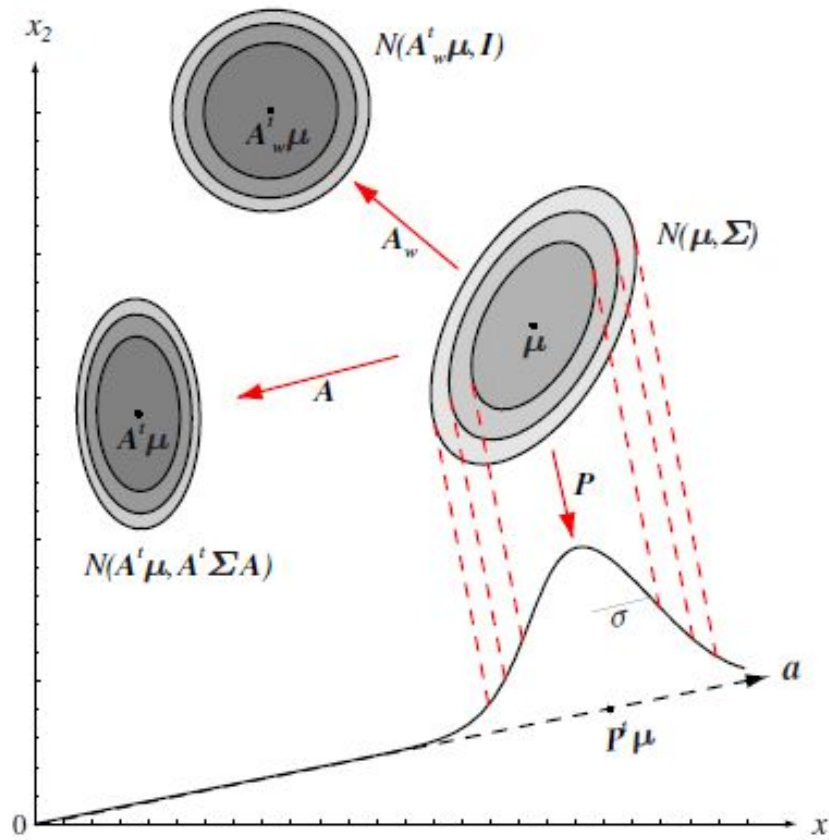


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, A , takes the source distribution into distribution $N(A^T \mu, A^T \Sigma A)$. Another linear transformation—a projection P onto a line defined by vector a —leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original $x_1 x_2$ -space. A whitening transform, A_w , leads to a circularly symmetric Gaussian, here shown displaced. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

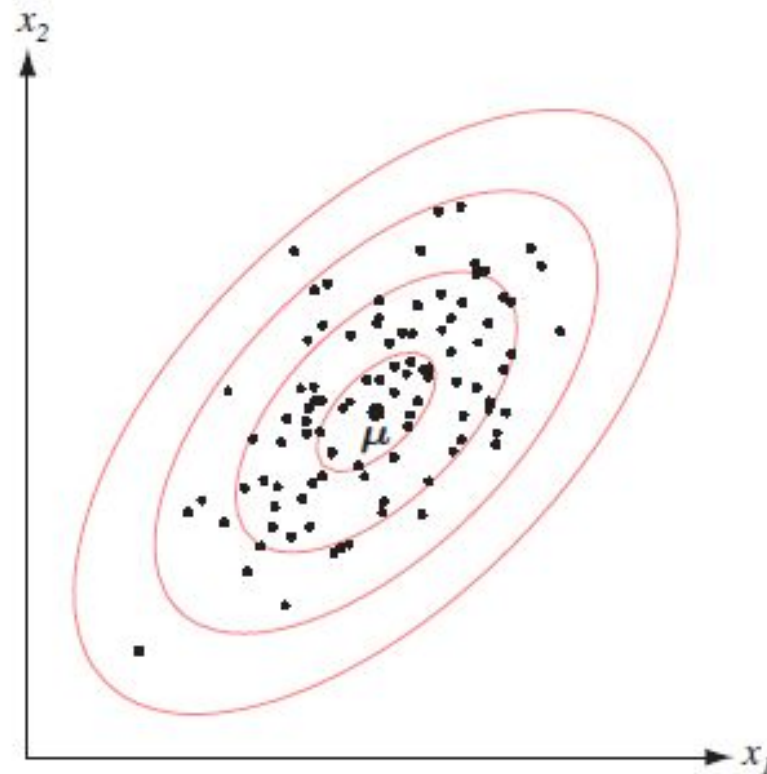


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ . The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

6. Discriminant Functions for the Normal Density

- We saw that the minimum error-rate classification can be achieved by the discriminant function

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

- Case of multivariate normal $p(x) \sim N(\mu, \Sigma)$

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

1. Case $\Sigma_i = \sigma^2 I$ (I stands for the identity matrix)

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

where :

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

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

- A classifier that uses linear discriminant functions is called “a linear machine”
- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

$$g_i(x) = g_j(x)$$

- If equal priors for all classes, this reduces to the minimum-distance classifier where an unknown is assigned to the class of the nearest mean

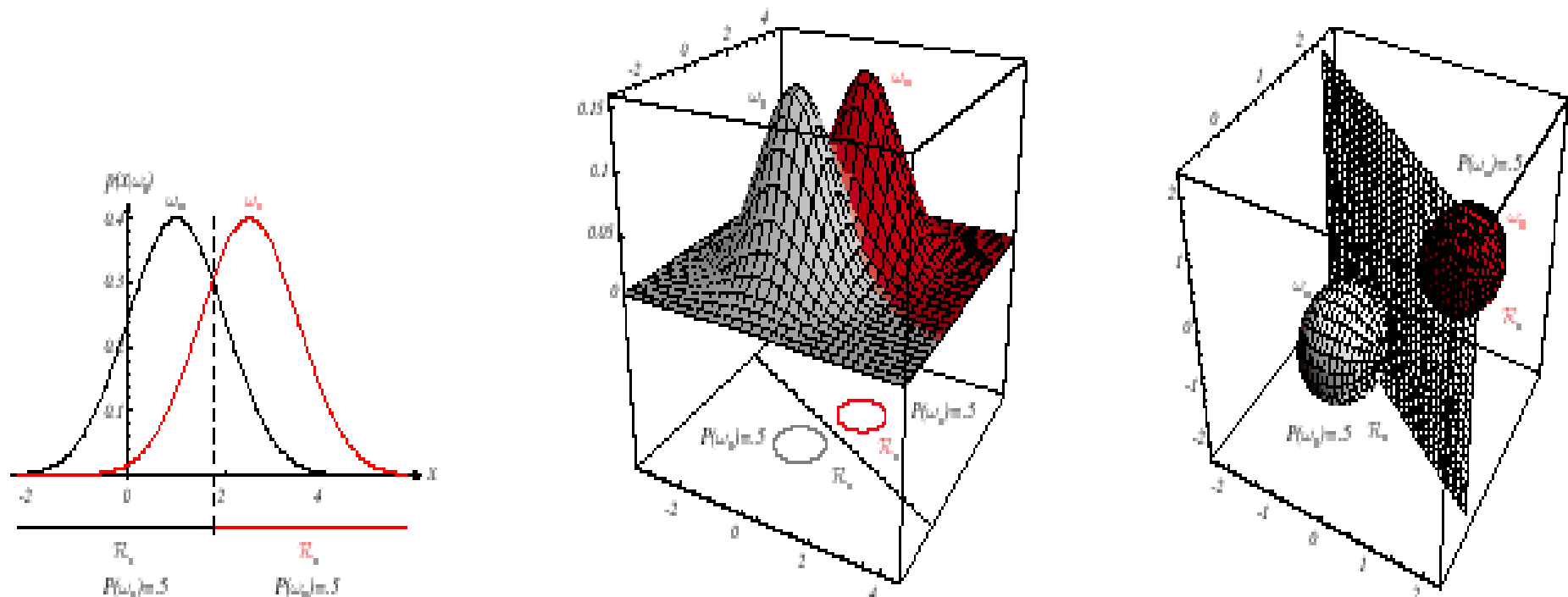


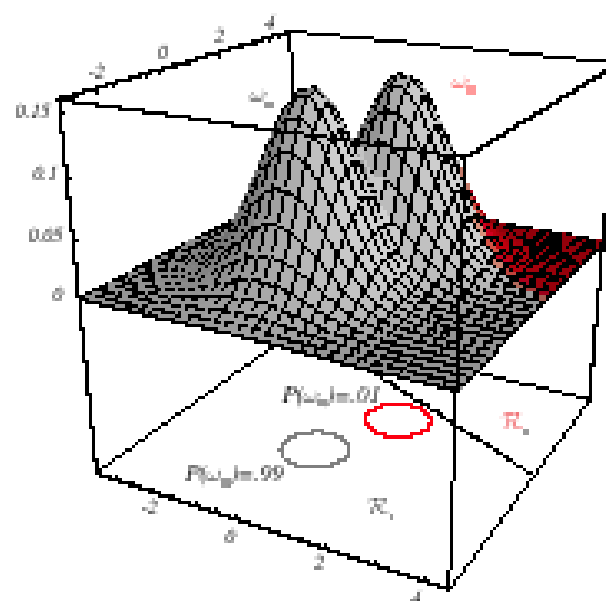
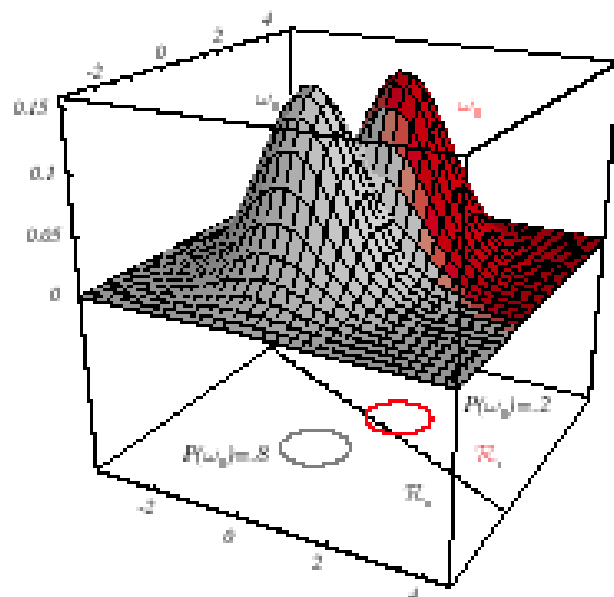
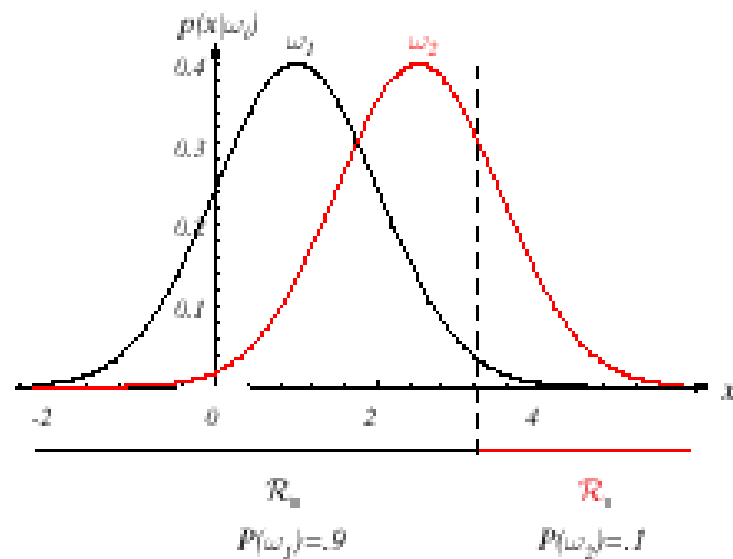
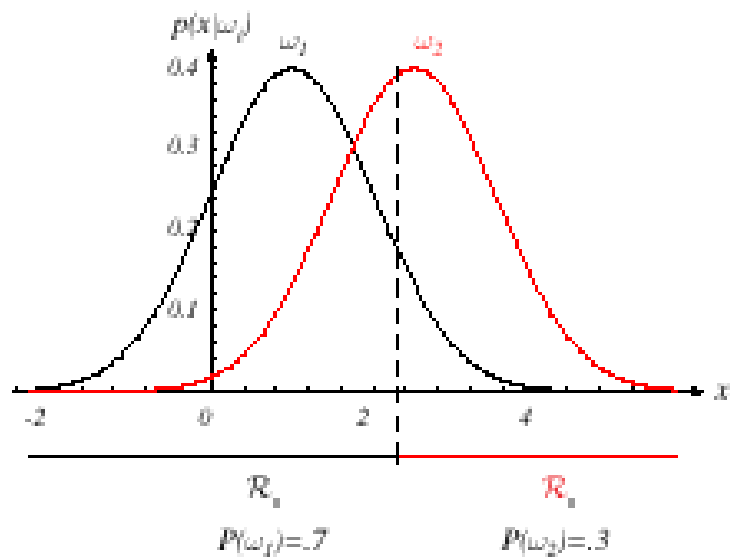
FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d - 1$ dimensions, perpendicular to the line separating the means. In these one-, two-, 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.

- The hyperplane separating R_i and R_j

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

always orthogonal to the line linking the means!

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



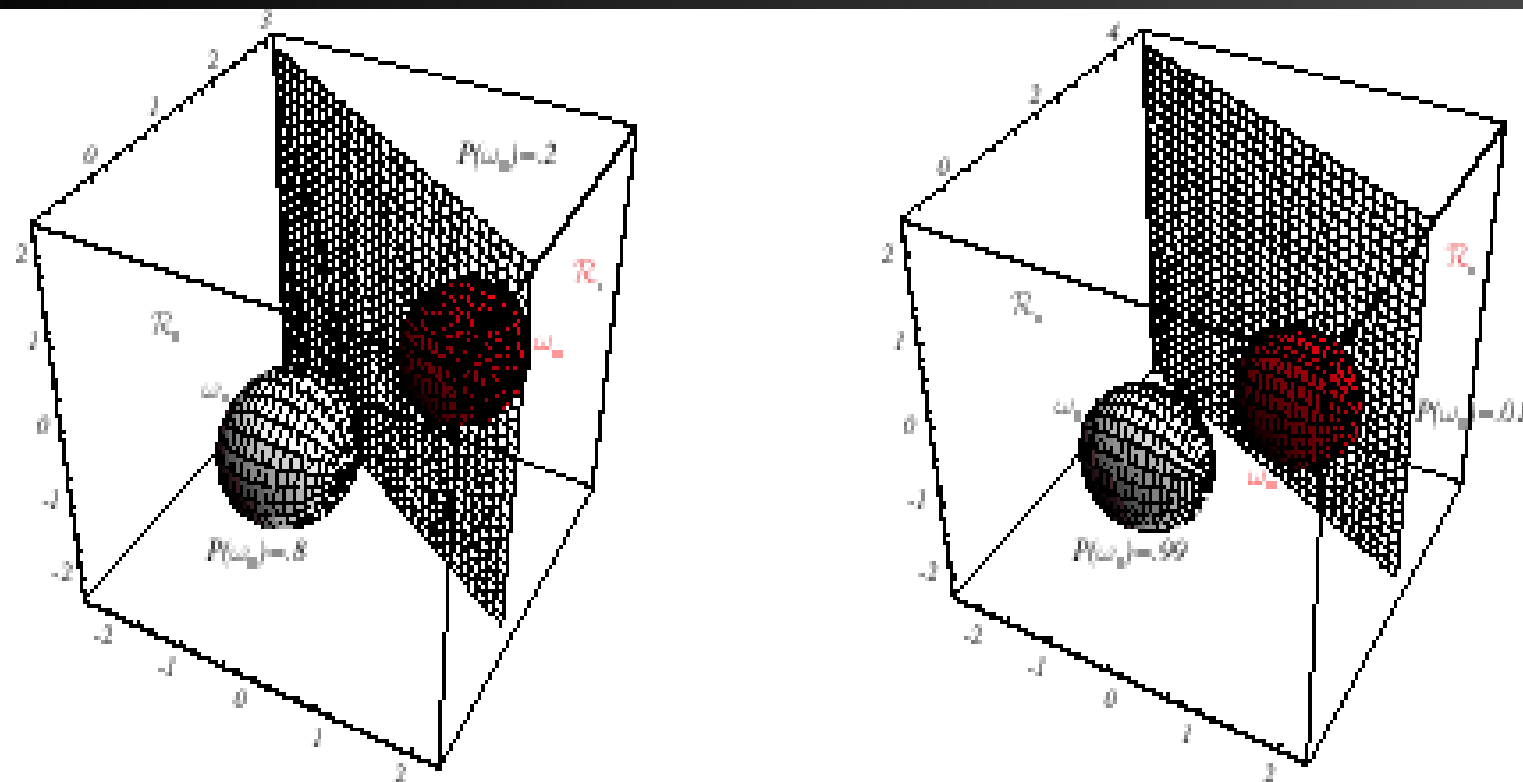


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

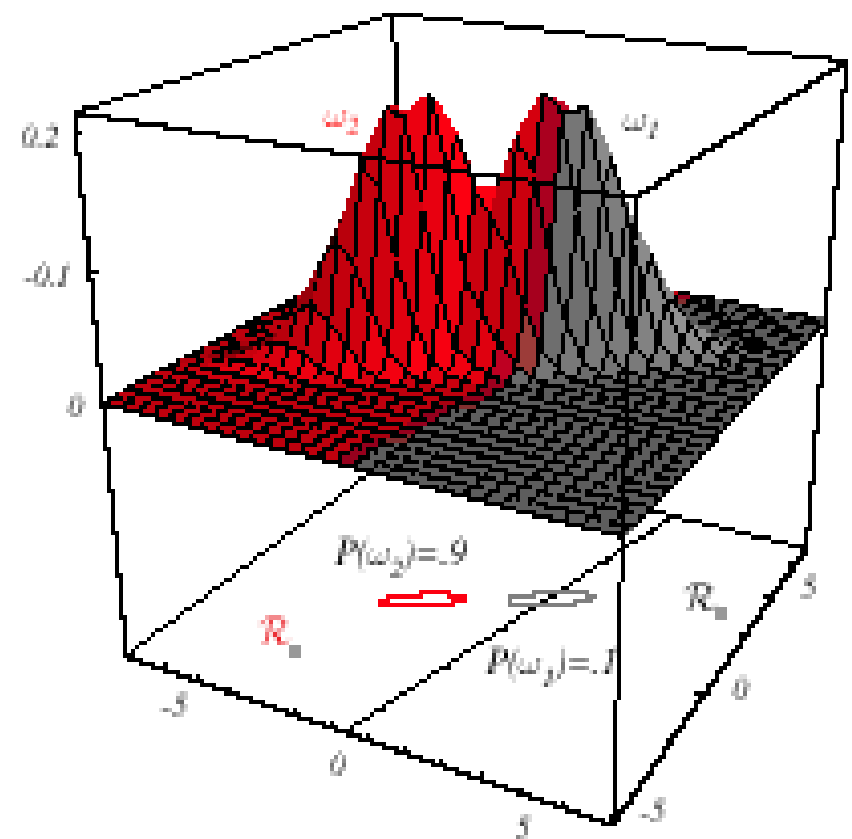
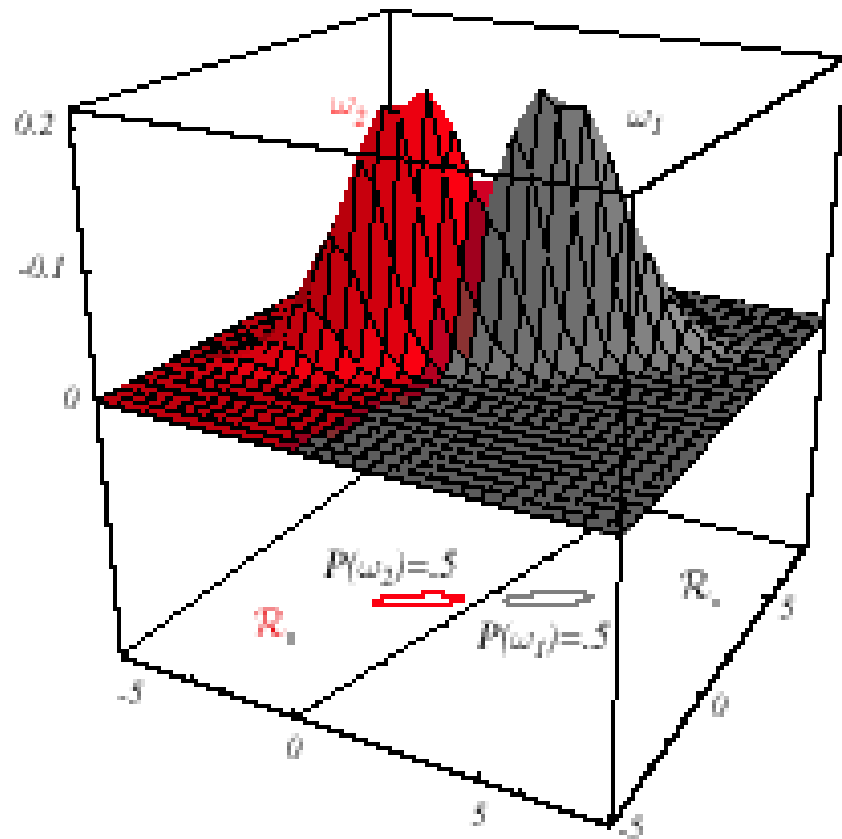
2. Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

- Hyperplane separating R_i and R_j

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

(the hyperplane separating R_i and R_j is generally not orthogonal to the line between the means!)

If priors equal, reduces to squared Mahalanobis distance



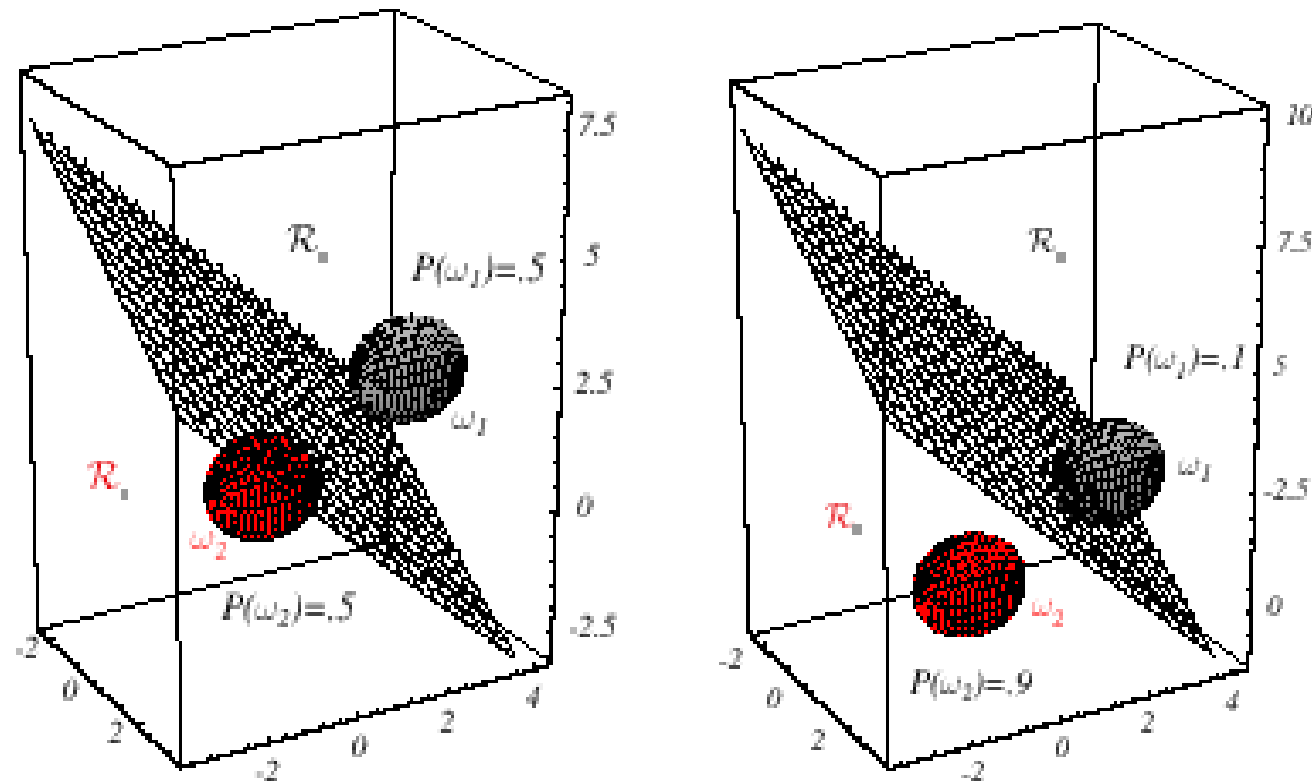


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

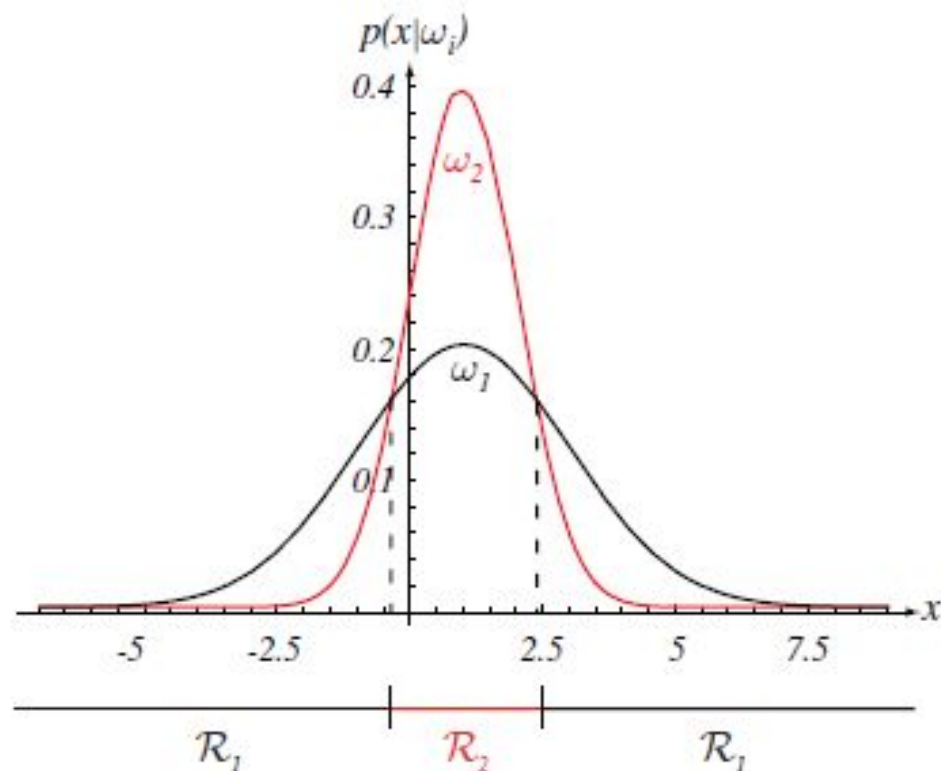


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

3. Case $\Sigma_i = \text{arbitrary}$

- The covariance matrices are different for each category

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

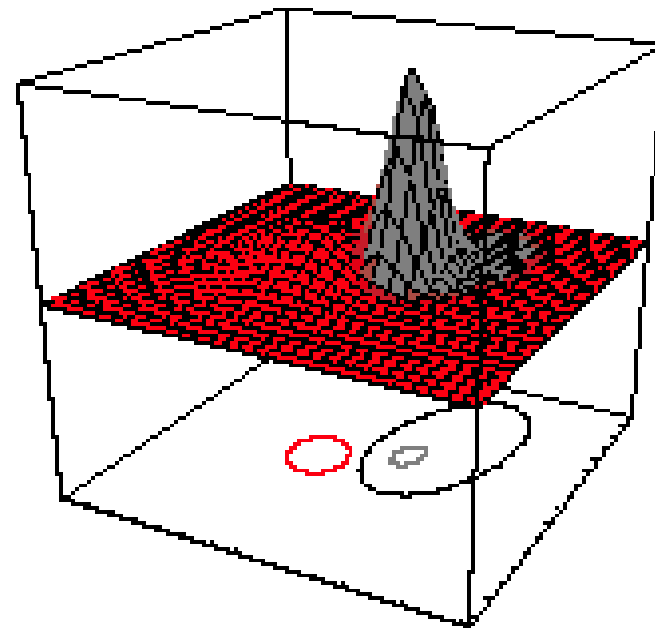
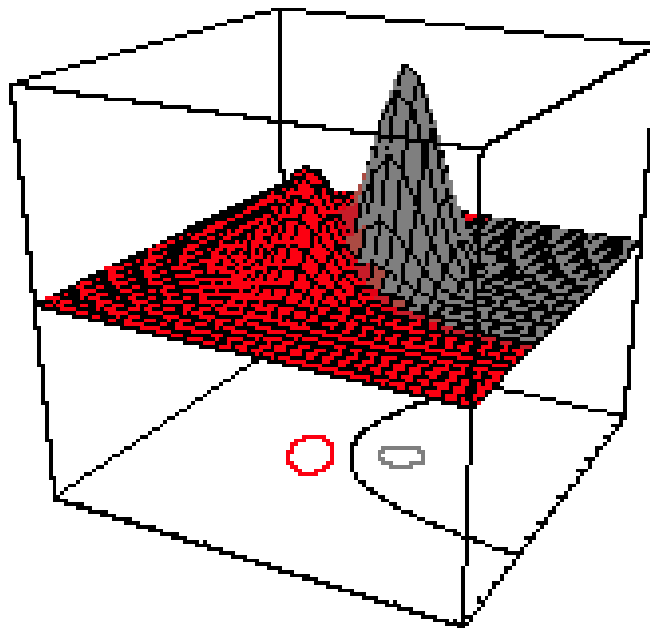
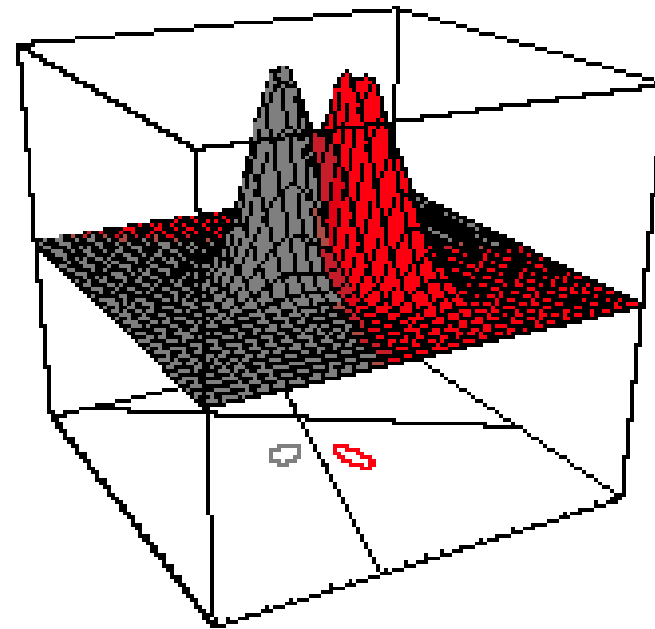
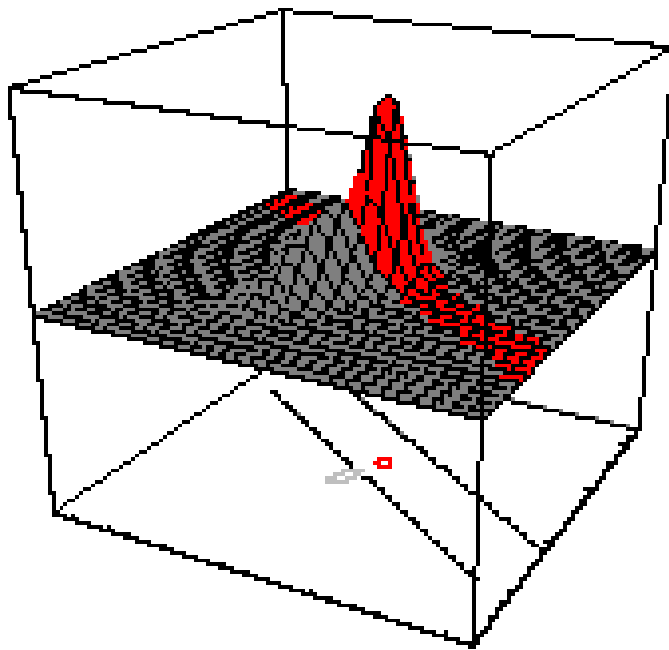
where :

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

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

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

(**Hyperquadrics** which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)



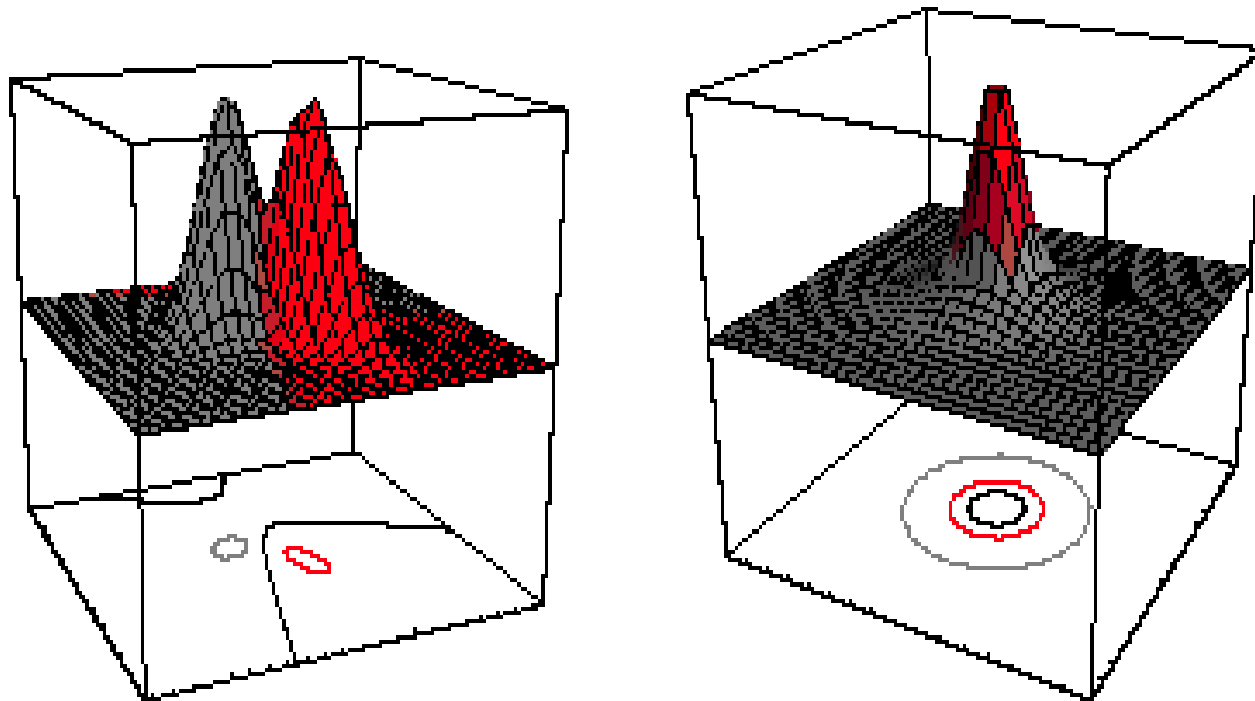
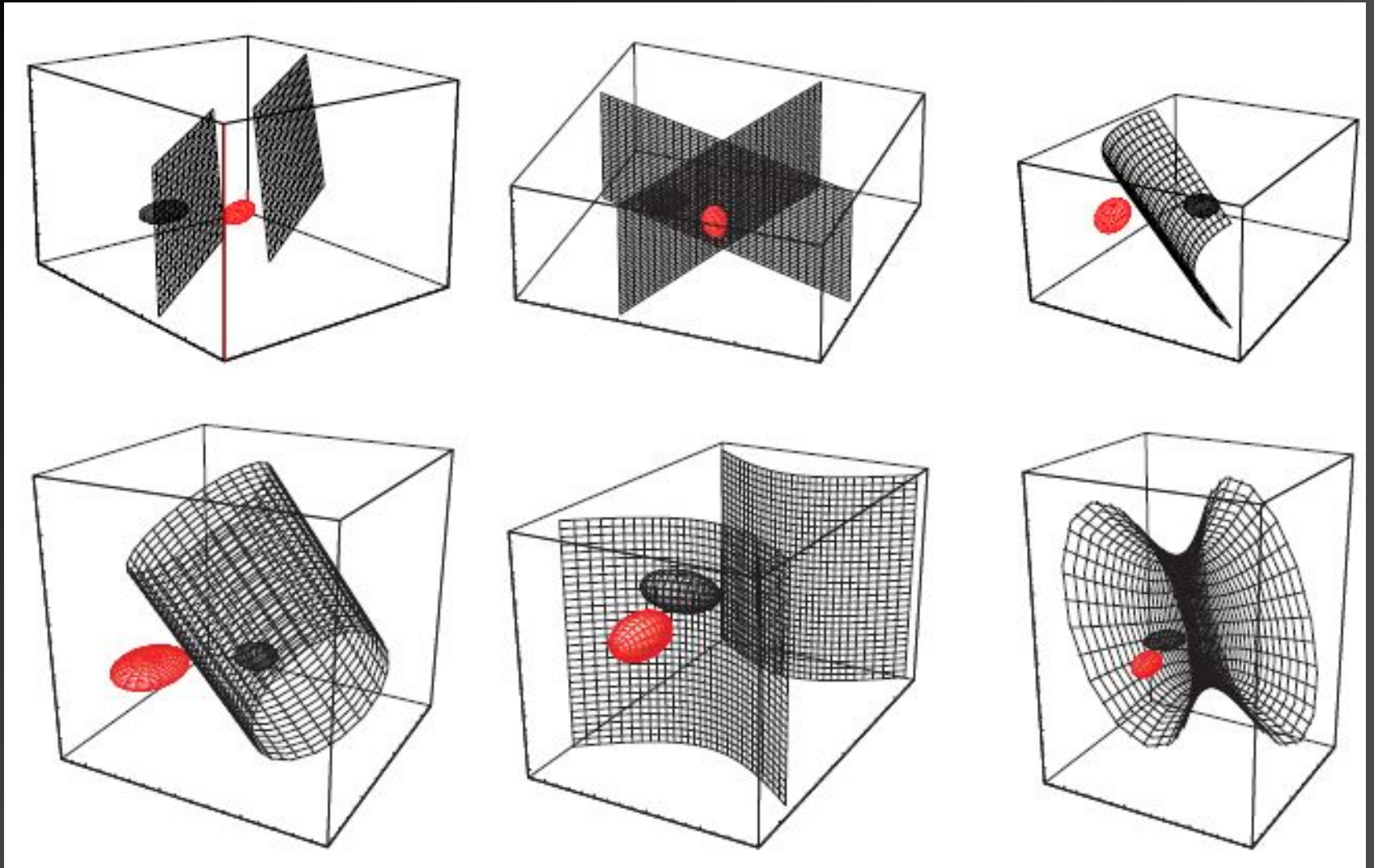


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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



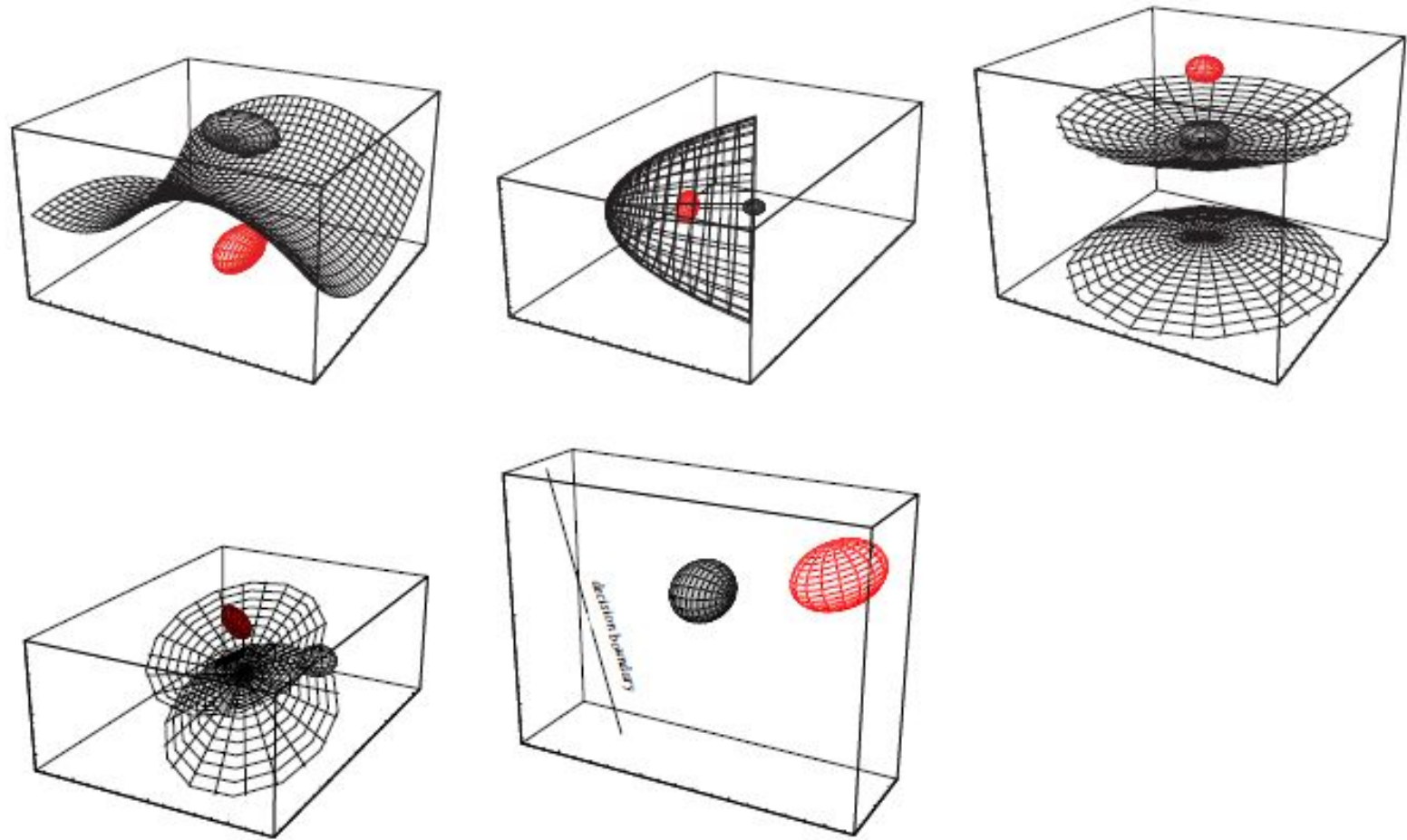


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

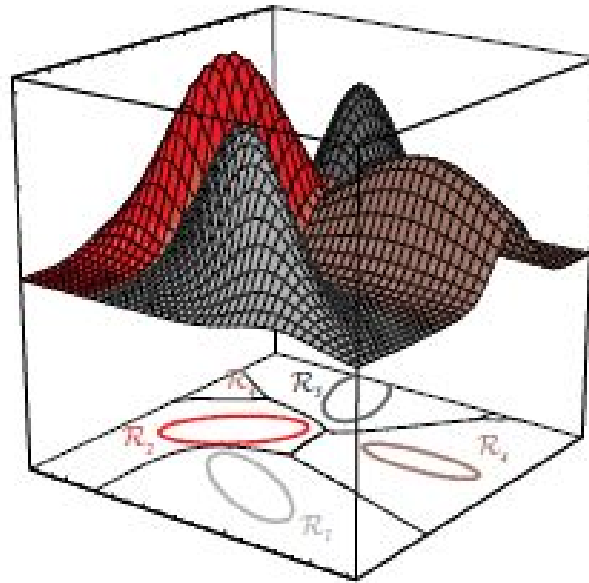
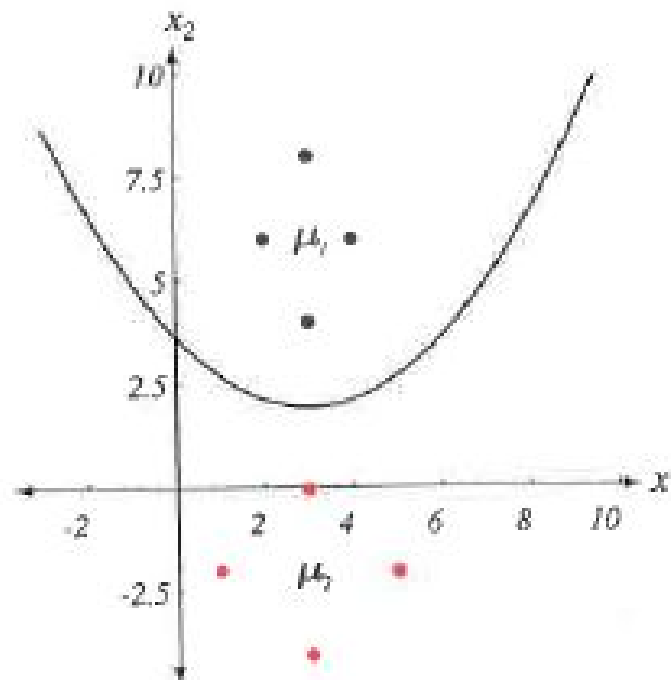


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

EXAMPLE 1 Decision Regions for Two-Dimensional Gaussian Data

To clarify these ideas, we explicitly calculate the decision boundary for the two-category two-dimensional data in the Example figure.



The computed Bayes decision boundary for two Gaussian distributions, each based on four data points.