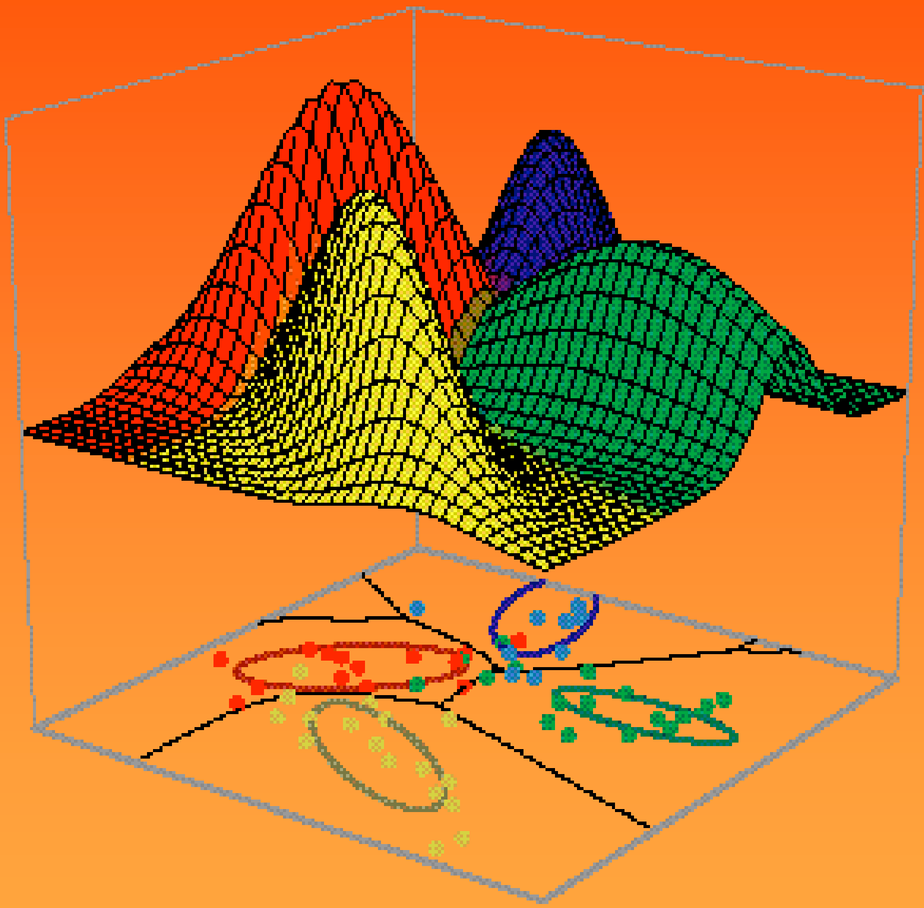
Pattern Classification

David G.Stork
Rambus



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

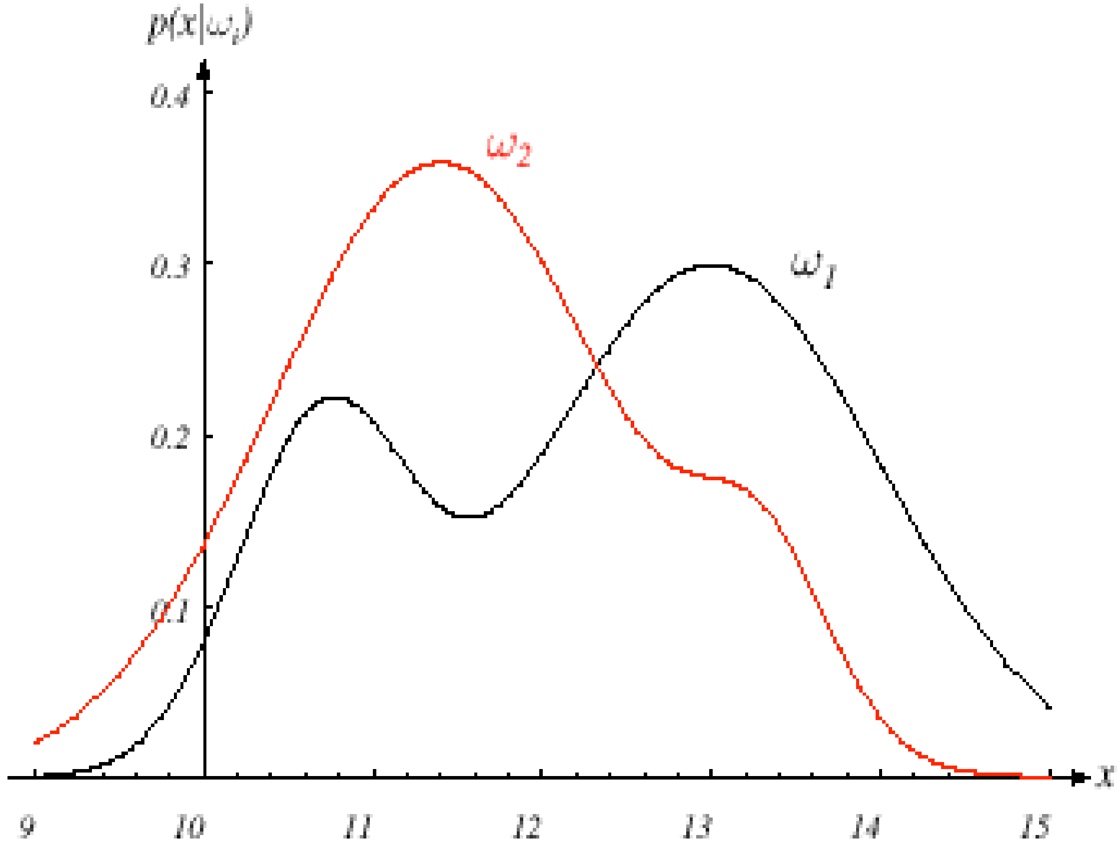


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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
  - $P(\omega_j \mid x) = P(x \mid \omega_j) \, . \, P(\omega_j) \, / \, P(x)$
  - Where in case of two categories

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

  - Posterior = (Likelihood. Prior) / Evidence

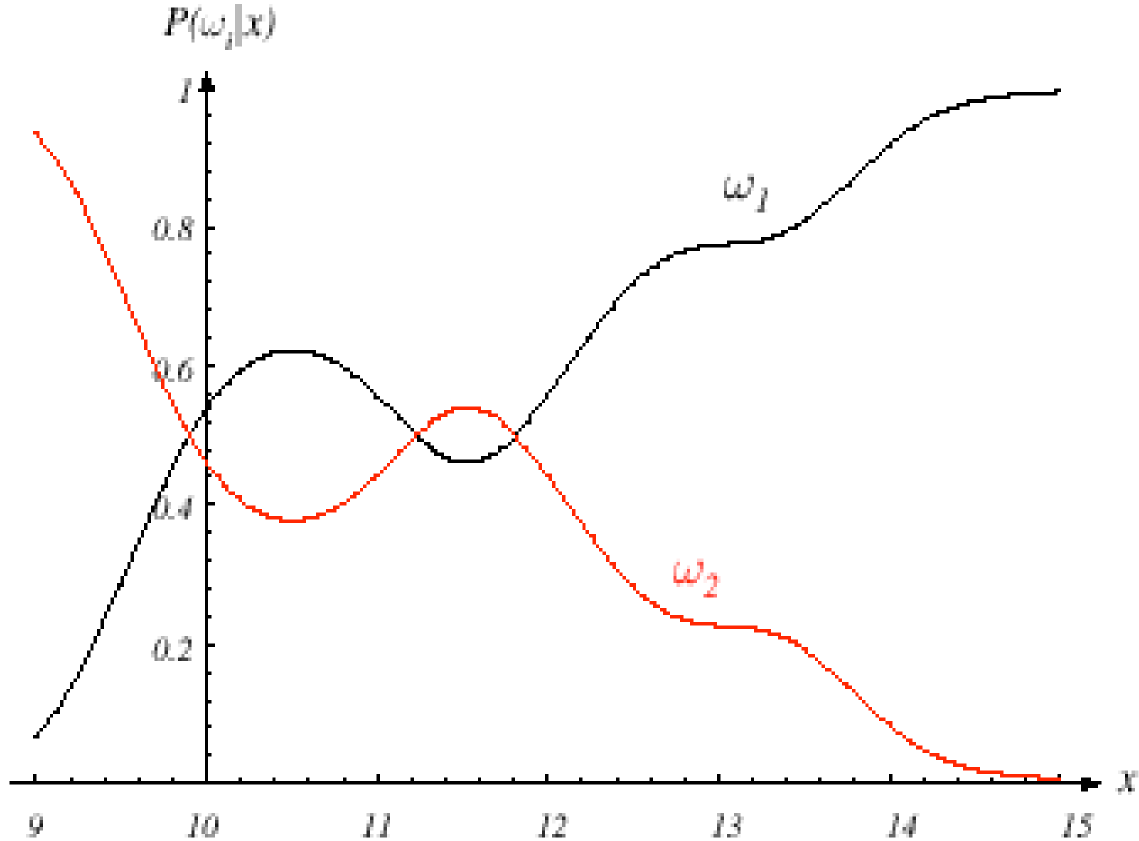


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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*.

- Decision given the posterior probabilities

X is an observation for which:

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

Therefore:

whenever we observe a particular x, the probability of error is :

$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$
$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

- Minimizing the probability of error

- Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

  Therefore:

$$P(error \mid x) = \min\,[P(\omega_1 \mid x),\ P(\omega_2 \mid x)]$$

(Bayes decision)

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions

Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking

action α_i when the state of nature is ω_j

Overall risk

R = *Sum of all* $\underbrace{R(\alpha_i \mid x)}_{\text{Conditional risk}}$ *for* $i = 1,\ldots,a$

Minimizing R $\Longleftrightarrow$ Minimizing $R(\alpha_i \mid x)$ *for* $i = 1,\ldots, a$

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

for i = 1,…,a

- Two-category classification

α_1 : deciding ω_1

α_2 : deciding ω_2

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

decide ω_1 if:

$$(\lambda_{21} - \lambda_{11})\, P(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22})\, P(x \mid \omega_2)\, P(\omega_2)$$

and decide ω_2 otherwise

Likelihood ratio:

The preceding rule is equivalent to the following rule:

$$if \; \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

Optimal decision property

“If the likelihood ratio exceeds a threshold value independent of the input pattern x, we can take optimal actions”

Minimum-Error-Rate Classification

- Actions are decisions on classes

  If action α_i is taken and the true state of nature is ω_j then:

  the decision is correct if $i = j$ and in error if $i \neq j$

- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

- Introduction of the zero-one loss function:

$$\lambda(\alpha_i,\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i,j = 1,\ldots,c$$

Therefore, the conditional risk is:

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

$$= \sum_{j \neq 1} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

"The risk corresponding to this loss function is the average probability error"

- Minimize the risk requires maximize $P(\omega_i \mid x)$ (since $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$)

- For Minimum error rate
  - Decide ω_i if $P(\omega_i \mid x) > P(\omega_j \mid x)\ \forall j \neq i$

- Regions of decision and zero-one loss function, therefore:

$$Let\ \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}.\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda\ then\ decide\ \omega_1\ \ if:\ \frac{P(x\,|\,\omega_1)}{P(x\,|\,\omega_2)}>\theta_\lambda$$

- If λ is the zero-one loss function which means:

$$\lambda=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$

$$then\ \ \theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}=\theta_a$$

$$if\ \ \lambda=\begin{pmatrix}0 & 2\\ 1 & 0\end{pmatrix}\ then\ \ \theta_\lambda=\frac{2P(\omega_2)}{P(\omega_1)}=\theta_b$$

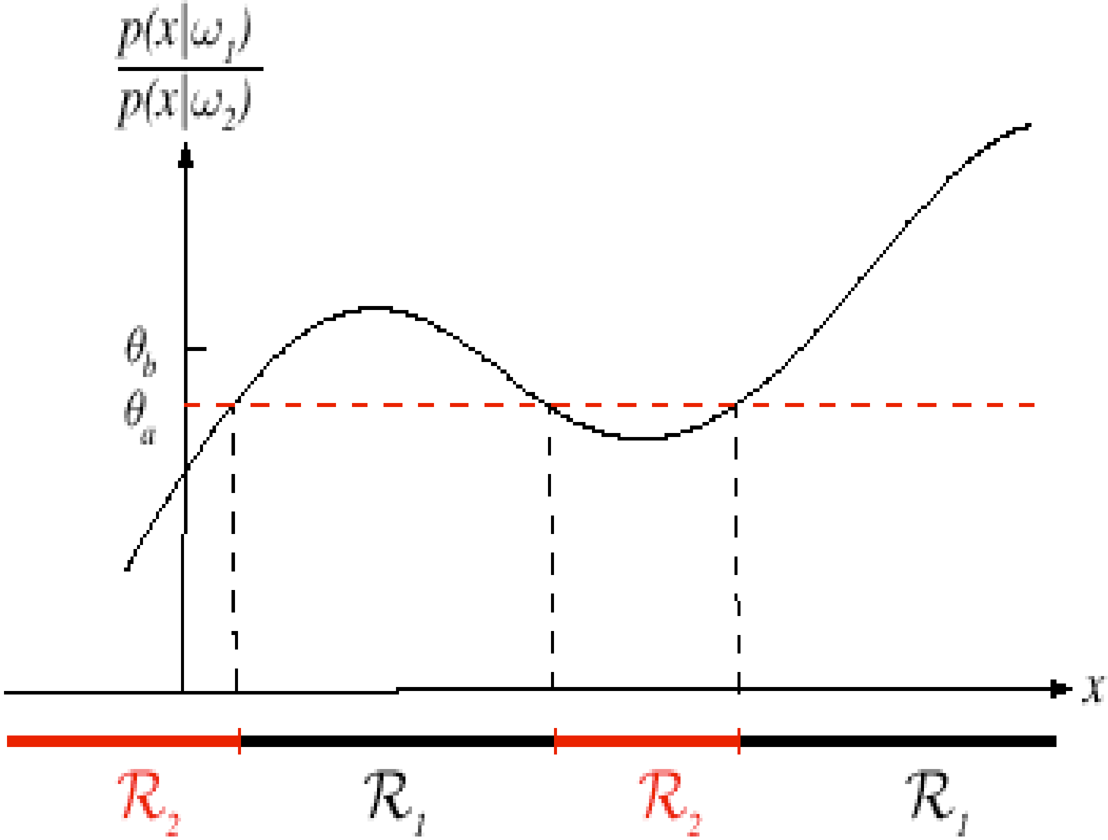


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classifiers, Discriminant Functions and Decision Surfaces

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

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

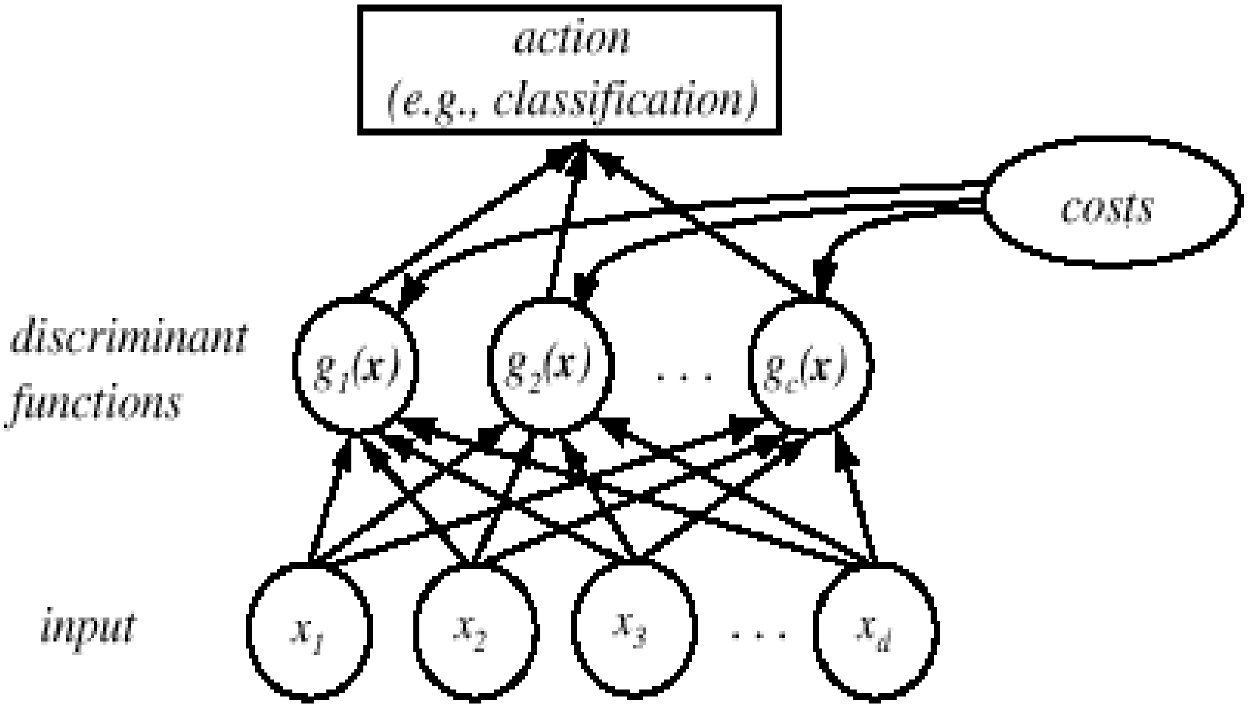


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction 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.

- *Let* $g_i(x) = - R(\alpha_i \mid x)$

  (max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take

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

  (max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$$

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

(ln: natural logarithm!)

- Feature space divided into c decision regions

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

(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)

$$g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$
$$= \ln \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

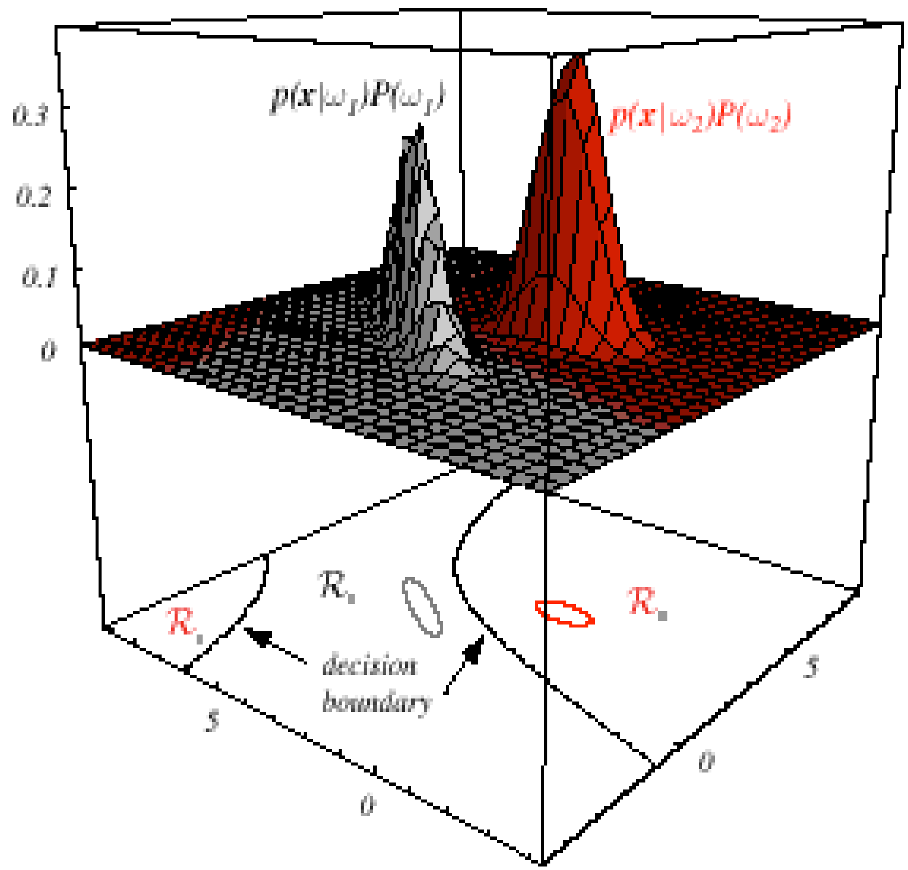


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

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

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

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

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

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 ith 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:

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

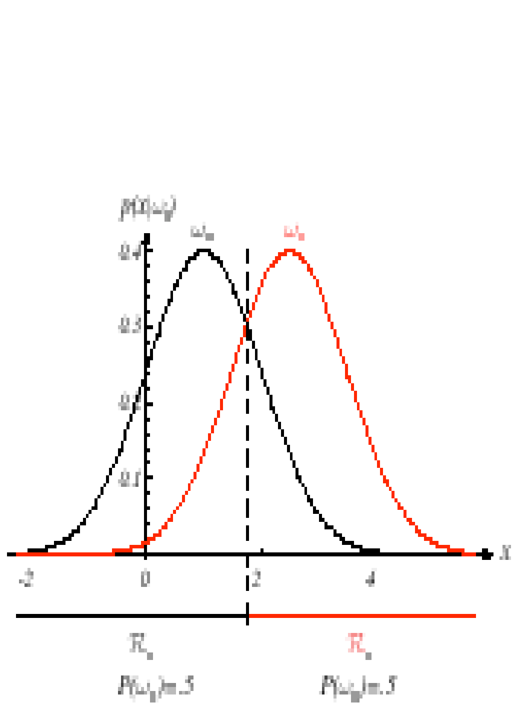


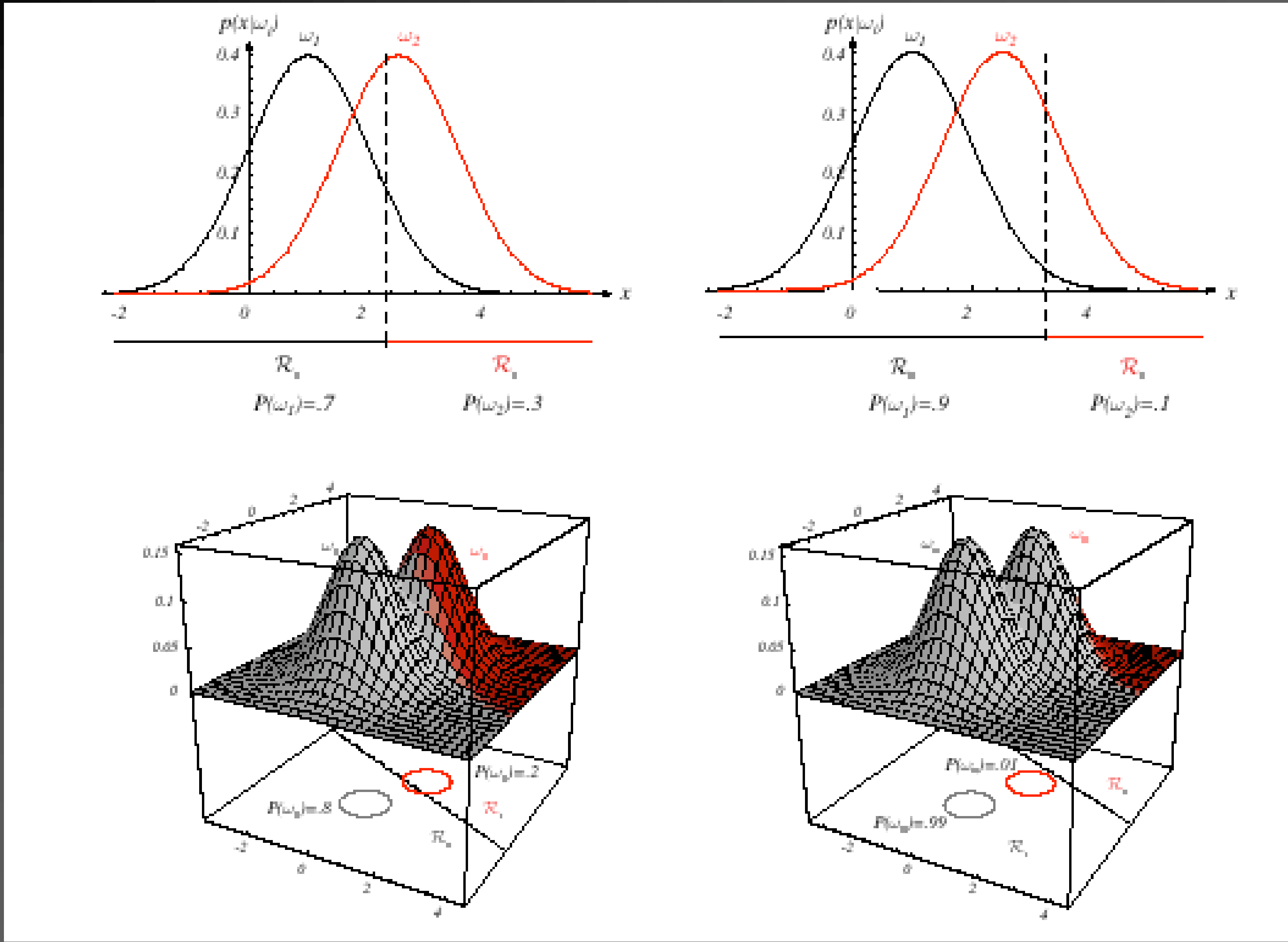
FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

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



$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
0
2
4
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.7$
$P(\omega_2)=.3$
$P(\omega_1)=.9$
$P(\omega_2)=.1$
0.15
0.1
0.05
0

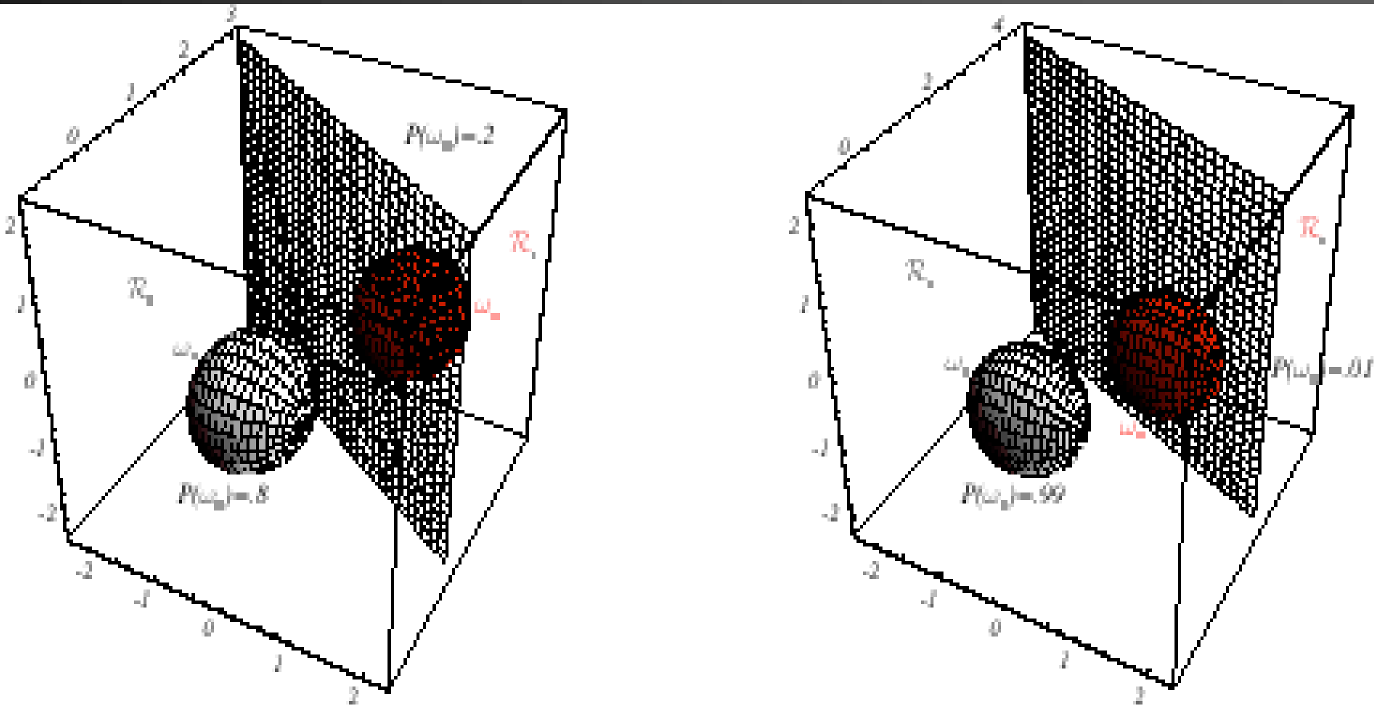


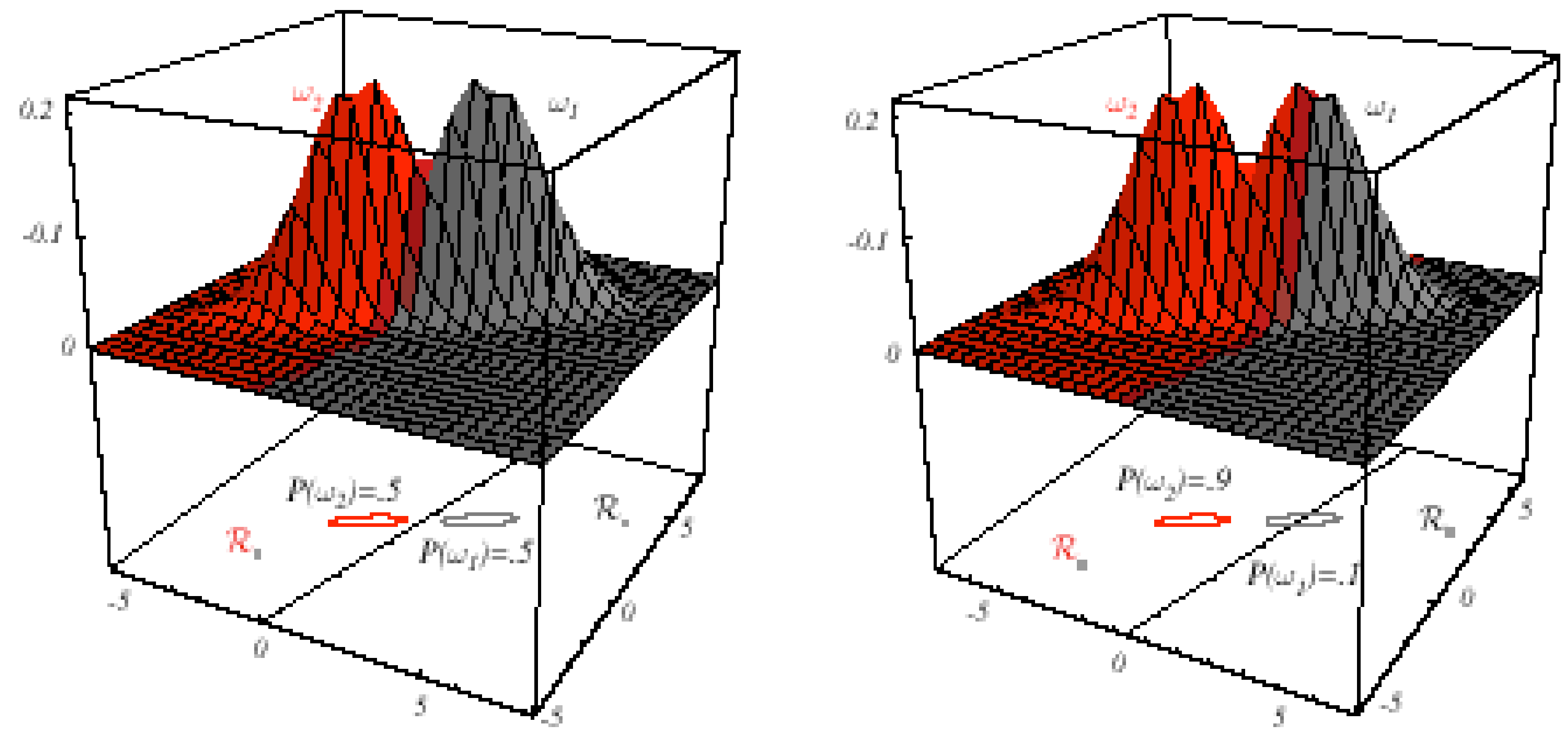
FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

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

  - Hyperplane separating R_i and R_j

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

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



0.2
0.1
0
ω_2
ω_1
$P(\omega_2)=.5$
$P(\omega_1)=.5$
-5
0
5
0.2
0.1
0
ω_2
ω_1
$P(\omega_2)=.9$
$P(\omega_1)=.1$
-5
0
5

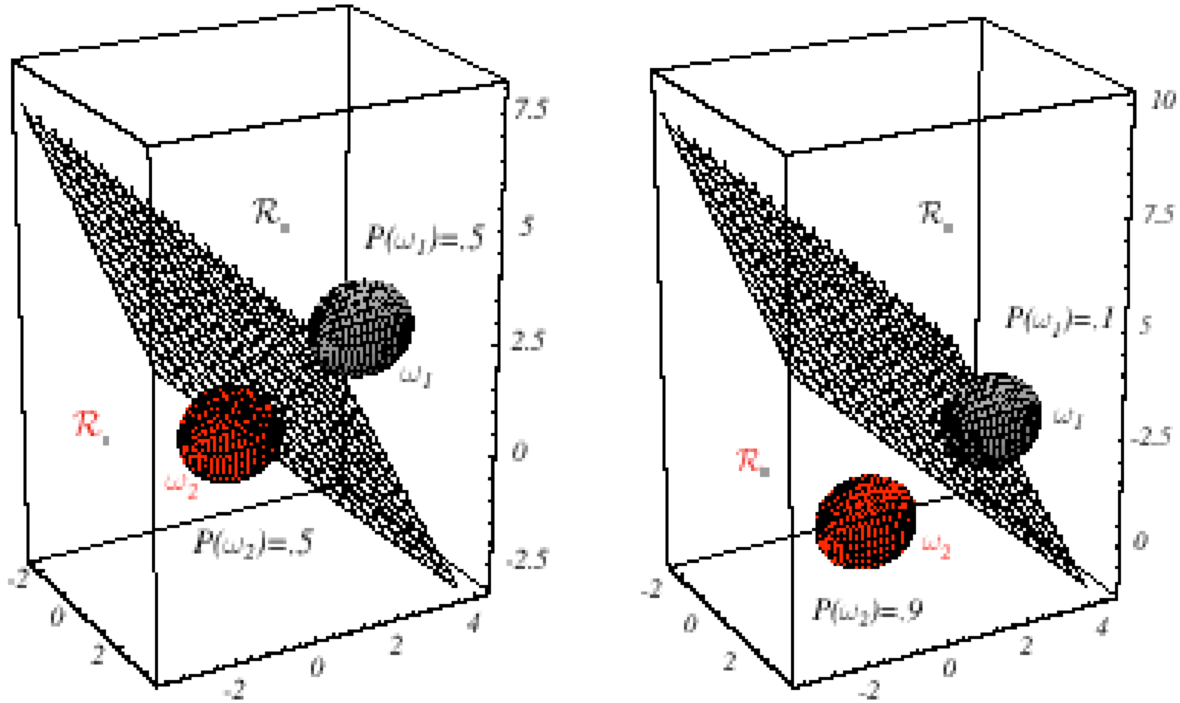


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

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

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

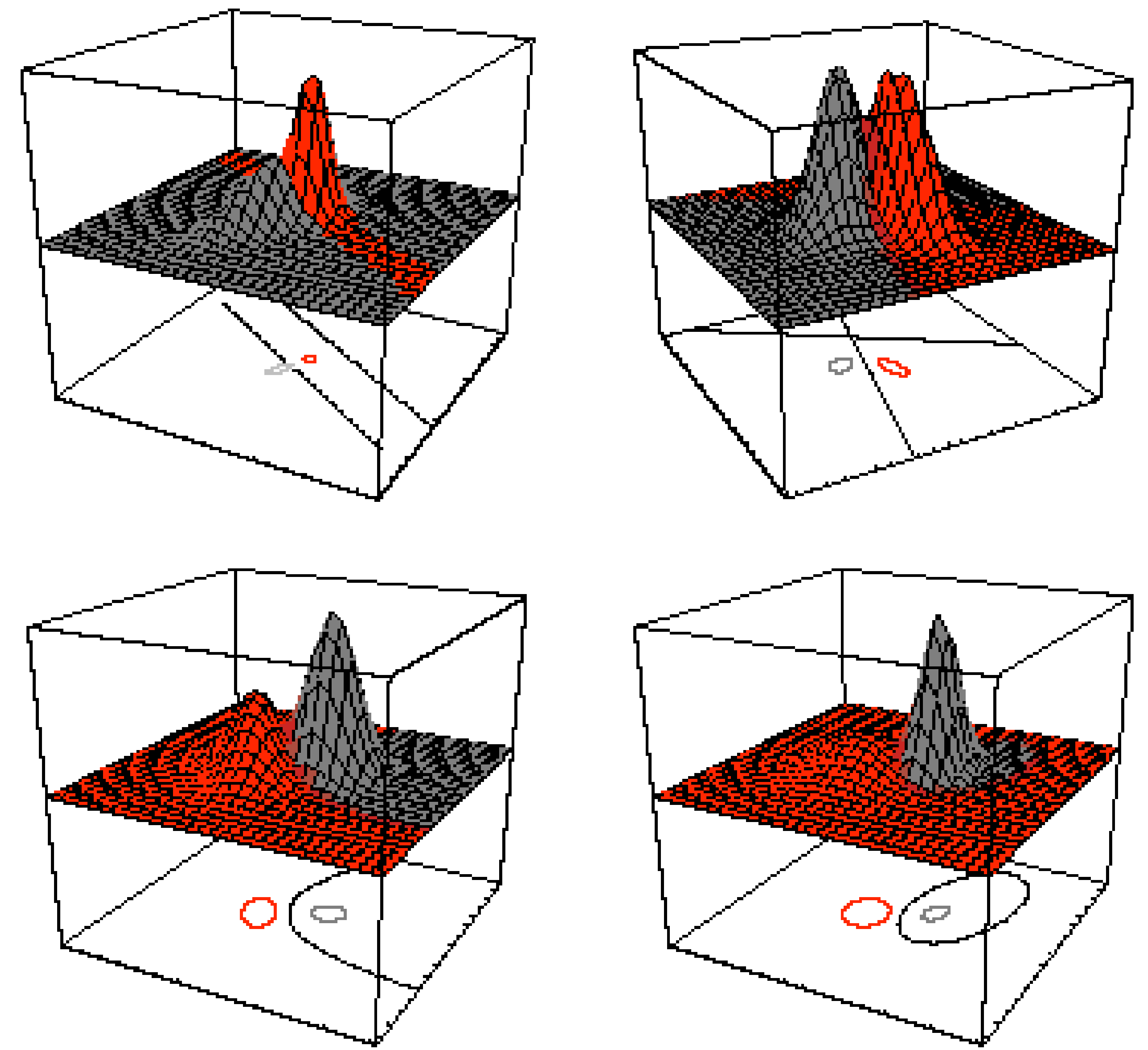
where :

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

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

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

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



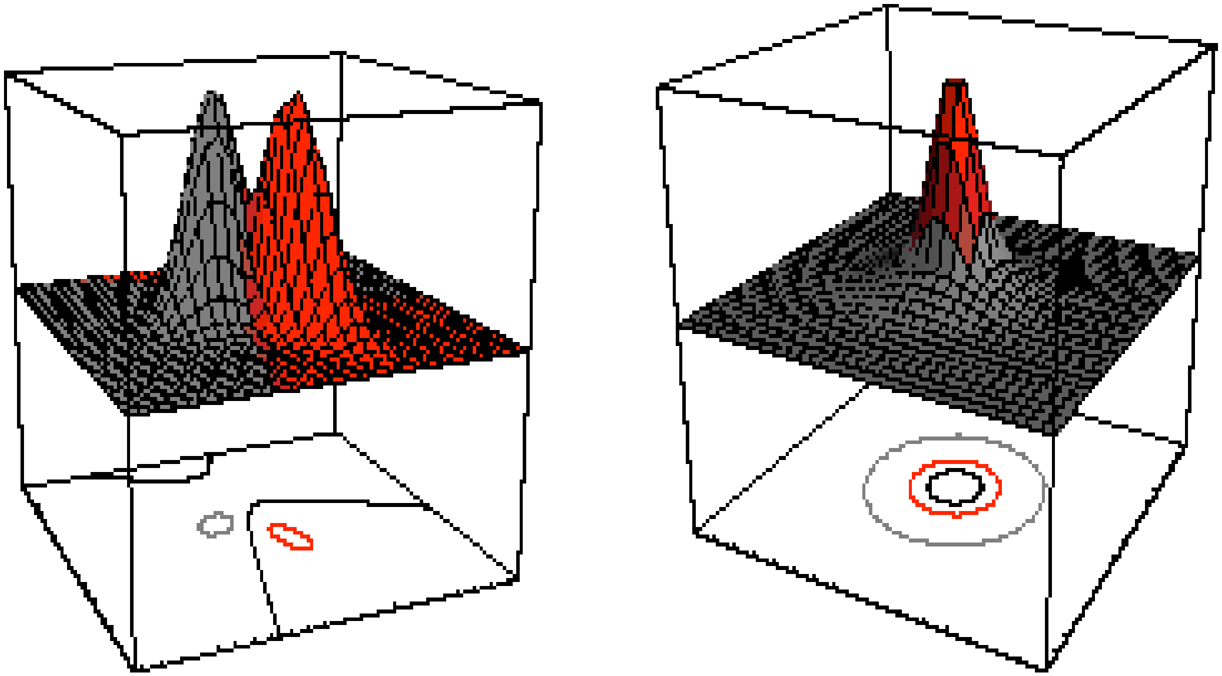


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Bayes Decision Theory – Discrete Features

- Components of x are binary or integer valued, x can take only one of m discrete values

$$v_1, v_2, \ldots, v_m$$

- Case of independent binary features in 2 category problem

  Let $x = [x_1, x_2, \ldots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 \mid \omega_1)$$
$$q_i = P(x_i = 1 \mid \omega_2)$$

- The discriminant function in this case is:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

where :

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1,\ldots,d$$

and :

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

decide ω_1 *if* $g(x) > 0$ *and* ω_2 *if* $g(x) \leq 0$