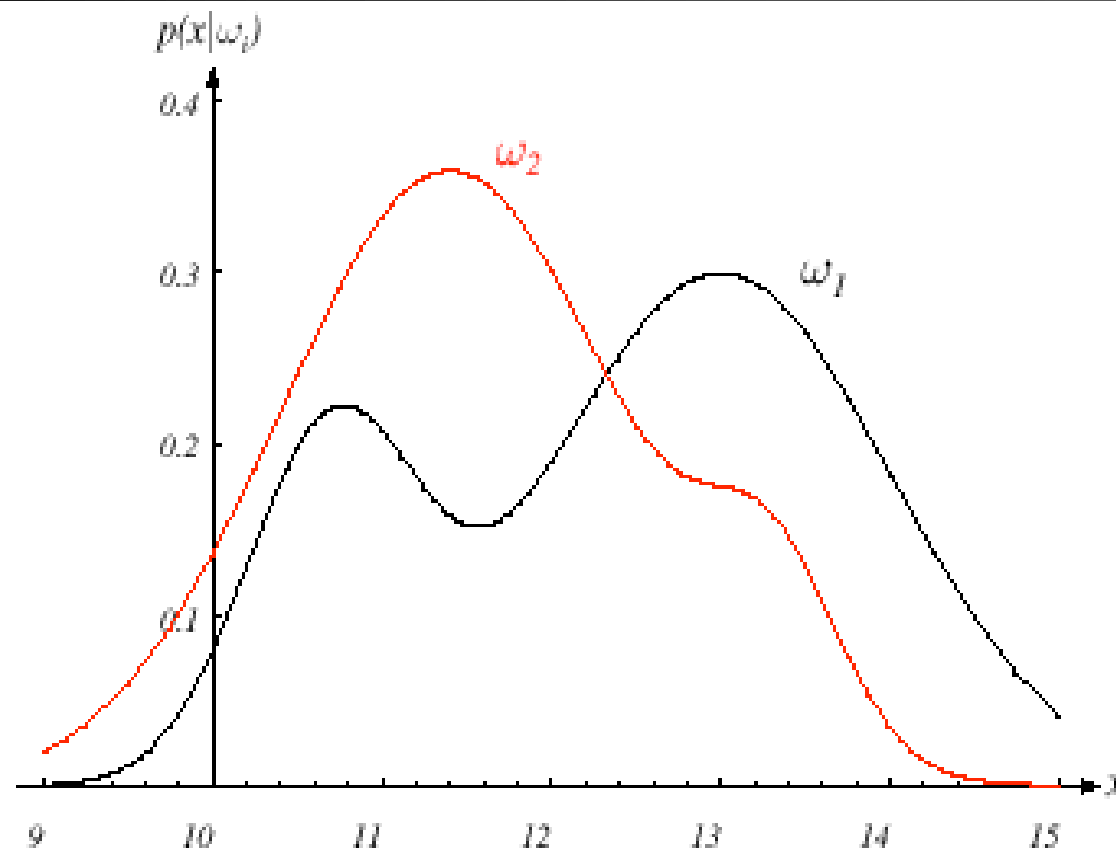


# Pattern Classification

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  $\omega_i$ . 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.

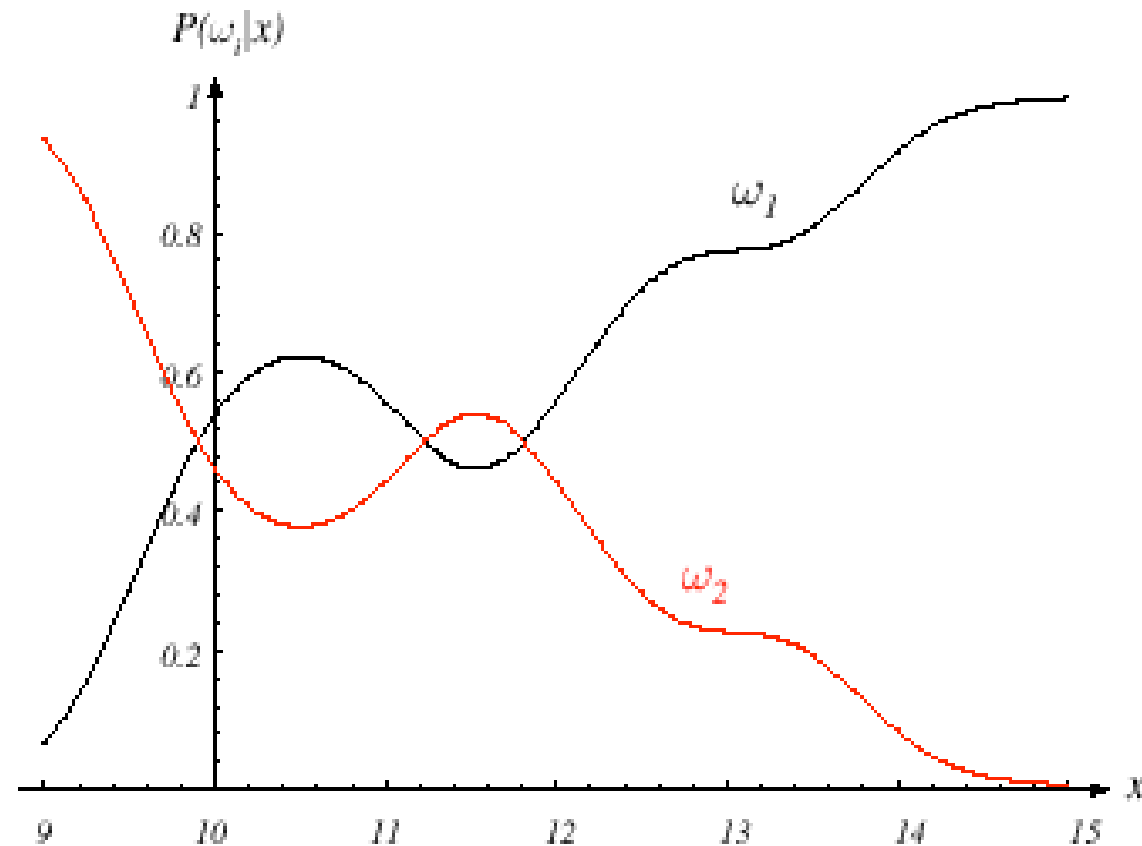
- Posterior, likelihood, evidence

- $P(\omega_j | \mathbf{x}) = P(\mathbf{x} | \omega_j) \cdot P(\omega_j) / P(\mathbf{x})$

- Where in case of two categories

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

- 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  $\omega_2$  is roughly 0.08, and that it is in  $\omega_1$  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  $P(\omega_1 | x) > P(\omega_2 | x)$   $\implies$  True state of nature =  $\omega_1$

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

Therefore:

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

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

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

- Minimizing the probability of error
- Decide  $\omega_1$  if  $P(\omega_1 | x) > P(\omega_2 | x)$ ;  
otherwise decide  $\omega_2$

Therefore:

$$P(\text{error} | x) = \min [P(\omega_1 | x), P(\omega_2 | x)]$$

(Bayes decision)

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

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

Let  $\lambda(\alpha_i | \omega_j)$  be the loss incurred for taking  
action  $\alpha_i$  when the state of nature is  $\omega_j$

Overall risk

$R = \text{Sum of all } R(\alpha_i | x) \text{ for } i = 1, \dots, a$

  
Conditional risk

Minimizing  $R \iff$  Minimizing  $R(\alpha_i | x)$  for  $i = 1, \dots, a$

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

for  $i = 1, \dots, a$



- Two-category classification

$\alpha_1$  : deciding  $\omega_1$

$\alpha_2$  : deciding  $\omega_2$

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

loss incurred for deciding  $\omega_i$  when the true state of nature is  $\omega_j$

Conditional risk:

$$R(\alpha_1 | x) = \lambda_{11}P(\omega_1 | x) + \lambda_{12}P(\omega_2 | x)$$

$$R(\alpha_2 | x) = \lambda_{21}P(\omega_1 | x) + \lambda_{22}P(\omega_2 | x)$$

Our rule is the following:

if  $R(\alpha_1 | x) < R(\alpha_2 | x)$   
action  $\alpha_1$ : “decide  $\omega_1$ ” is taken

This results in the equivalent rule :

decide  $\omega_1$  if:

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

and decide  $\omega_2$  otherwise

## Likelihood ratio:

The preceding rule is equivalent to the following rule:

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

Then take action  $\alpha_1$  (decide  $\omega_1$ )

Otherwise take action  $\alpha_2$  (decide  $\omega_2$ )

## 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  $\alpha_i$  is taken and the true state of nature is  $\omega_j$  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:

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

Therefore, the conditional risk is:

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{j=1}^{j=c} \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x}) \\ &= \sum_{j \neq i} P(\omega_j | \mathbf{x}) = 1 - P(\omega_i | \mathbf{x}) \end{aligned}$$

*“The risk corresponding to this loss function is the average probability error”*

- Minimize the risk requires maximize  $P(\omega_i | x)$   
(since  $R(\alpha_i | x) = 1 - P(\omega_i | x)$ )
- For Minimum error rate
  - Decide  $\omega_i$  if  $P(\omega_i | x) > P(\omega_j | x) \forall j \neq i$

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

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

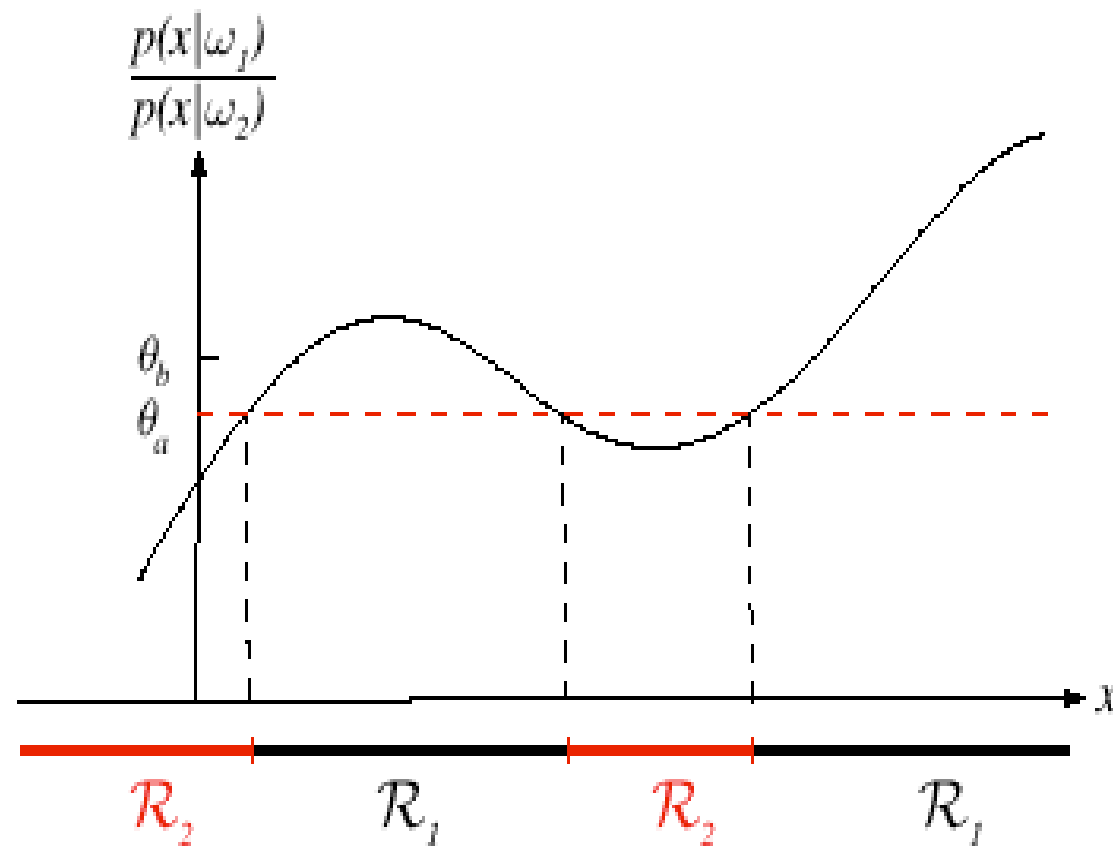
- If  $\lambda$  is the zero-one loss function which means:

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

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

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



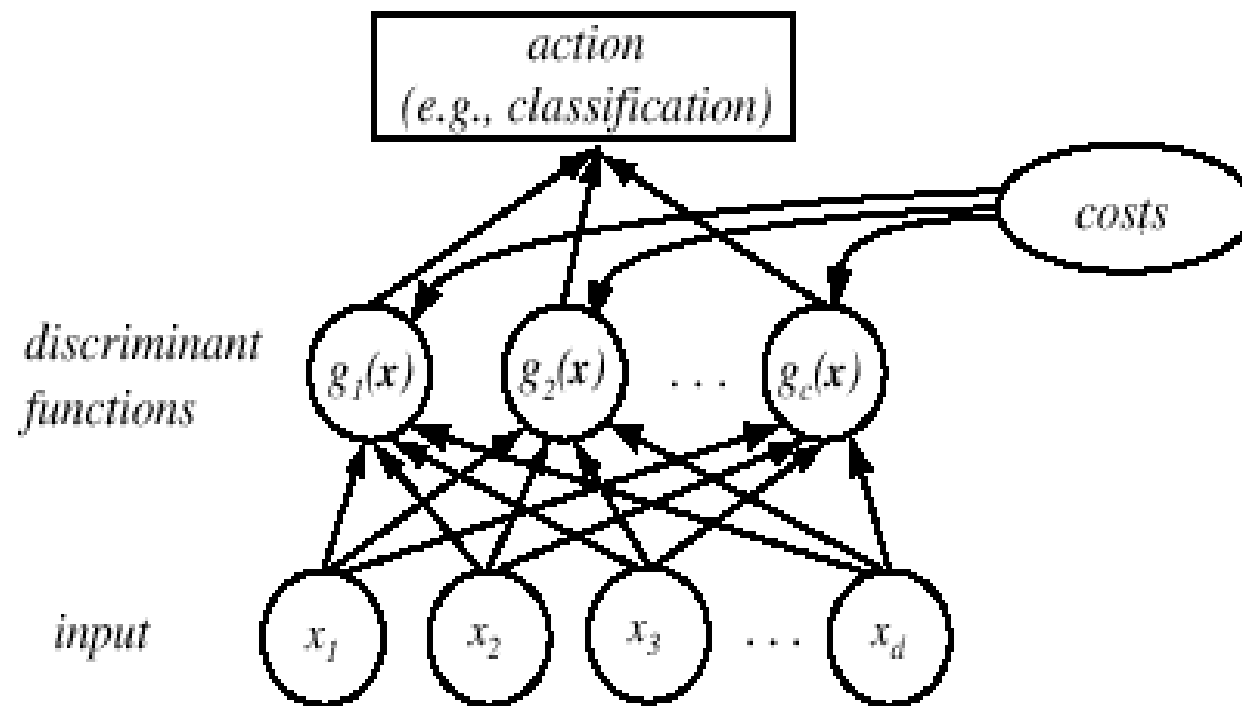


**FIGURE 2.3.** The likelihood ratio  $p(x|\omega_1)/p(x|\omega_2)$  for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold  $\theta_a$ . If our loss function penalizes miscategorizing  $\omega_2$  as  $\omega_1$  patterns more than the converse, we get the larger threshold  $\theta_b$ , 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 17

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

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



**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  $g_i(x) = -R(\alpha_i | x)$   
(max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take

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

(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!)

- 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  $\omega_i$ )*

- The two-category case

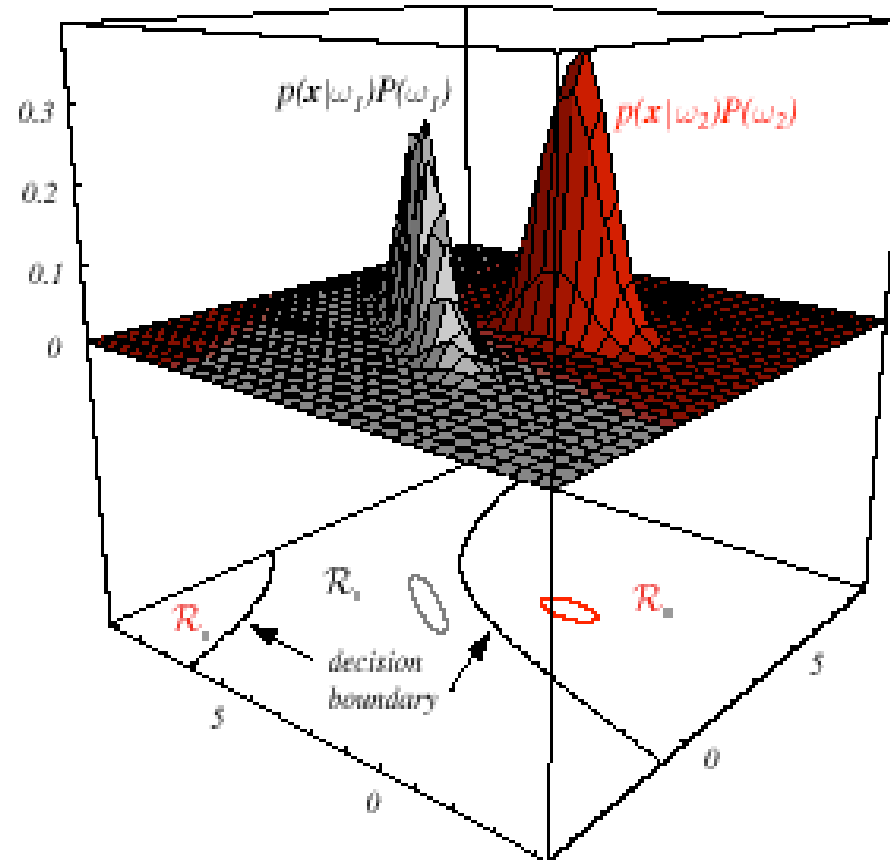
- A classifier is a “dichotomizer” that has two discriminant functions  $g_1$  and  $g_2$

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

**Decide  $\omega_1$  if  $g(x) > 0$  ; Otherwise decide  $\omega_2$**

- The computation of  $g(\mathbf{x})$

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



**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.

# Discriminant Functions for the Normal Density

23

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

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

- Case of multivariate normal

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$



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

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

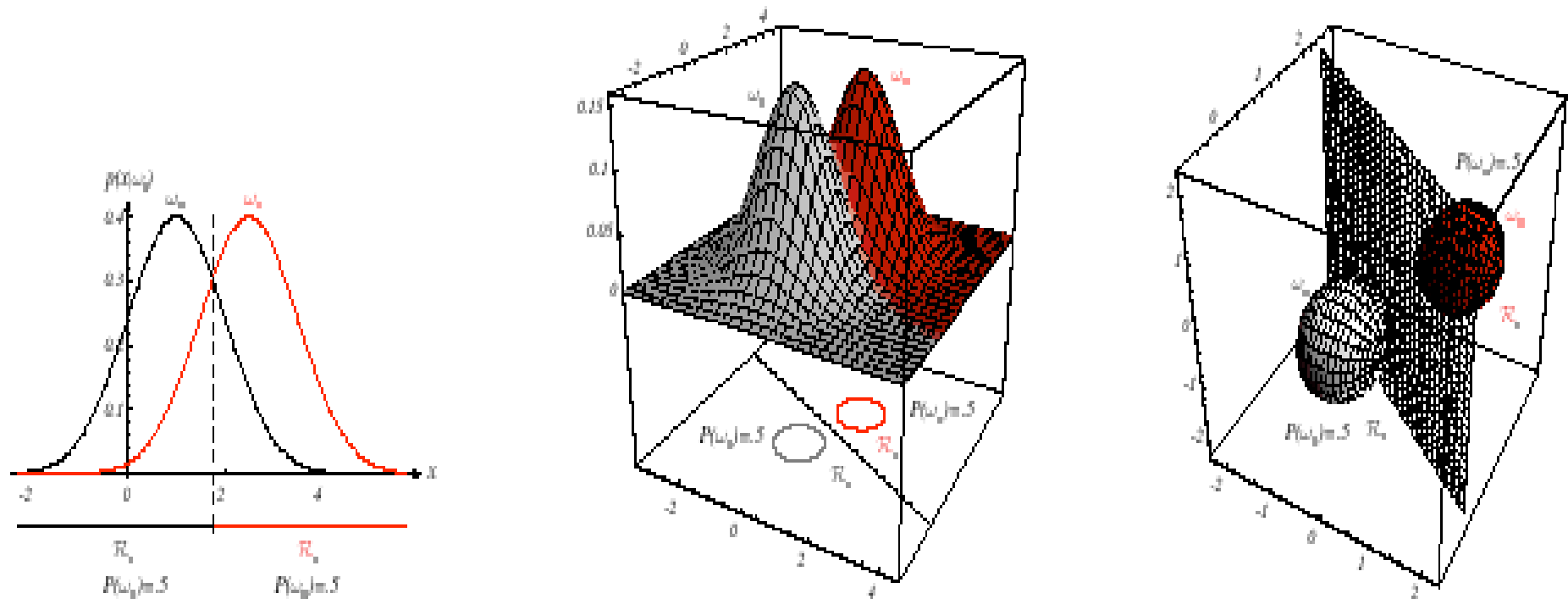
*where :*

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

*( $w_{i0}$  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:

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



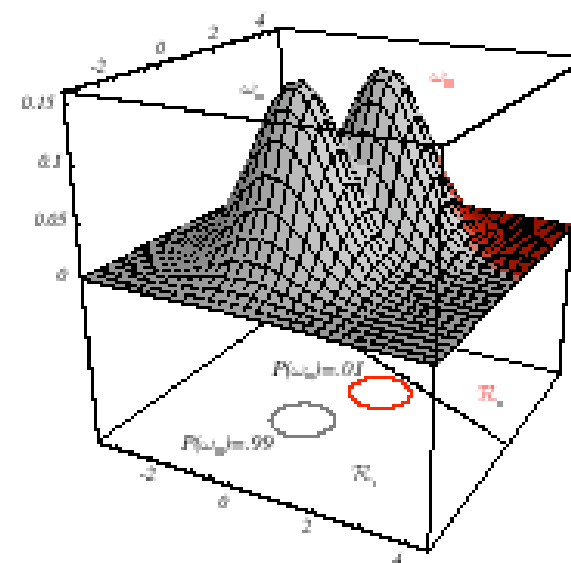
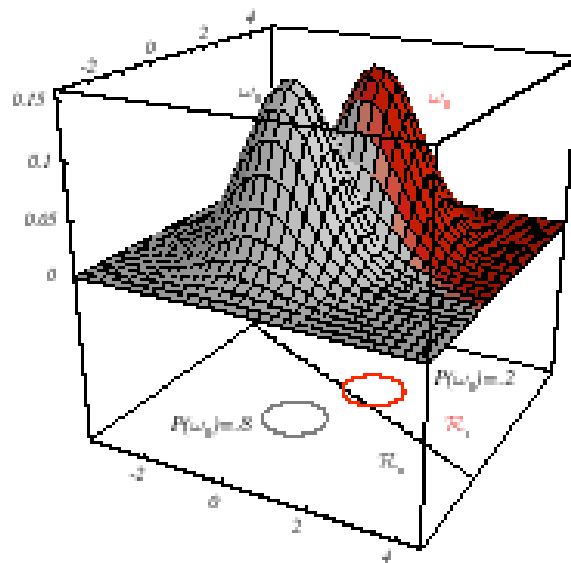
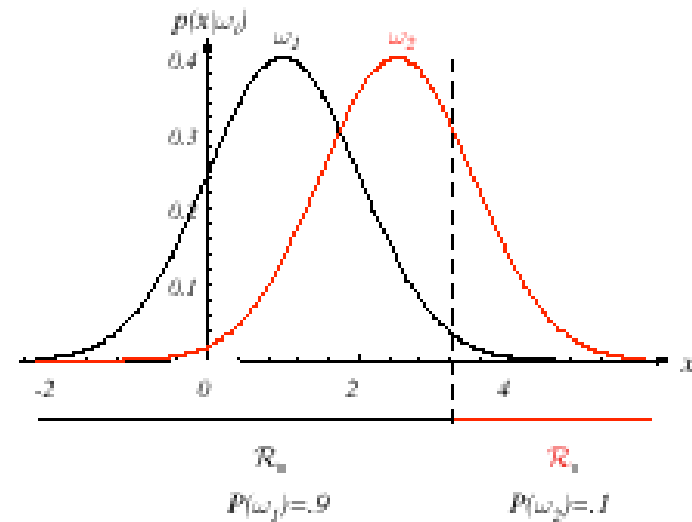
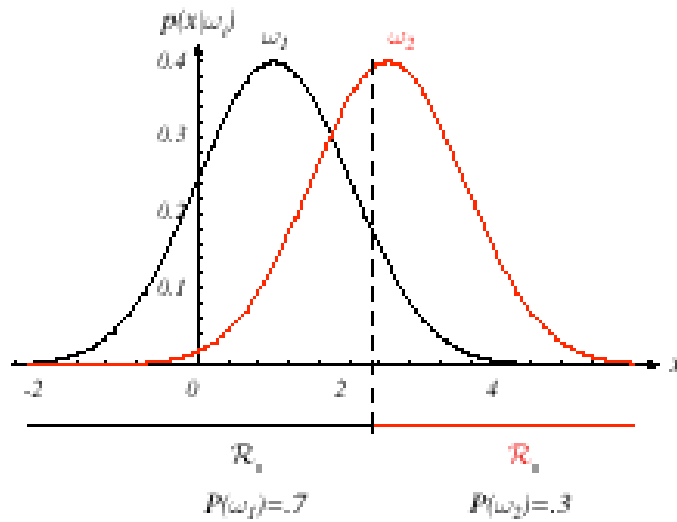
**FIGURE 2.10.** If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in  $d$  dimensions, and the boundary is a generalized hyperplane of  $d - 1$  dimensions, perpendicular to the line separating the means. In these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

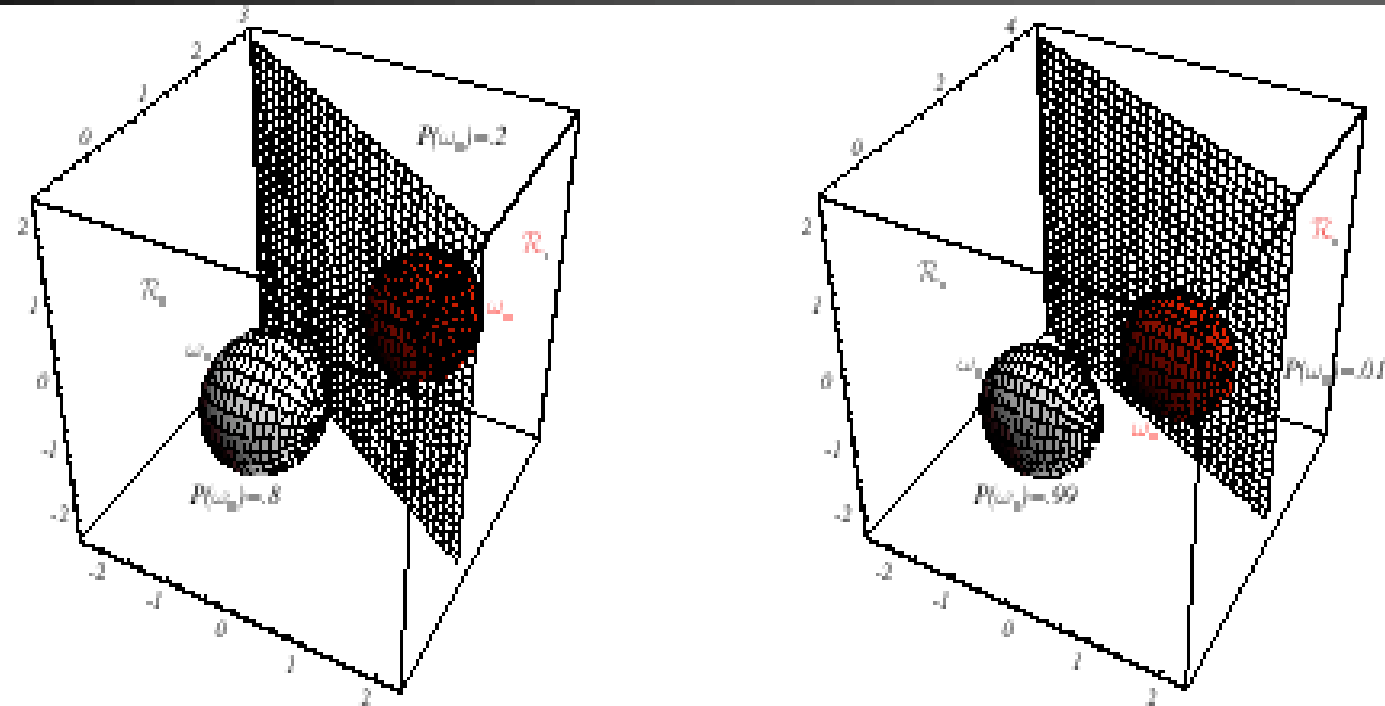
- The hyperplane separating  $R_i$  and  $R_j$

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

always orthogonal to the line linking the means!

$$\text{if } P(\omega_i) = P(\omega_j) \text{ then } \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$$



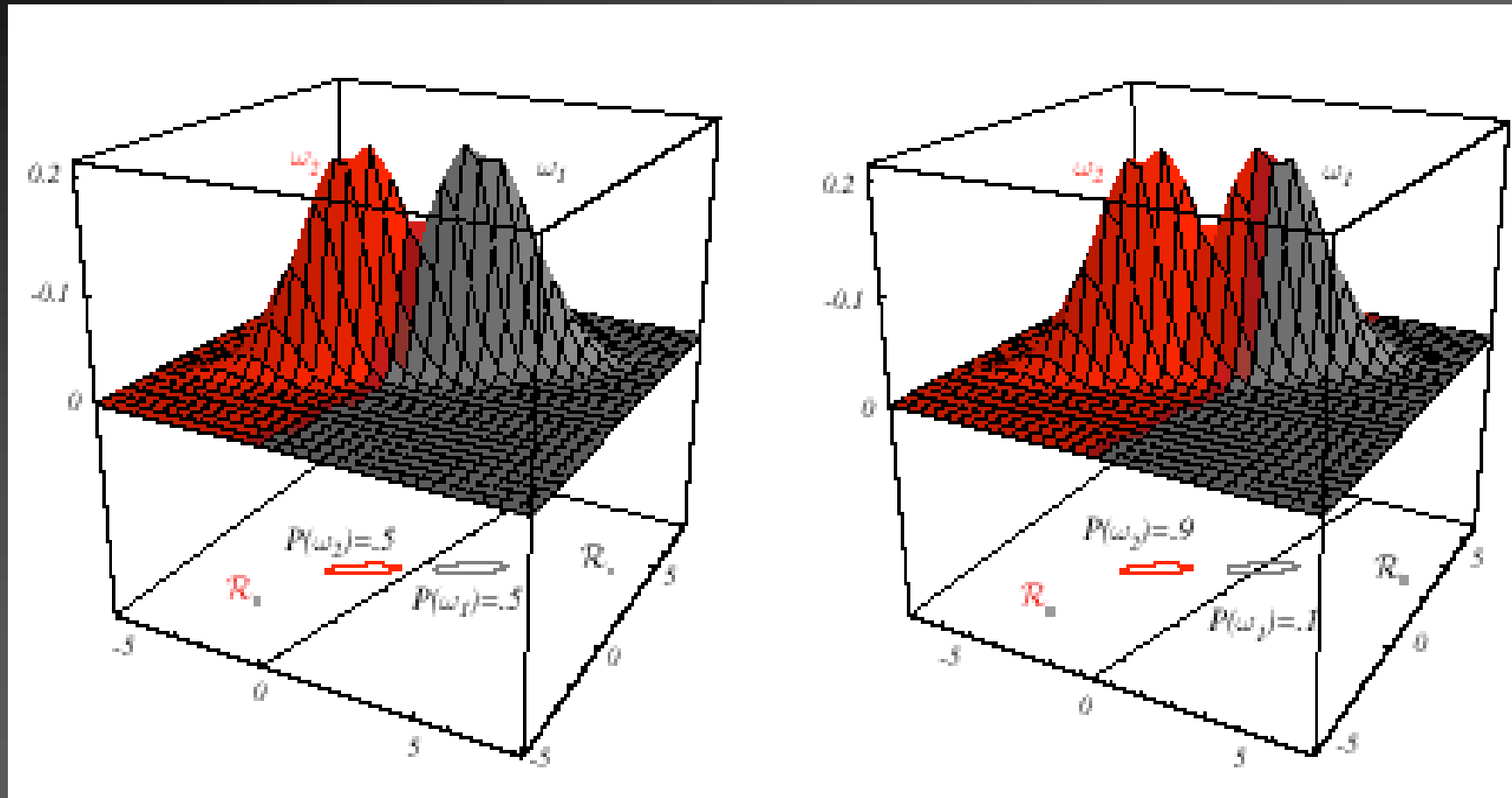


**FIGURE 2.11.** As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

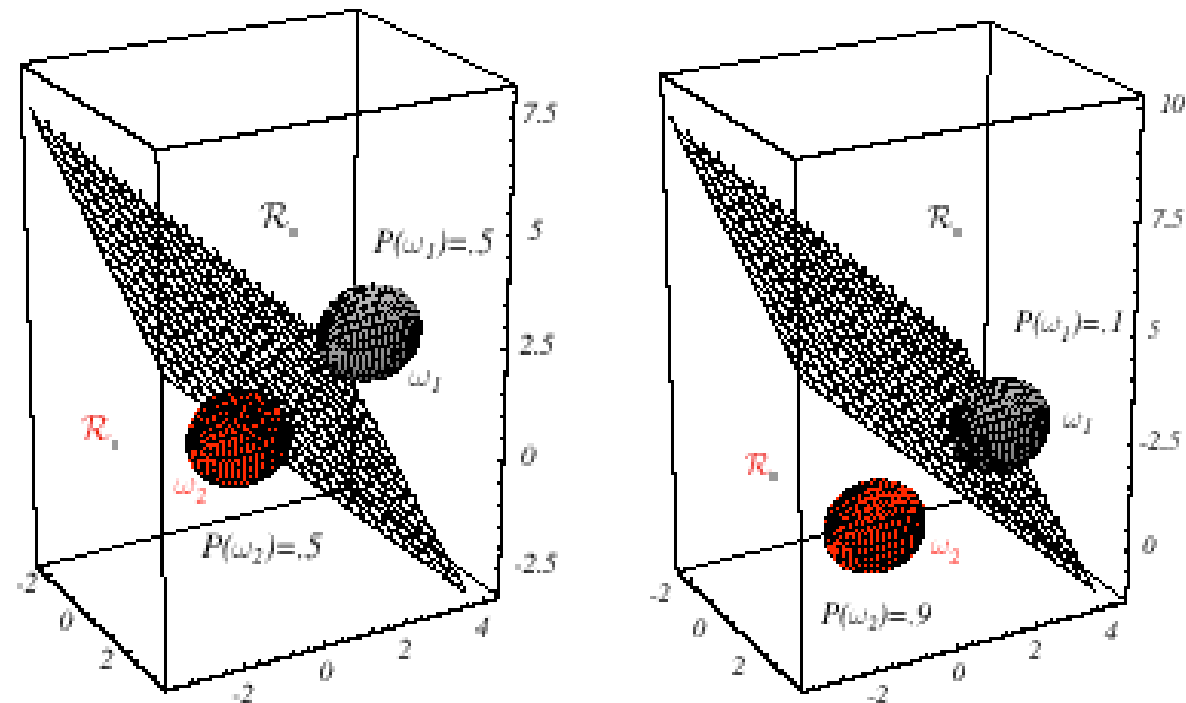
- **Case  $\Sigma_i = \Sigma$**  (covariance of all classes are identical but arbitrary!)
- Hyperplane separating  $R_i$  and  $R_j$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln[P(\omega_i) / P(\omega_j)]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \cdot (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

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







**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  $\Sigma_i = \text{arbitrary}$ 
  - The covariance matrices are different for each category

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} = w_{i0}$$

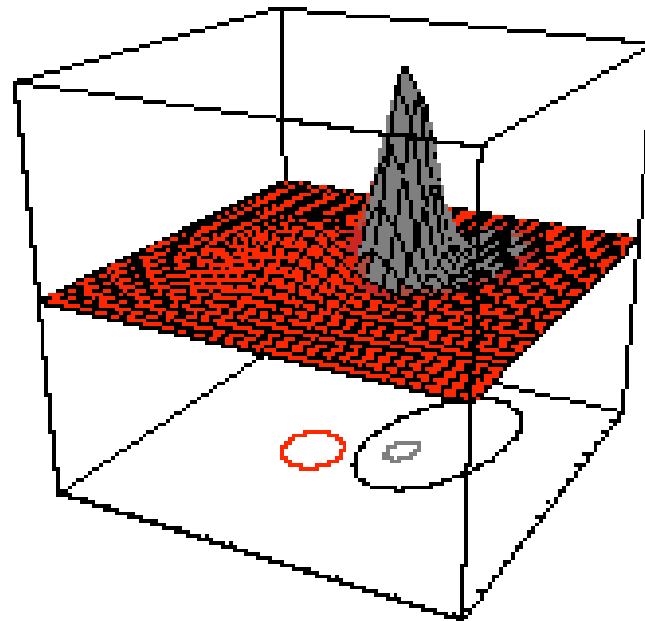
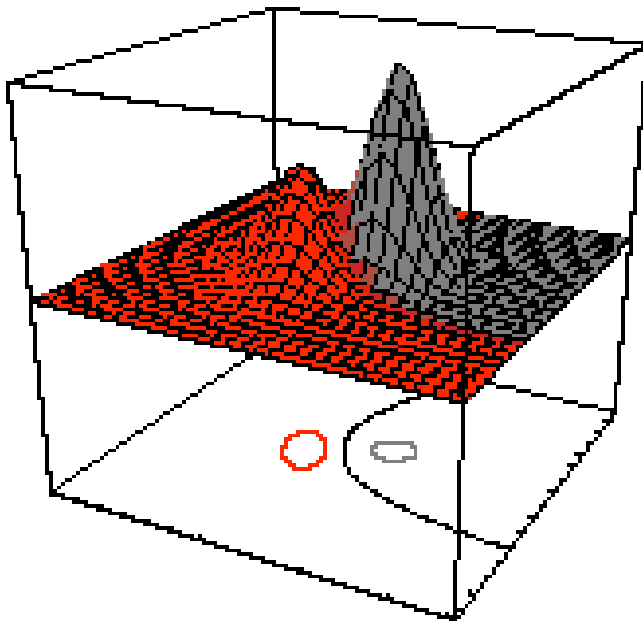
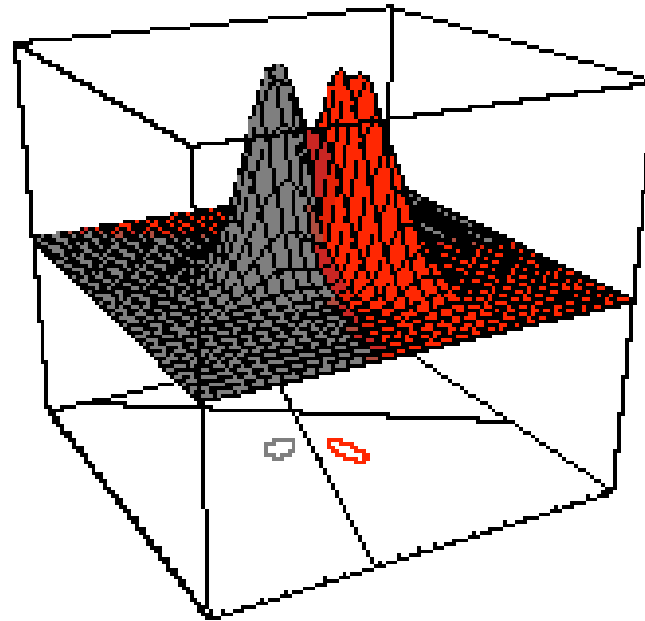
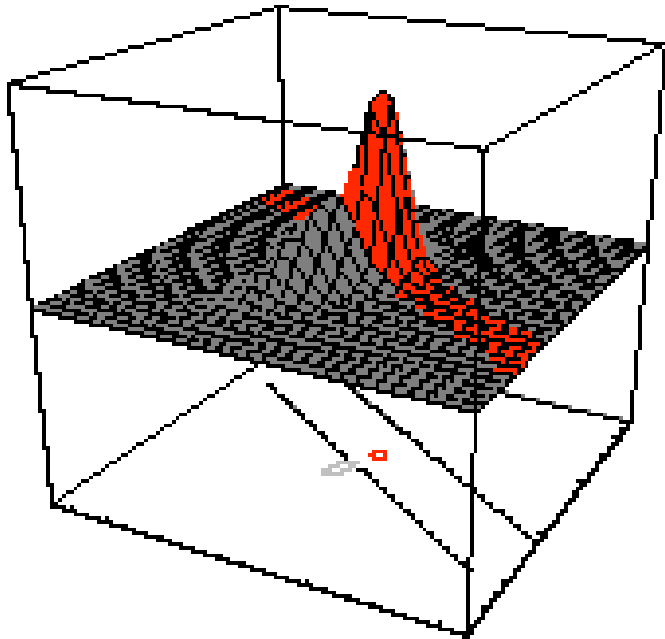
where :

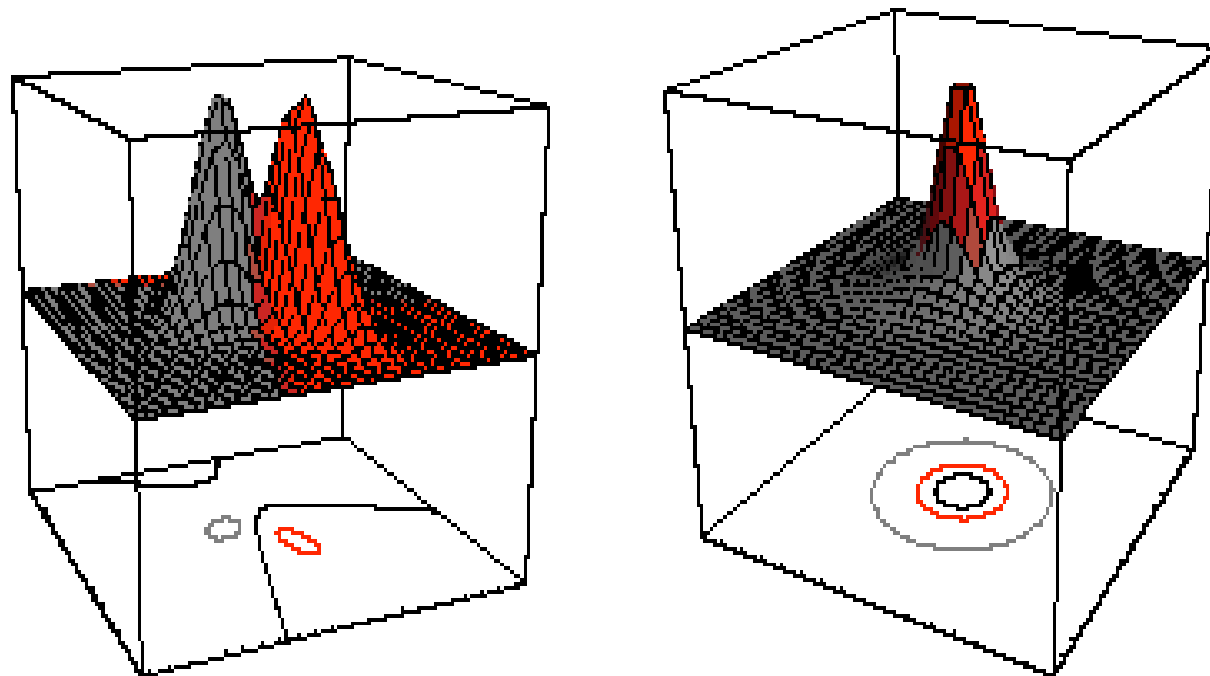
$$\mathbf{W}_i = -\frac{1}{2} \Sigma_i^{-1}$$

$$\mathbf{w}_i = \Sigma_i^{-1} \boldsymbol{\mu}_i$$

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

(**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.

# Bayes Decision Theory – Discrete Features

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

$$V_1, V_2, \dots, V_m$$

- Case of independent binary features in 2 category problem  
Let  $x = [x_1, x_2, \dots, x_d]^t$  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:

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

where :

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

and :

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

decide  $\omega_1$  if  $g(\mathbf{x}) > 0$  and  $\omega_2$  if  $g(\mathbf{x}) \leq 0$