Chapter 2 (part 3)
Bayesian Decision Theory
(Sections 2-6,2-9)

- Discriminant Functions for the Normal Density
- Error probabilities
- Bayes Decision Theory – Discrete Features

Discriminant Functions for the Normal Density

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

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

- Case of multivariate normal

$$P(x) = \frac{1}{(2p)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)\right]$$

$$g_i(x) = -\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m_i) - \frac{d}{2}\ln 2p - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

$$P(x) = \frac{1}{(2p)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)\right]$$

- Mahalanobis distance

The points of constant density are hyperellipsoids for which the quadratic form is constant

$$r^2 = \left[(x-m)^T \Sigma^{-1}(x-m)\right]$$

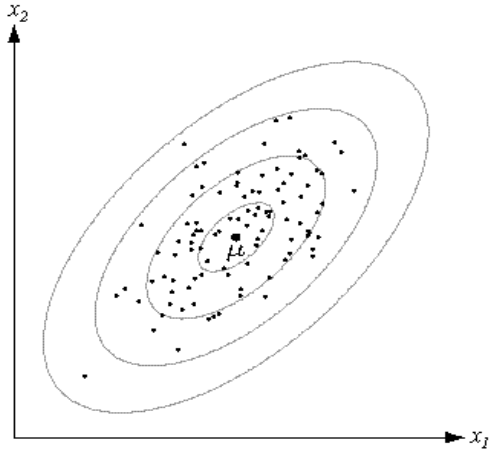


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Euclidean distance

$$r^2 = \left[(x-m)^T (x-m)\right]$$

First (independent) case

- Case $S_i = s^2.I$ (I stands for the identity matrix)

$$g_i(x) = -\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m_i) - \frac{d}{2}\ln 2p - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Ignoring quadratic term X^TX, the same for all classes, we get a linear discriminant function:

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

where :

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

(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:

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

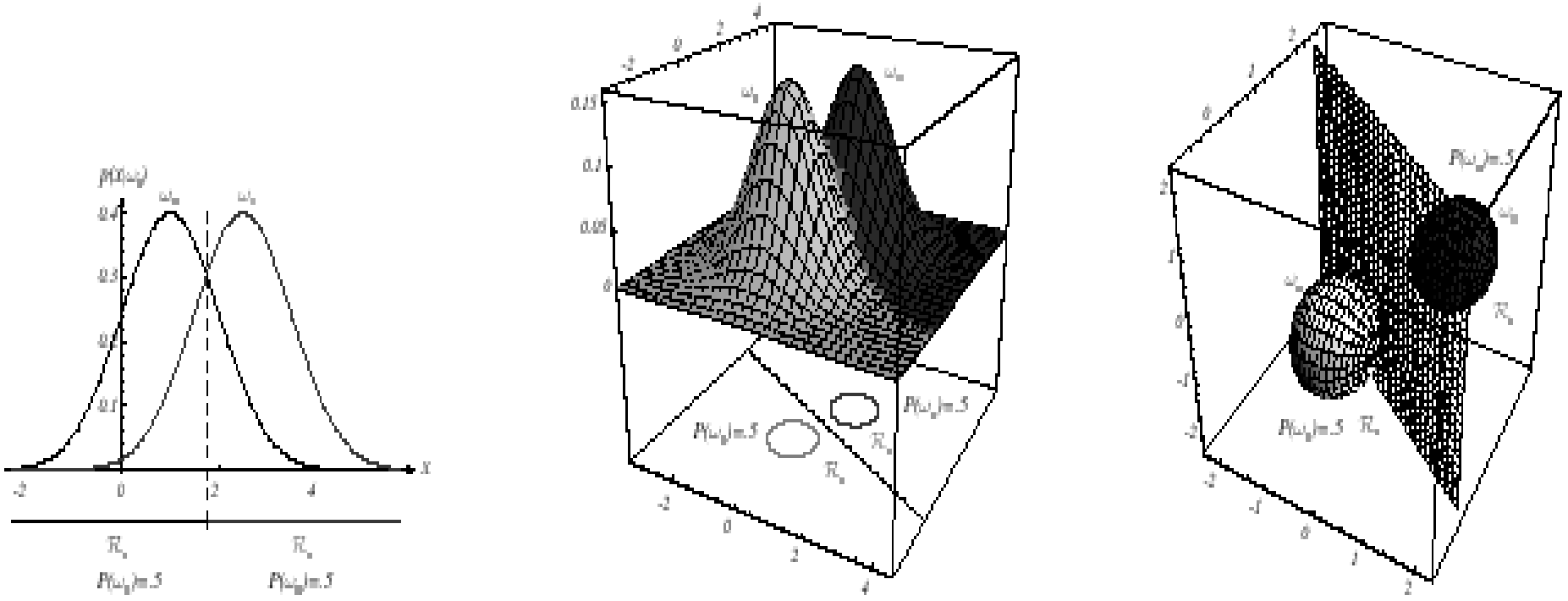


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

$$x_0 = \frac{1}{2}(m_i + m_j) - \frac{s^2}{\|m_i - m_j\|^2} \ln \frac{P(w_i)}{P(w_j)} (m_i - m_j)$$

always orthogonal to the line linking the means!

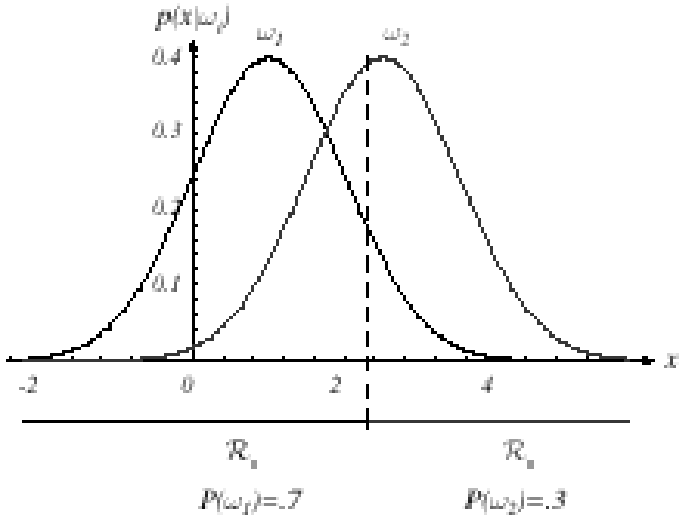
if $P(w_i) = P(w_j)$ then $g_i(x) = (x - m_i)^t (x - m_i)$

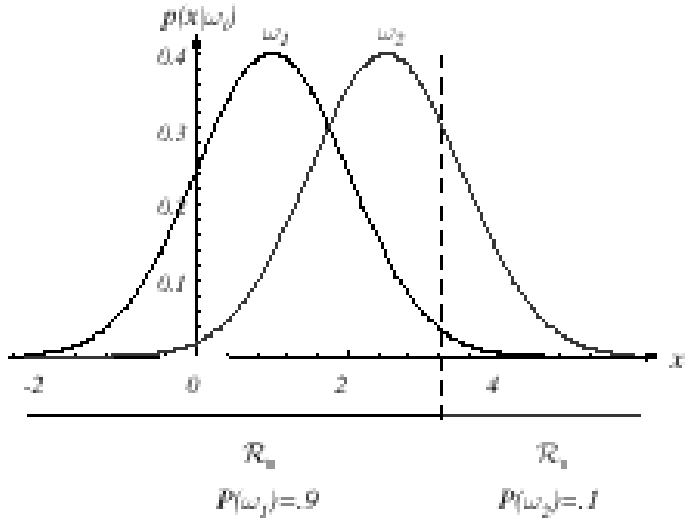
and $x_0 = \frac{1}{2}(m_i + m_j)$

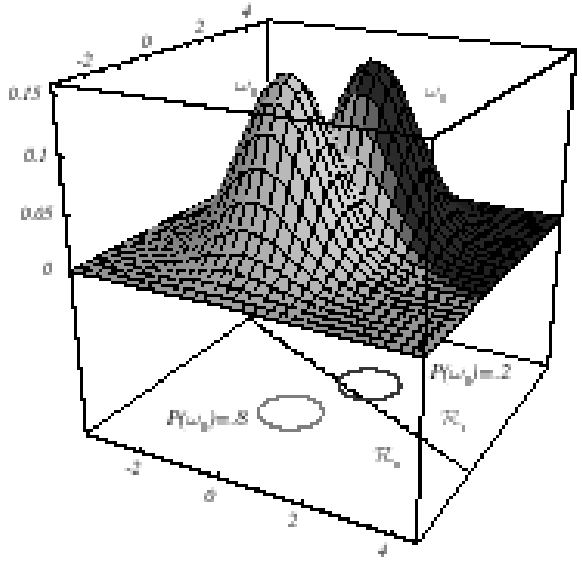
If prior probabilities are the same for all classes, then the input pattern sample should be classified into the class minimizing the Euclidean distance to its mean.

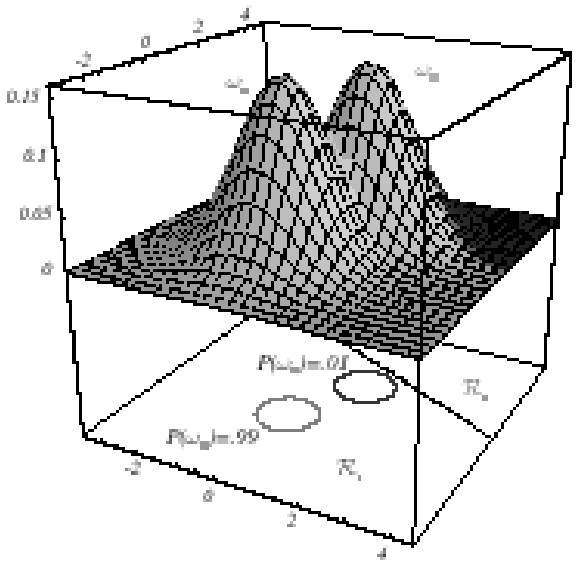
Minimum Distance Classifier :

- adopts a template-matching approach. The mean μ_k for each class C_k is assigned during training.
- For each new pattern sample, extract the feature vector and compute its Euclidean distance to class mean. Then, classify sample into the class minimizing this distance.


p(x|ω_j)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
R1
R2
P(ω1)=.7
P(ω2)=.3


p(x|ω_j)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
R1
R2
P(ω1)=.9
P(ω2)=.1


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8
R1
R2


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99
R1
R2

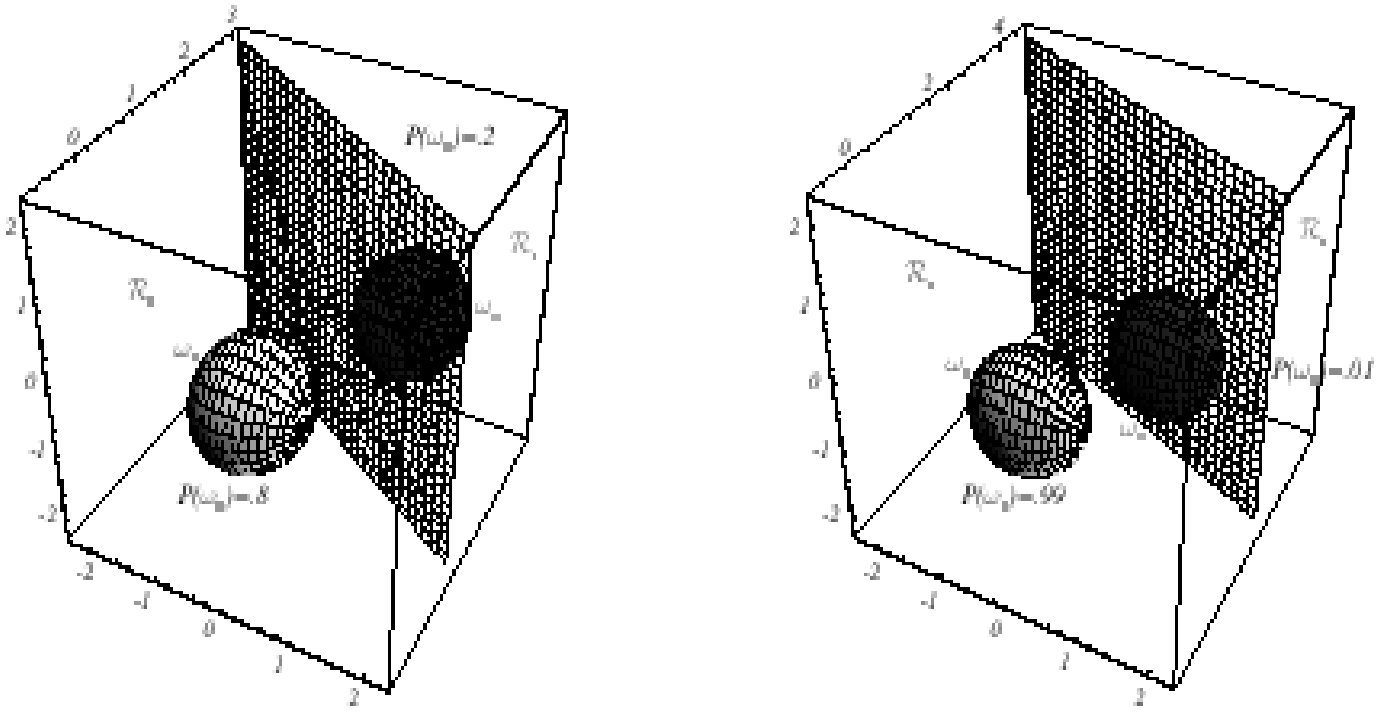


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Second Case

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

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

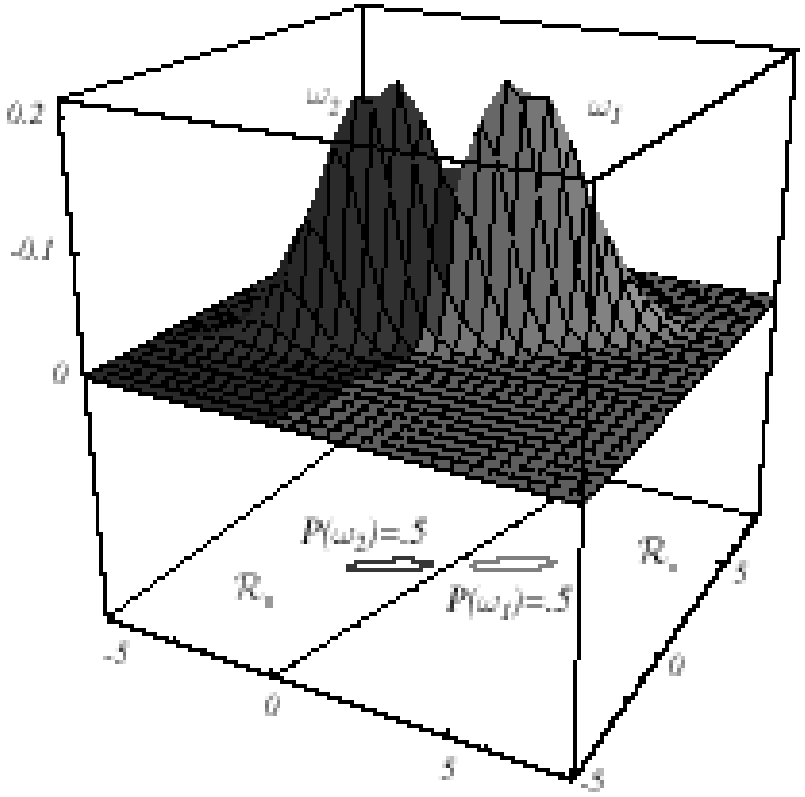
where :

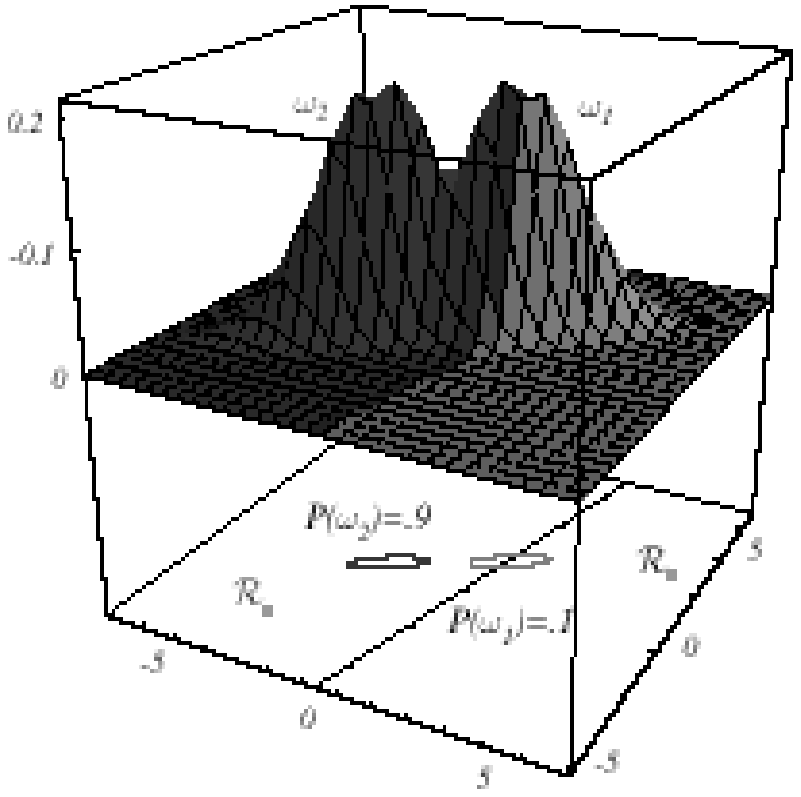
$$w_i = \Sigma^{-1} m_i; \quad w_{i0} = -\frac{1}{2} m_i^t \Sigma^{-1} m_i + \ln P(w_i)$$

  - Hyperplane separating R_i and R_j

$$x_0 = \frac{1}{2}(m_i + m_j) - \frac{\ln\left[P(w_i)/P(w_j)\right]}{(m_i - m_j)^t \Sigma^{-1} (m_i - m_j)} .(m_i - m_j)$$

(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
$P(\omega_2)=.5$
$P(\omega_1)=.5$
$\mathcal{R}_2$
$\mathcal{R}_1$
-5
0
5
-5
0
5


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

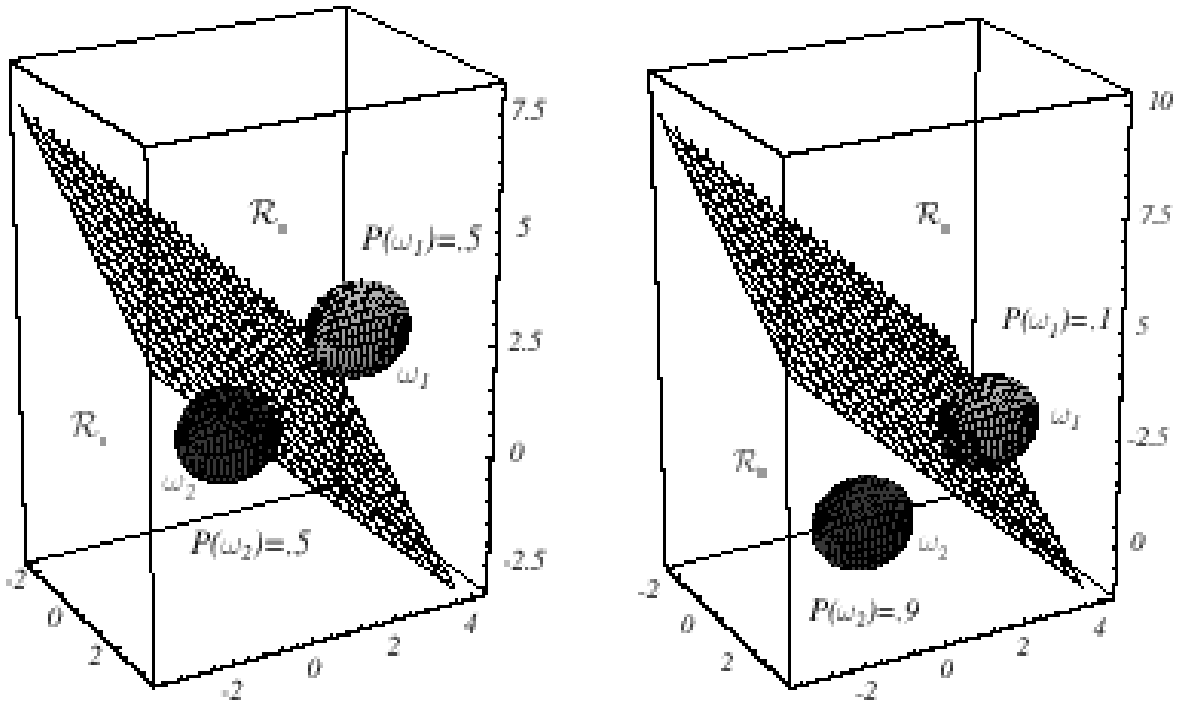


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, 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

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

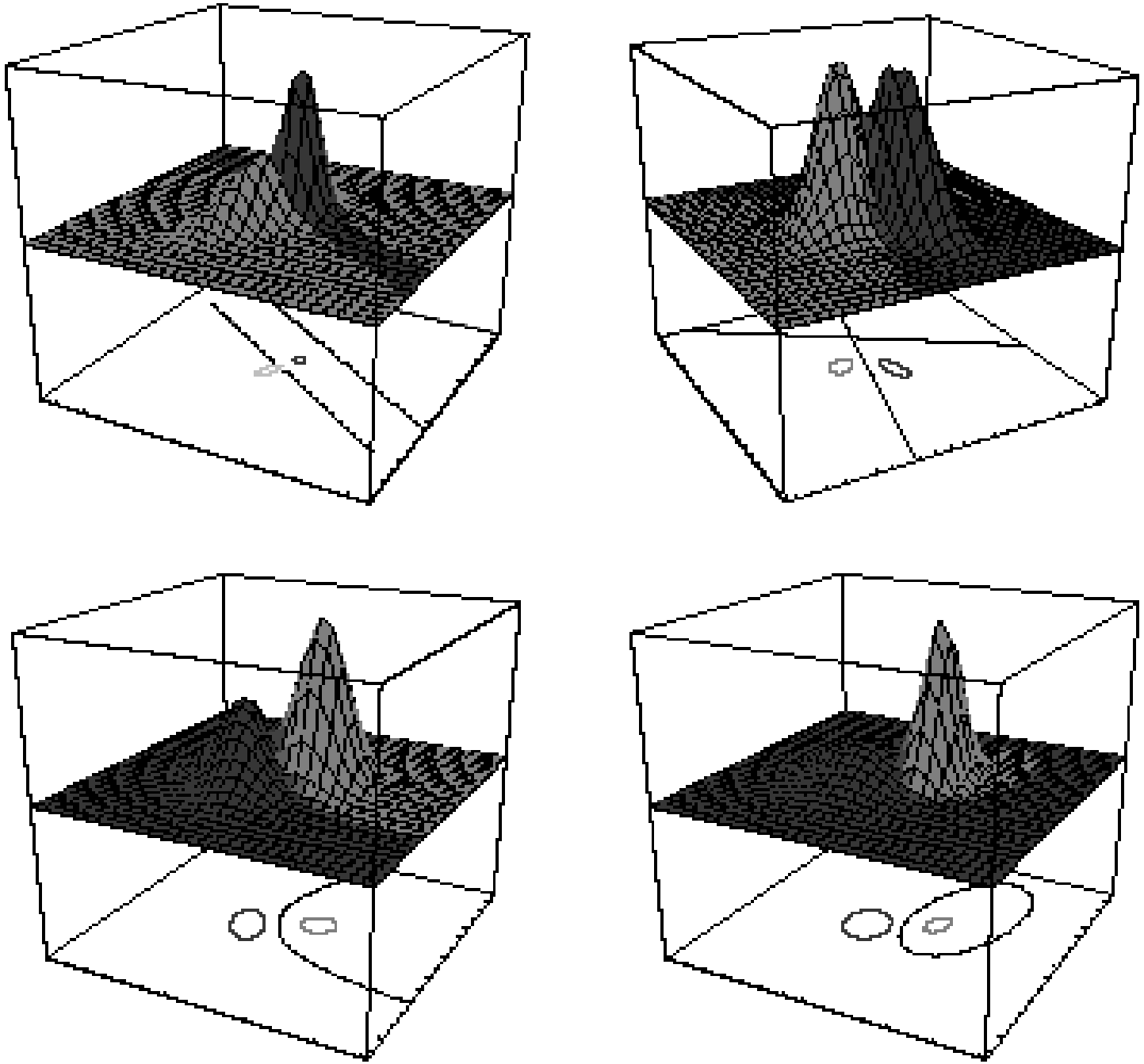
where :

$$\mathrm{W}_\mathrm{i} = -\frac{1}{2}\Sigma_i^{-1}$$

$$\mathrm{w}_\mathrm{i} = \Sigma_i^{-1} m_i$$

$$\mathrm{w}_{i0} = -\frac{1}{2} m_i^t \Sigma_i^{-1} m_i - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(w_i)$$

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



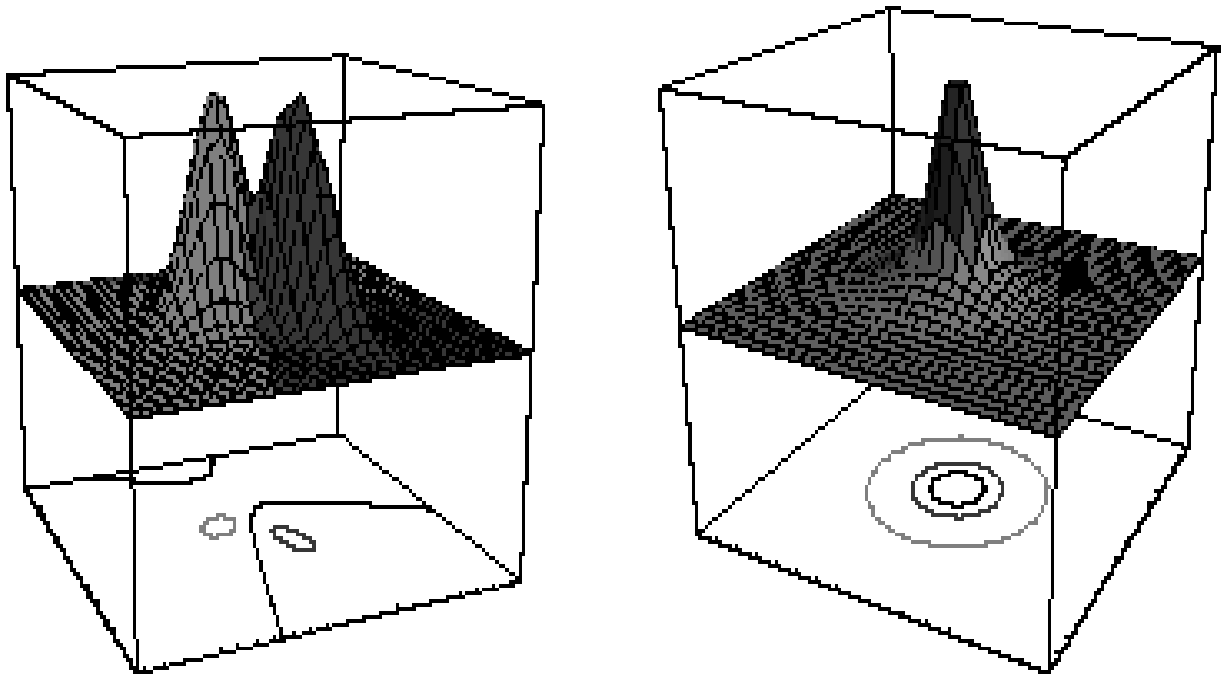


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Decision Regions Case Σ_i = arbitrary

$$g(x) = g_i(x) - g_j(x) = 0$$

$$g(x) = x^t A x + b^t x + c = 0$$

$$A = \sum_j^{-1} - \sum_i^{-1}$$

$$b = -2\sum_j^{-1} m_j + 2\sum_i^{-1} m_i$$

$$c = 2\ln\frac{P(w_i)}{P(w_j)} + \ln\frac{\left|\sum_j\right|}{\left|\sum_i\right|} + m_j{}^t \sum_j^{-1} m_j - m_i{}^t \sum_i^{-1} m_i$$

Example: 2 features, 2 categories

$$m_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix} \quad , \quad m_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad and \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\Sigma_1^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \quad , \quad \Sigma_2^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad P(w_1) = P(w_2)$$

$$x^t A x + b^t x + c = 0$$

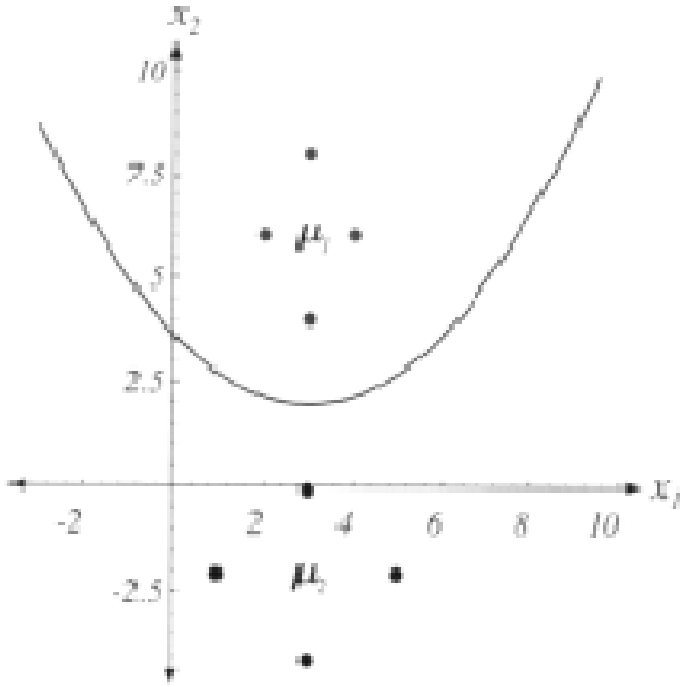
$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} -3/2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \left[2 \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \end{bmatrix} - 2 \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} \right]^t \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$+ \ln(4) + \begin{bmatrix} 3 & -2 \end{bmatrix} \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} - \begin{bmatrix} 3 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \end{bmatrix} = 0$$

$$-\frac{3}{2} x_1^2 + 9x_1 + 8x_2 - 28.1137 = 0$$

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

Example (cont.)



Discriminant functions for a multiple category system

Between two categories the boundary decision is a hyperquadric

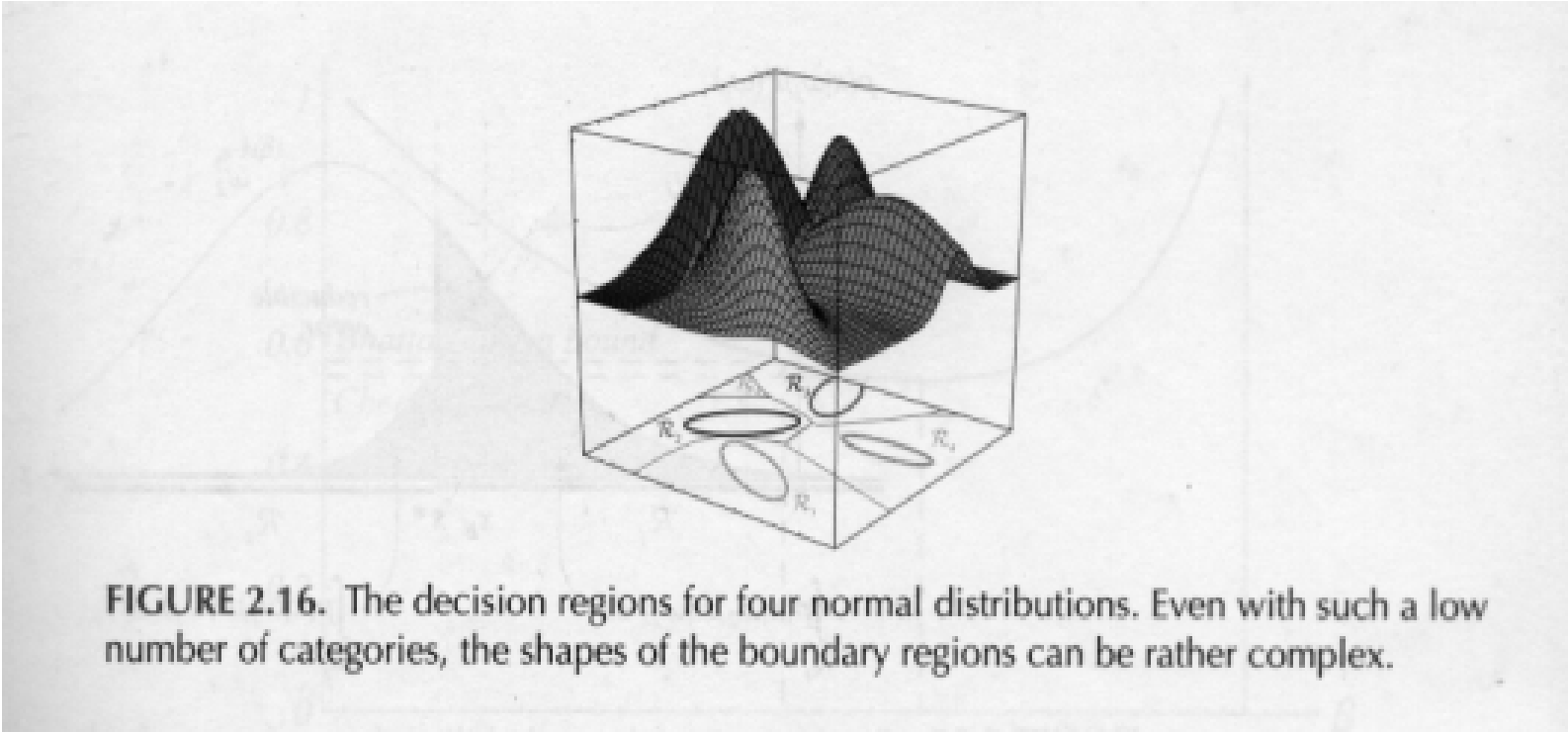


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.

Error Probabilities

P(error)=Error total=Sum of all error probabilities

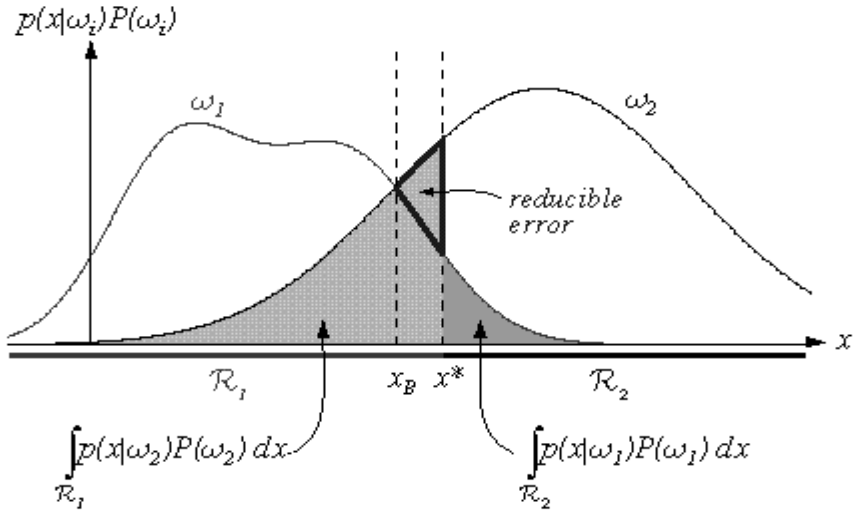


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible; this is the Bayes decision and gives the Bayes error rate. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Error Bounds

In most cases, it is difficult to compute the integrals for complex forms

What is the error bound?

Chernoff bound: $\min[a,b] \le a^{\beta} b^{1-\beta}$, $0 \le \beta \le 1$, $a,b \ge 0$

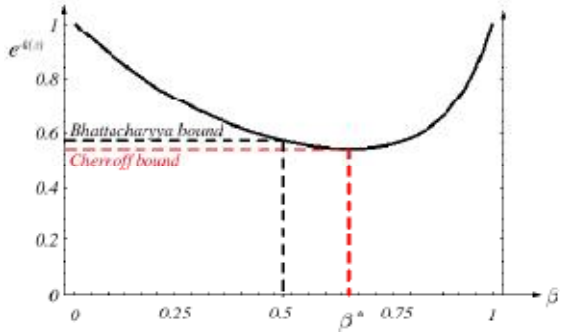
$$P(\text{error}) = \min(P(x|\omega_1)P(\omega_1), P(x|\omega_2)P(\omega_2))$$
$$\le \int P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)\, P^{\beta}(x|\omega_1)P^{1-\beta}(x|\omega_2)dx$$
$$= P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)e^{-k(\beta)} \text{ for normal distributions}$$

K(β): Equation 75

Bhattacharyya Bound

β=1/2

for the Gaussian case:



$$k(\frac{1}{2}) = \frac{1}{8}(m_2 - m_1)^t \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1}(m_2 - m_1) + \frac{1}{2}\ln\frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

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

$$P(\mathbf{x} \mid \omega_j) = \prod_{i=1}^{d} P(x_i \mid \omega_j)$$

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

$$p_i = P(x_i = 1 \mid w_1)$$
$$q_i = P(x_i = 1 \mid w_2)$$

Case of independent binary features in 2 categories

$$P(\mathbf{x}\,|\,\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

$$P(\mathbf{x}\,|\,\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$$

$$p_i = P(x_i = 1\,|\,\omega_1) \quad et \quad q_i = P(x_i = 1\,|\,\omega_2)$$

$$g(x) = P(\omega_1\,|\,\mathbf{x}) - P(\omega_2\,|\,\mathbf{x}) \Rightarrow g(x) = \ln\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

$$\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} = \prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

$$g(x) = \sum_{i=1}^{d}\left[x_i \ln\frac{p_i}{q_i} + (1-x_i)\ln\frac{1-p_i}{1-q_i}\right] + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

- The discriminant function in this case is:

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

where :

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

and :

$$w_0 = \sum_{i=1}^{d} \ln \frac{1 - p_i}{1 - q_i} + \ln \frac{P(w_1)}{P(w_2)}$$

decide w_1 ***if*** $g(x) > 0$ ***and*** w_2 ***if*** $g(x) \leq 0$

§ If $p_i=q_i$, g(x) depends only on $p(\omega_1)$ and $p(\omega_2)$

§ If pi>qi, 1-pi<1-qi so $w_i>0$. xi=1 results in favor of ω1.