CpE 646 Pattern Recognition and Classification

Prof. Hong Man

Department of Electrical and Computer Engineering
Stevens Institute of Technology

Bayesian Decision Theory

Chapter 2 (Section 2.5 – 2.9) Outline:

- The Normal Density
- Discriminant Functions for the Normal Density
- Error Probabilities and Integrals
- Error Bounds for Normal Densities
- Bayes Decision Theory – Discrete Features

The Normal Density

- Univariate normal density
  - Normal density which is analytically tractable
  - Continuous density
  - Many processes are asymptotically Gaussian or prototype corrupted by random processes
  - Definition $N(\mu, \sigma^2)$

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

The Normal Density

- μ : expected value of *x* or mean

$$\mu \triangleq E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

- σ^2 : expected squared deviation or variance

$$\sigma^2 \triangleq E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$$

- $H(p(x))$: uncertainty of sample values or entropy

$$H(p(x)) = -\int_{-\infty}^{\infty} p(x)\ln(p(x))dx$$

The Normal Density

– Central Limit Theorem: If the independent random variables $X_1, \ldots, X_n$ form a random sample of size n from a given distribution with mean μ and variance σ^2 $(0 < \sigma^2 < \infty)$, then for each fixed number z,

$$\lim_{n\to\infty} \Pr\left[\frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma} \le z\right] = \int_{-\infty}^{z} p(u)du,$$

$$\text{where } p(u) \sim N(0,1)$$

- If a large random sample is taken from any distribution with mean μ and variance σ^2, regardless of continuous or discrete distribution, the distribution of sample mean $\bar{X}_n$ will be approximately a normal distribution with mean μ and variance σ^2/n

The Normal Density

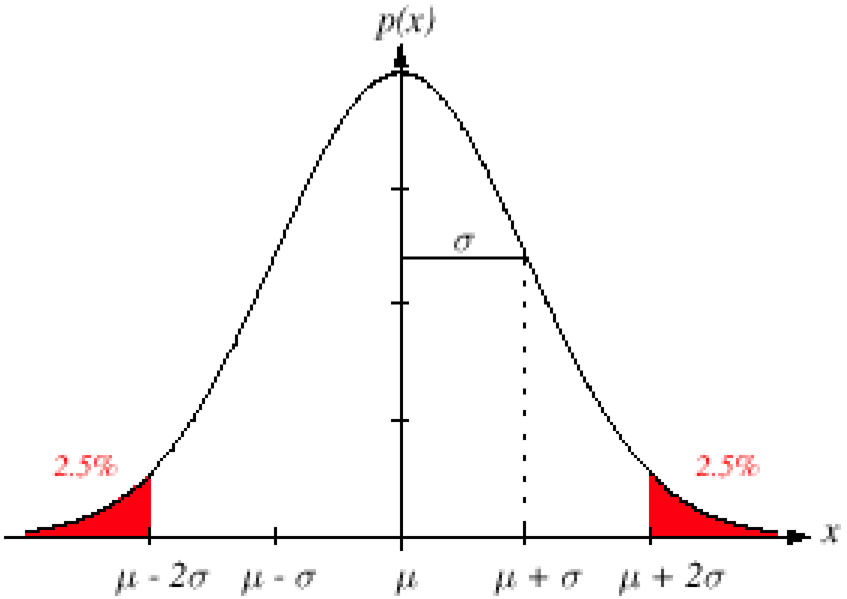


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

The Normal Density

- Multivariate normal density in d-dimension is $N(\mu, \Sigma)$:

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

where:

$\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^t$ (^{t} stands for transpose)

$\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_d]^t$ is the mean vector

$\boldsymbol{\Sigma}$ is the covariance matrix with size $d \times d$

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

$\boldsymbol{\Sigma}$ is always symmetric and positive semidefinite. We focus on cases in which $\boldsymbol{\Sigma}$ is positive definite, $|\boldsymbol{\Sigma}| > 0$.

The Normal Density

- Mean vector

$$\boldsymbol{\mu} \triangleq E[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}$$

  - The *i*-th element in the mean vector is

$$\mu_i \triangleq E[x_i] = \int_{-\infty}^{\infty} x_i p(x_i) dx_i$$

- Covariance matrix

$$\boldsymbol{\Sigma} \triangleq E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^t] = \int (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^t p(\boldsymbol{x}) d\boldsymbol{x}$$

  - The *ij*-th element in the covariance matrix is

$$\sigma_{ij}^2 \triangleq E[(x_i-\mu_i)(x_j-\mu_j)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i-\mu_i)(x_j-\mu_j) p(x_i, x_j) dx_i dx_j$$

The Normal Density

- The diagonal elements σ_{ii}^2 of covariance matrix are variances of x_i
- The off-diagonal elements σ_{ij}^2 are covariance of x_i and x_j.
- If x_i and x_j are statistically independent, then $\sigma_{ij}^2 = 0$.

The ***i*** and ***j*** here are the indices of the elements of the feature vector, and the corresponding elements of the mean vector and the covariance matrix

The Normal Density

- Linear combination (transform) of jointly normally distributed random variables, independent or not, are also normally distributed
  - Let **A** be a $d \times k$ matrix, and $y=\mathbf{A}^t x$ is a k-dimensional vector.
  - If $p(x) \sim N(\mu, \Sigma)$, then $p(y) \sim N(\mathrm{A}^t \mu, \mathrm{A}^t \Sigma \mathrm{A})$

The Normal Density

- A whitening transform converts an arbitrary multivariate normal distribution into a spherical one with covariance matrix $\Sigma \Rightarrow I$ (identity matrix)

  $$A_w = \Phi \Lambda^{-1/2}$$

  where columns of matrix Φ are eigenvectors of Σ, and the diagonal elements of matrix Λ are corresponding eigenvalues.

The Normal Density

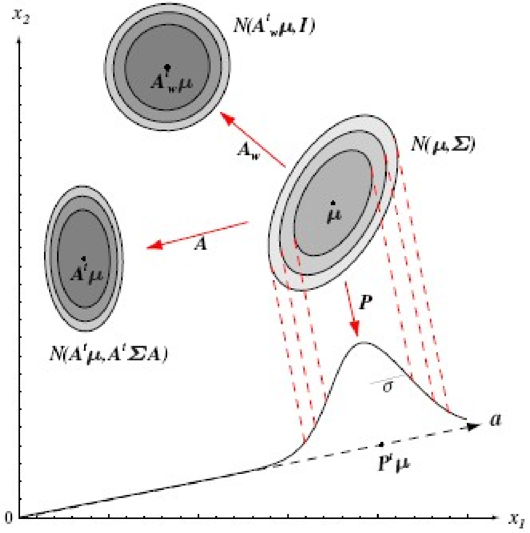


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, A, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\Sigma\mathbf{A})$. Another linear 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_1x_2-space. A whitening transform, $\mathbf{A}_w$, 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.

Discriminant Functions for Normal Density

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

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

- If the densities $p(x|\omega_i)$ are multivariate normal, $p(x|\omega_i) \sim N(\mu_\mathrm{i},\Sigma_i)$

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

Here after, the ***i*** and ***j*** are the indices of the classes

Discriminant Functions for Normal Density

- Case 1. $\Sigma_i = \sigma^2 I$, where I is identity matrix
  - – random variables (features) are independent and with the same variance σ^2
  - – $|\Sigma_i| = \sigma^{2d}$, $\Sigma_i^{-1} = (1/\sigma^2)I$
  - – Let $||x-\mu_i||^2 = (x-\mu_i)^t (x-\mu_i)$ be the Euclidean norm

Discriminant Functions for Normal Density

$$g_i(x) = -\frac{\left\| x - \mu_i \right\|^2}{2\sigma^2} + \ln P(\omega_i)$$

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

Because $x^t x / 2\sigma^2$ is the same for all classes of *x*, so it can be ignored, then we have

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

where:

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

(ω_{i0} is called the threshold or bias for the *i*th class)

Discriminant Functions for Normal Density

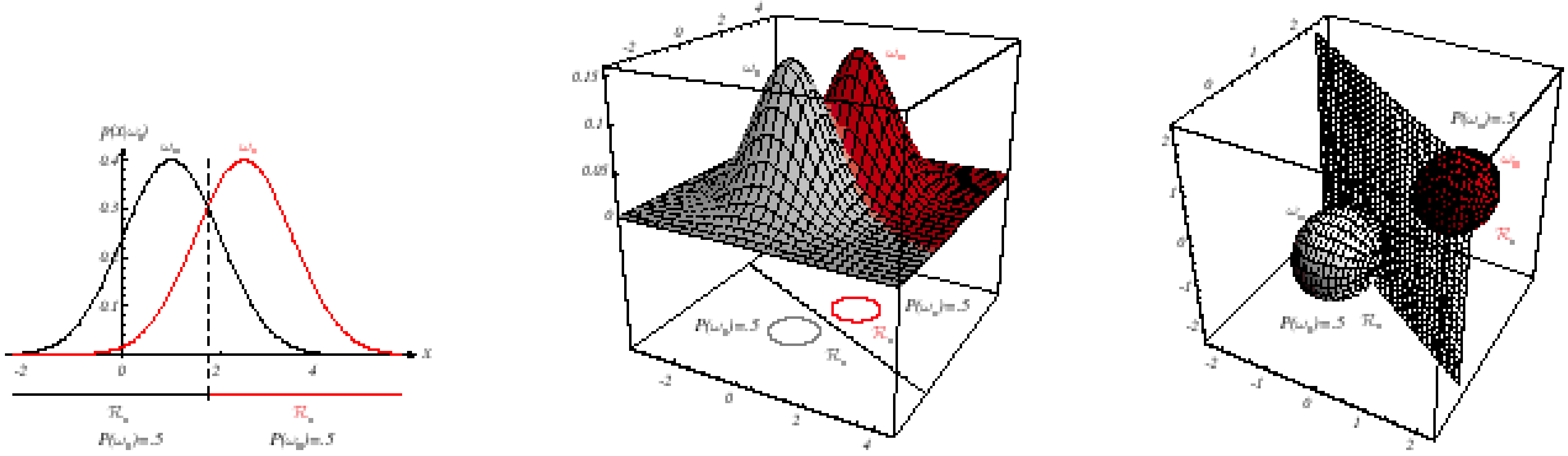


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

Discriminant Functions for Normal Density

- 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:

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

- In this case the hyperplane can be written as

$$\boldsymbol{w}^t(\boldsymbol{x} - \boldsymbol{x}_0) = 0$$

where $\boldsymbol{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, **and**

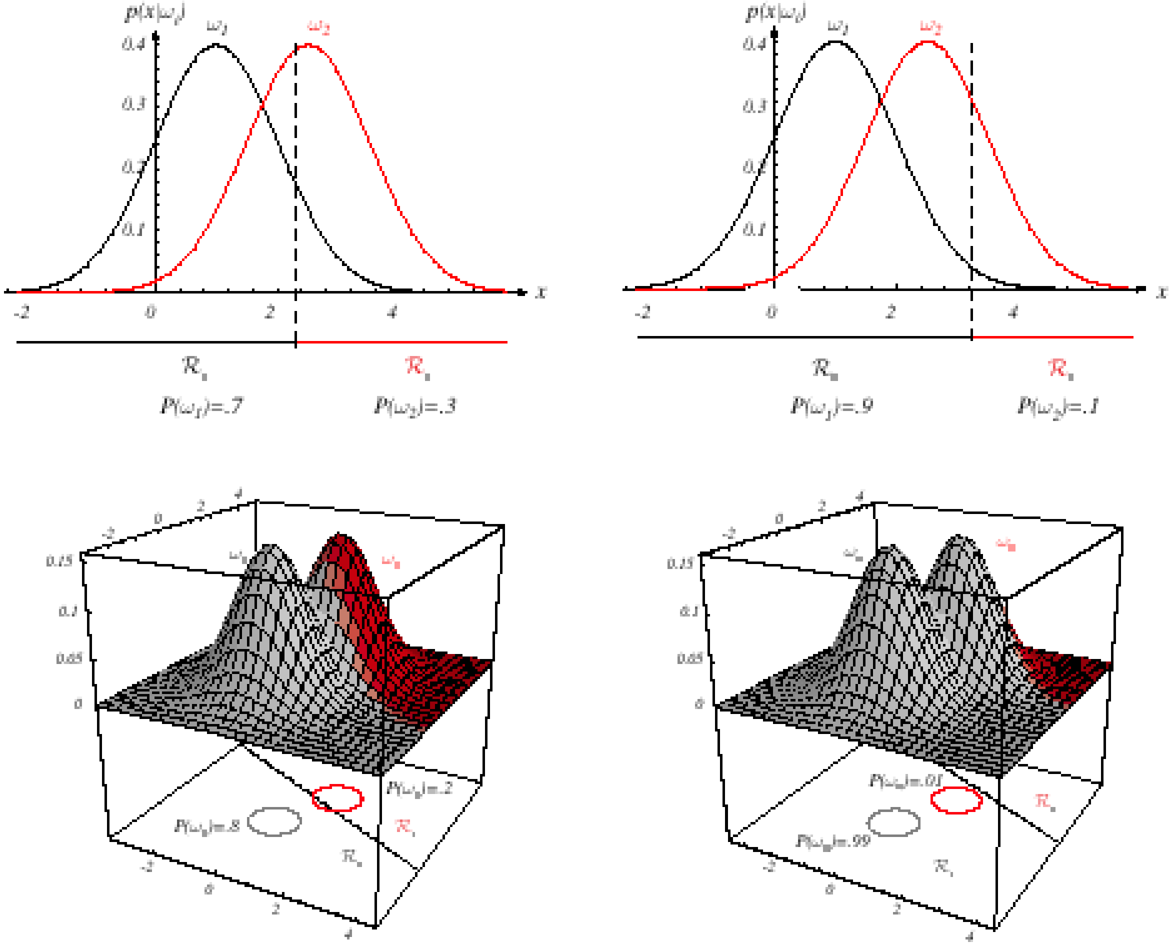
$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) - \frac{\boldsymbol{\sigma}^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \ln \frac{\boldsymbol{P}(\boldsymbol{\omega}_i)}{\boldsymbol{P}(\boldsymbol{\omega}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

Discriminant Functions for Normal Density

- The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ passing through the point $\boldsymbol{x_0}$ and orthogonal to the vector $\boldsymbol{w}$ (the line linking the mean vectors)

- **If** $\boldsymbol{P(\omega_i) = P(\omega_j)}$ **then** $\boldsymbol{x_0 = \frac{1}{2}(\mu_i + \mu_j)}$
  In such case, the optimum decision rule is based on Euclidean distance $||\boldsymbol{x - \mu_i}||$. Such classifier is called minimum distance classifier.

Discriminant Functions for Normal Density



Discriminant Functions for Normal Density

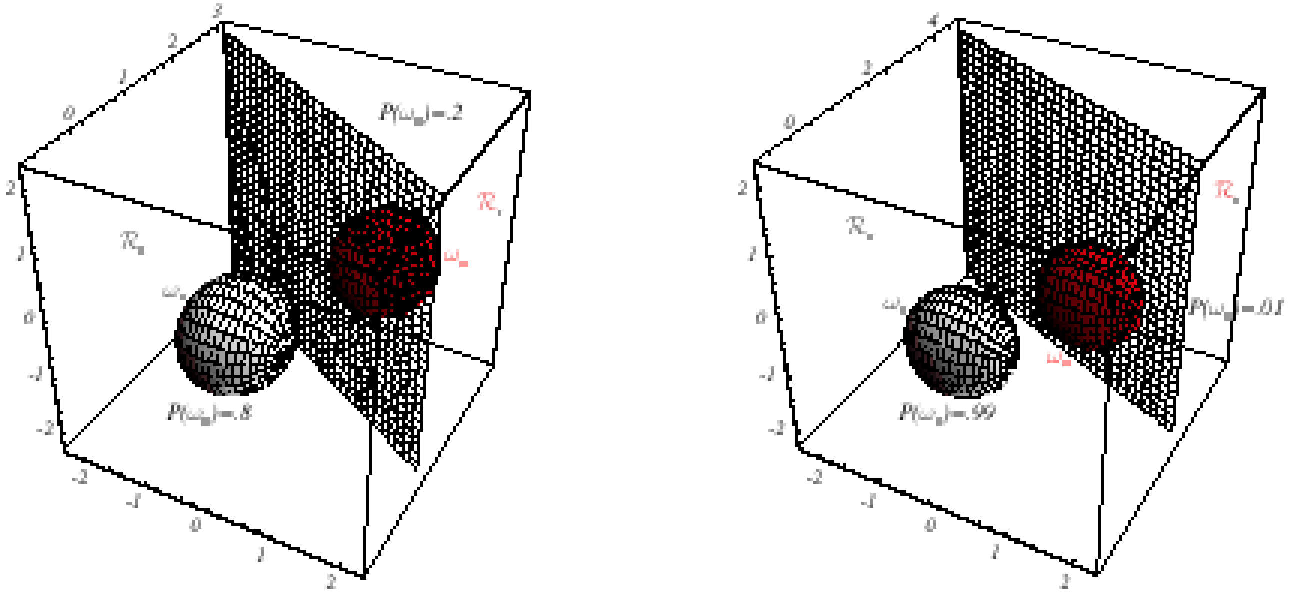


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Discriminant Functions for Normal Density

- Case 2 $\Sigma_i = \Sigma$
  - Covariance of all classes are identical but arbitrary

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

Because $x^t \Sigma^{-1} x$ is the same for all classes of x

and can be ignored, then we have

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

where:

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

Discriminant Functions for Normal Density

– Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ can be written as

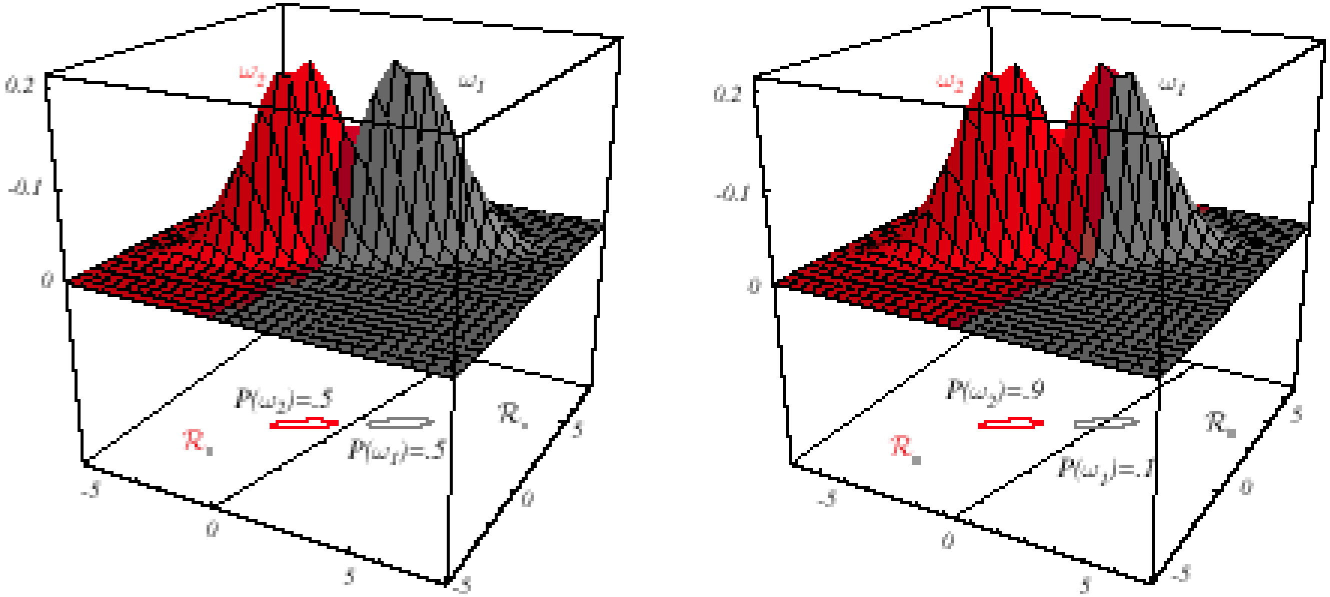
$$\boldsymbol{w}^t(\boldsymbol{x}-\boldsymbol{x}_0)=0$$

$$\textbf{where } \boldsymbol{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\textbf{, and}$$

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

(this hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is generally not orthogonal to the line linking the means)

Discriminant Functions for Normal Density



Discriminant Functions for Normal Density

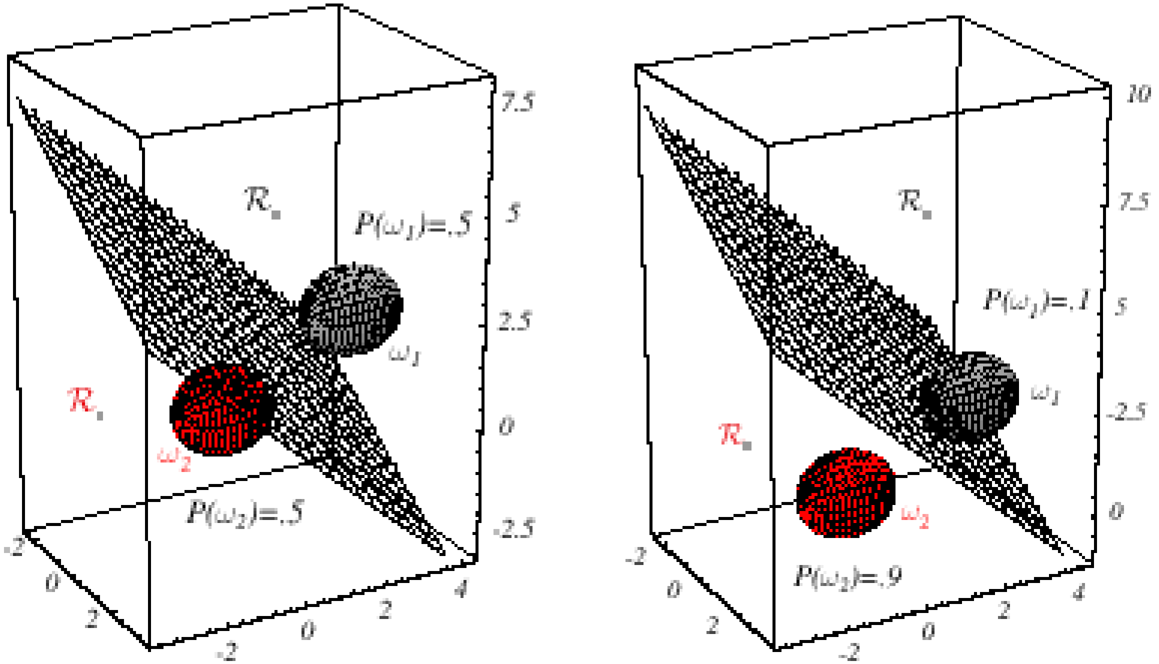


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Discriminant Functions for Normal Density

- Case 3 Σ_i is arbitrary
  - The covariance matrices are different for each category

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

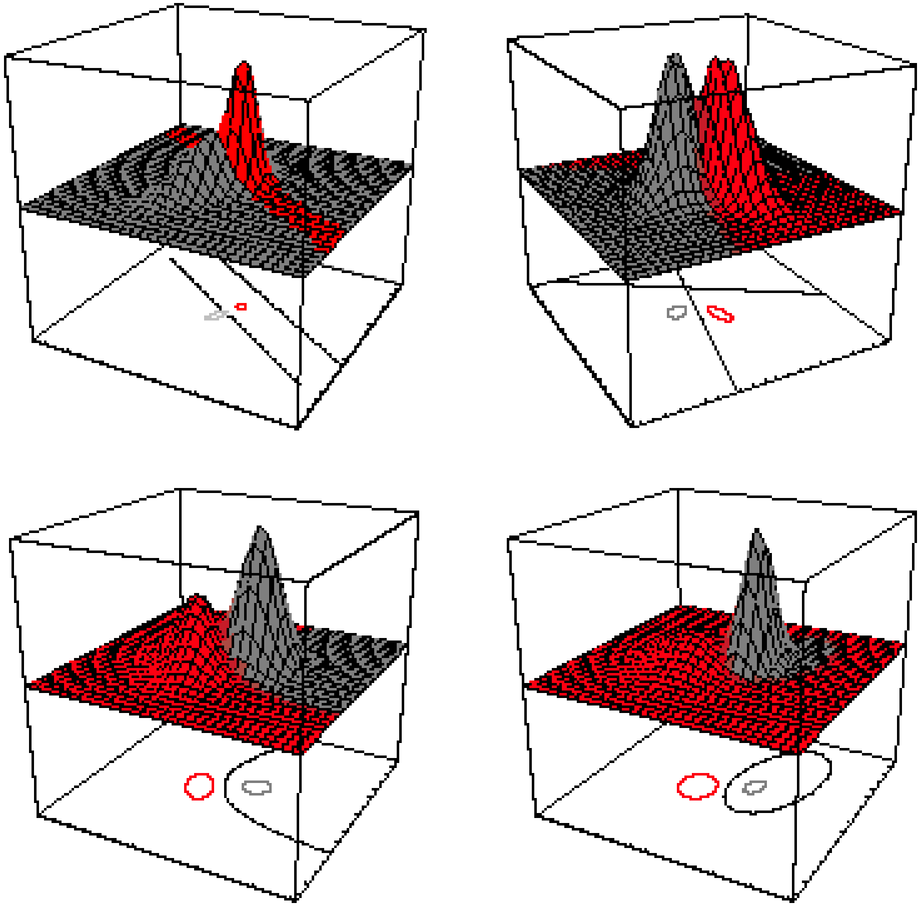
(a quadratic discriminant function) where:

$$W_i = -\frac{1}{2}\Sigma_i^{-1}, \quad w_i = \Sigma_i^{-1}\mu_i$$

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

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

Discriminant Functions for Normal Density



Discriminant Functions for Normal Density

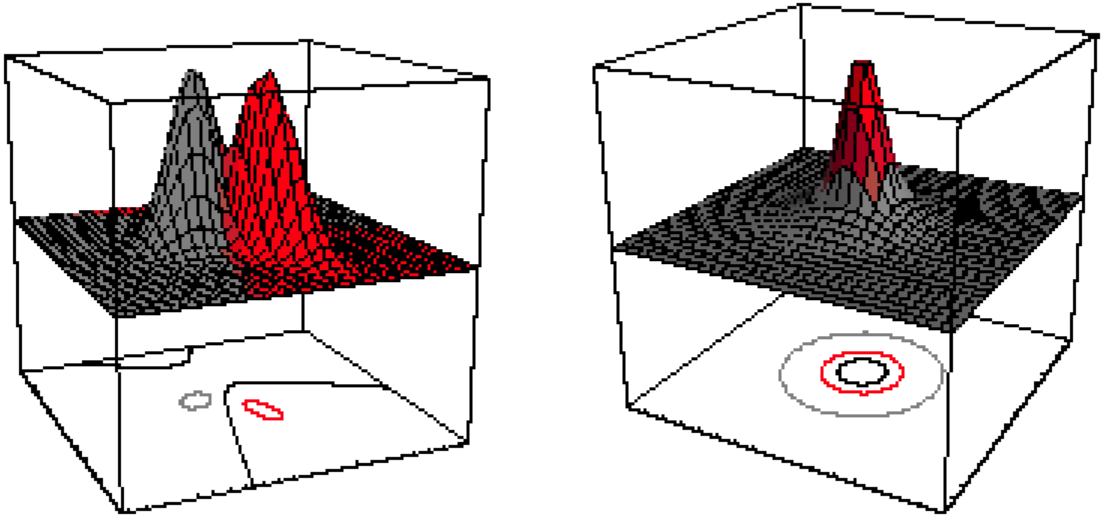


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Discriminant Functions for Normal Density

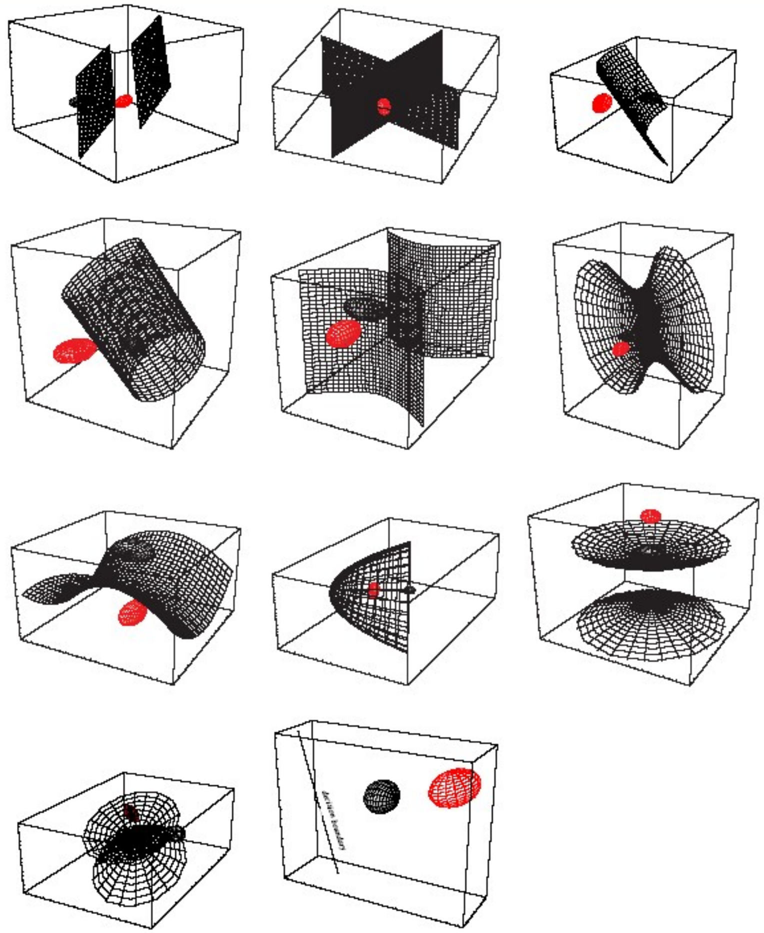
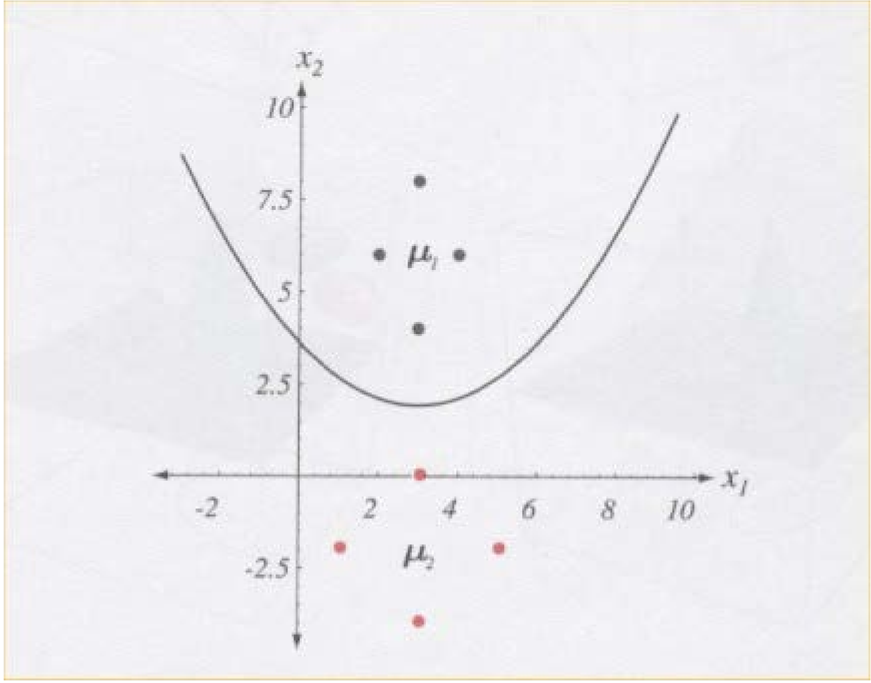


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright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

Example 2.1

- Let ω_1 be the set of four black points and ω_2 be the set of four red points



Example 2.1

let $\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix},$ **and** $\boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

we have $\boldsymbol{\Sigma}_1^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix}$ **and** $\boldsymbol{\Sigma}_2^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$

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

set $g_1(x) = g_2(x)$**, the decision boundary is**

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

Error Probabilities

- In a two-category case, the probability of error is

$$\begin{aligned} P(error) &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\ &= P(x \in R_2 \mid \omega_1) P(\omega_1) + P(x \in R_1 \mid \omega_2) P(\omega_2) \\ &= \int_{R_2} p(x \mid \omega_1) P(\omega_1) dx + \int_{R_1} p(x \mid \omega_2) P(\omega_2) dx \end{aligned}$$

Error Probabilities

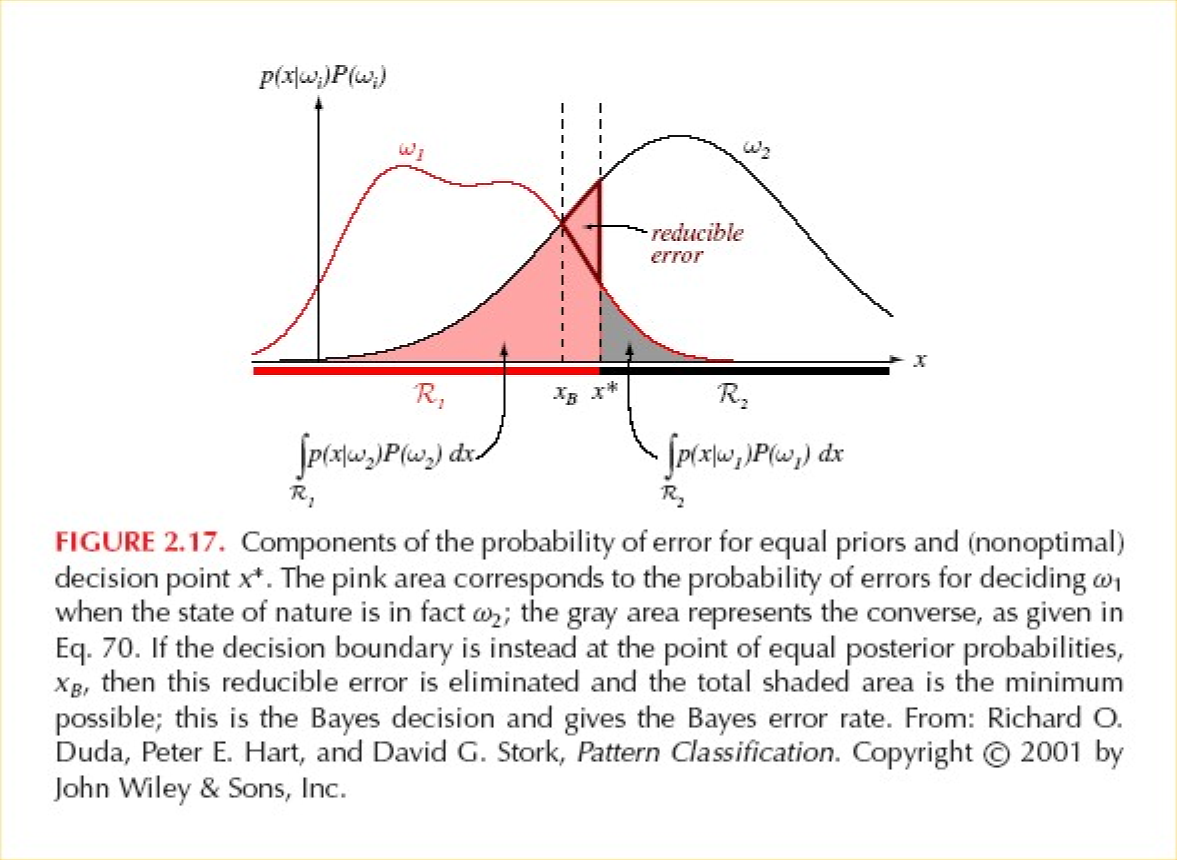


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary 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 Probabilities

- In a multi-category case, the probability of correction is

$$P(correct) = \sum_{i=1}^{c} P(x \in R_i, \omega_i)$$

$$= \sum_{i=1}^{c} P(x \in R_i \mid \omega_i) P(\omega_i)$$

$$= \sum_{i=1}^{c} \int_{R_i} p(x|\omega_i) P(\omega_i) dx$$

Error Bounds for Normal Densities

- The Bayes decision rule guarantees the lowest average error rate, but it does not tell the actual error.
- The calculation of the error rate can be complex, especially in high dimensions.
- Error bounds can be used to analytically approximate the upper limit of error rates.

Error Bounds

- Given the inequality

  $\min[a, b] \leq a^{\beta} b^{1-\beta}$ for $a, b \geq 0$ and $0 \leq \beta \leq 1$.

- Also we have

$$P(error) = \int_{-\infty}^{+\infty} P(error, x)dx = \int_{-\infty}^{+\infty} P(error \mid x) p(x)dx$$

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

- The error is bounded by

$$P(error) \leq P^{\beta}(\omega_1) P^{1-\beta}(\omega_2) \int p^{\beta}(x \mid \omega_1) p^{1-\beta}(x \mid \omega_2) dx$$

for $0 \leq \beta \leq 1$

  - This integral is over all feature space, regardless the decision boundary

Error Bounds

- If the conditional probabilities are normal

$$\int p^{\beta}(x|\omega_1)p^{1-\beta}(x|\omega_2)dx = e^{-k(\beta)}, \text{ where}$$

$$k(\beta)=\frac{\beta(1-\beta)}{2}(\mu_2-\mu_1)^{t}[\beta\Sigma_1+(1-\beta)\Sigma_2]^{-1}(\mu_2-\mu_1)$$

$$+\frac{1}{2}\ln\frac{|\beta\Sigma_1+(1-\beta)\Sigma_2|}{|\Sigma_1|^{\beta}|\Sigma_2|^{1-\beta}}$$

Error Bounds

- The Chernoff bound on $P(error)$ is found by analytically or numerically finding the value of β that minimize $e^{-k(\beta)}$ and then substituting this β into the inequality. The optimization here is on one-dimensional β space.
- Chernoff bound is loose for extreme values ($\beta \rightarrow 0$, and $\beta \rightarrow 1$). It is tight for intermediate values.
- By setting $\beta=1/2$, we have the Bhattacharyya bound. This is less tight but much simpler to compute.

Error Bounds

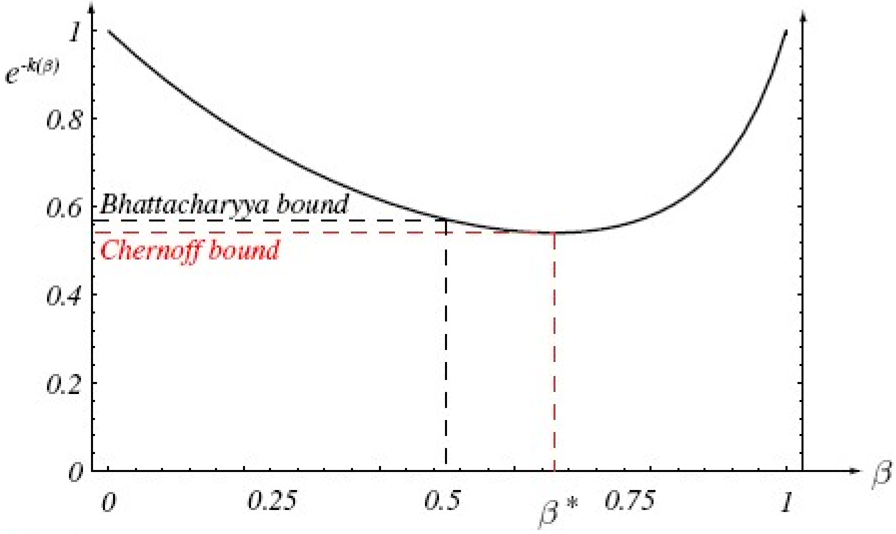


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Example 2.2

- Calculate the Bhattacharyya bound for the two class data in Example 2.1, we have

  $k(1/2)=4.116$, and

  $$P(error) \leq 0.0081910$$

- The Chernoff bound in this case is

  $$P(error) \leq 0.0081899$$

- The actual error calculated from numerical integration is 0.0021

Bayes Decision Theory – Discrete Features

- Discrete feature:
  - components of ***x*** are binary or integer valued, *x* can take only one of *m* discrete values $v_1, v_2, \ldots, v_m$
- Bayes formula involves probabilities instead of densities

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

- Bayes decision rule remains the same, to minimize the overall risk

$$\alpha^* = \arg\min_i R(\alpha_i \mid x)$$

Bayes Decision Theory – Discrete Features

- 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:

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

Here ***i*** is the index of the elements of the feature vector

Bayes Decision Theory – Discrete Features

$$P(x\,|\,\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i} \;\; \text{and} \;\; P(x\,|\,\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$$

the likelihood ratio is given by

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

therefore the discriminant function using log-likelihood is linear

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

Bayes Decision Theory – Discrete Features

- The discriminant function can be further written as:

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

where :

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

and :

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

$$\text{decide } \omega_1 \text{ if } g(x) > 0 \text{ and } \omega_2 \text{ if } g(x) \le 0$$

Example 2.3

- Consider a two class problem with three independent binary features

$$\boldsymbol{P}(\omega_1) = \boldsymbol{P}(\omega_2) = 0.5,$$

$$\boldsymbol{p_i} = 0.8,\ \boldsymbol{q_i} = 0.5\ \ \textbf{for}\ \ \boldsymbol{i} = 1, 2, 3\ \textbf{then}$$

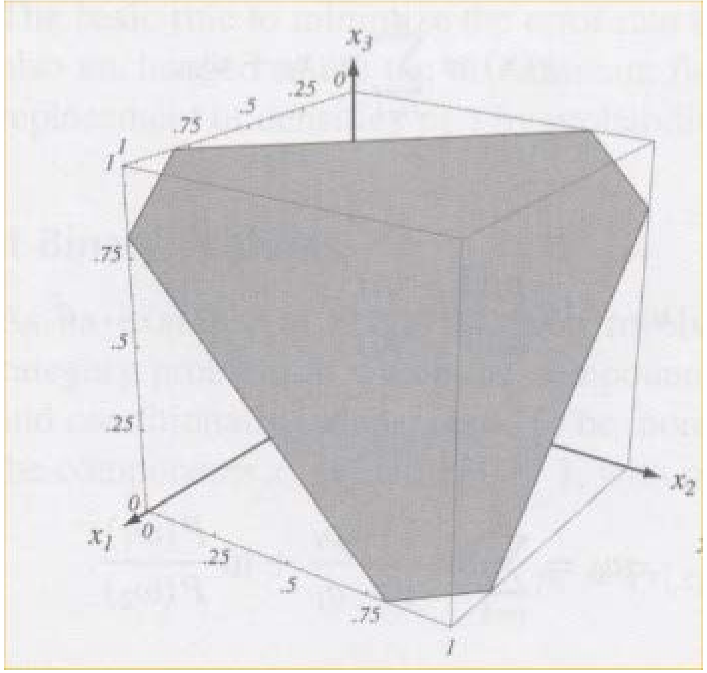
$$\boldsymbol{w_i} = \ln\frac{0.8(1-0.5)}{0.5(1-0.8)} = 1.3863$$

$$\boldsymbol{w}_0 = \sum_{i=1}^{3} \ln\frac{1-0.8}{1-0.5} + \ln\frac{0.5}{0.5} = -2.75$$

$$\boldsymbol{g}(\boldsymbol{x}) = \sum_{i=1}^{3} \boldsymbol{w}_i \boldsymbol{x}_i + \boldsymbol{w}_0,$$

$$\textbf{decide}\ \omega_1\ \textbf{if}\ \boldsymbol{g}(\boldsymbol{x}) > 0\ \textbf{and}\ \omega_2\ \textbf{if}\ \boldsymbol{g}(\boldsymbol{x}) \le 0$$

Example 2.3


x3
0
.25
.5
.75
1
1
.75
.5
.25
0
x1
0
.25
.5
.75
1
x2