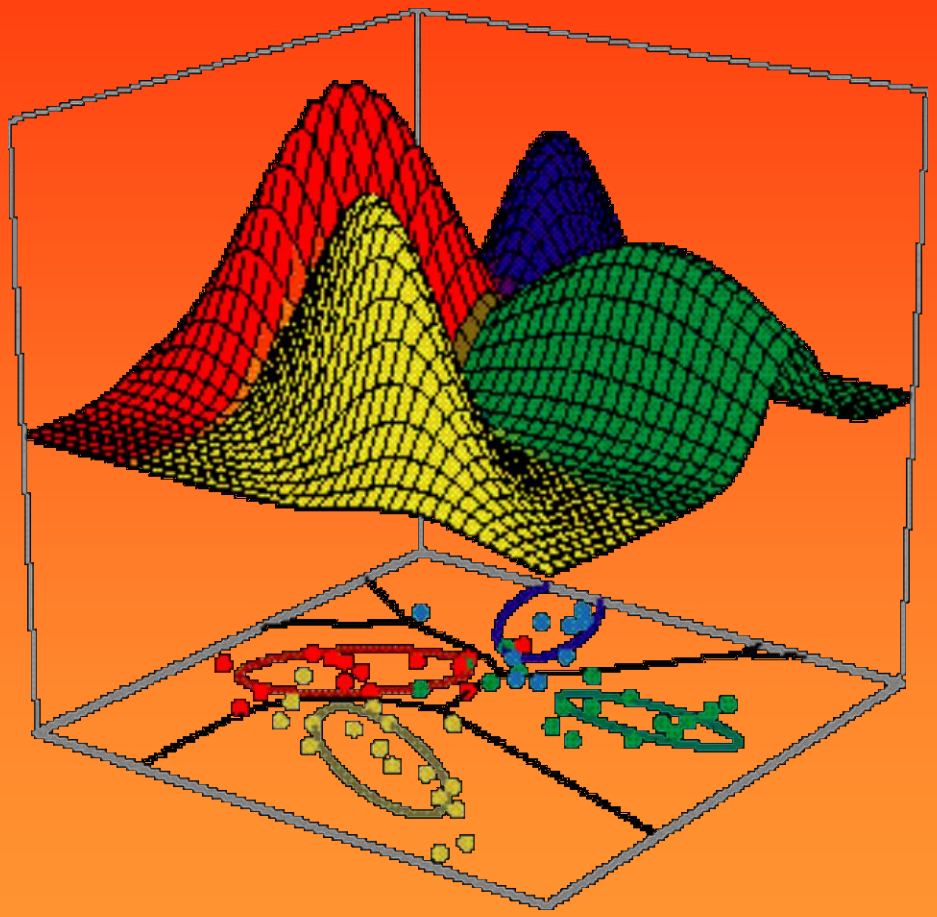


Pattern Classification

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

Capitulo 2
Teoria de Decision Bayesiana
(Sectiones 2-6,2-9)

Funciones discriminantes para la densidad normal

Errores y Cotas de Errores

Funciones discriminantes para la normal multivariada

- Vimos que la clasificación con tasa de error minima puede alcanzarse con la funcion de discriminación

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

- Caso normal multivariada

$$g_i(x) = \ln\left((2\pi)^{-d/2} \, |\Sigma|^{-1/2} \, \exp\left[-\frac{1}{2}(x-\mu_i)^t \Sigma^{-1} (x-\mu_i) \right] \right) + \ln\left(P(\omega_i)\right) =$$

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

Discriminantes en caso normal

- *Caso 1:* $\Sigma_i = \sigma^2.I$
  - Las variables X_i son i.i.d
  - $|\Sigma_i| = \sigma^{2d}$ y $\Sigma^{-1} = (1/\sigma^2)I$
  - decision se toma comparando discriminantes unos con otros, por lo cual los terminos no dependientes de la clase i-esima son constantes, y descartados de la formula

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

Caso 1: $\Sigma_i = \sigma^2.I$

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

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

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

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

Funcion discriminante lineal

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

$$= w_i^t x + w_{i0}$$

Funcion discriminante lineal, w_{io} i-esimo umbral

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

marcan el hiperplano que separa las regiones

Maquina lineal

- Un clasificador que usa funciones discriminantes lineales se llama " linear machine"
- Las superficies de decision de una maquina lineal son piezas de hiperplanos definidos por

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

para las dos categorias con mayor probabilidad a posteriori

- En el caso $\Sigma_i = \sigma^2.I$ esta ecuacion es

$$w^t (x-x_0)=0$$

donde $w=(\mu_i - \mu_j)$

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

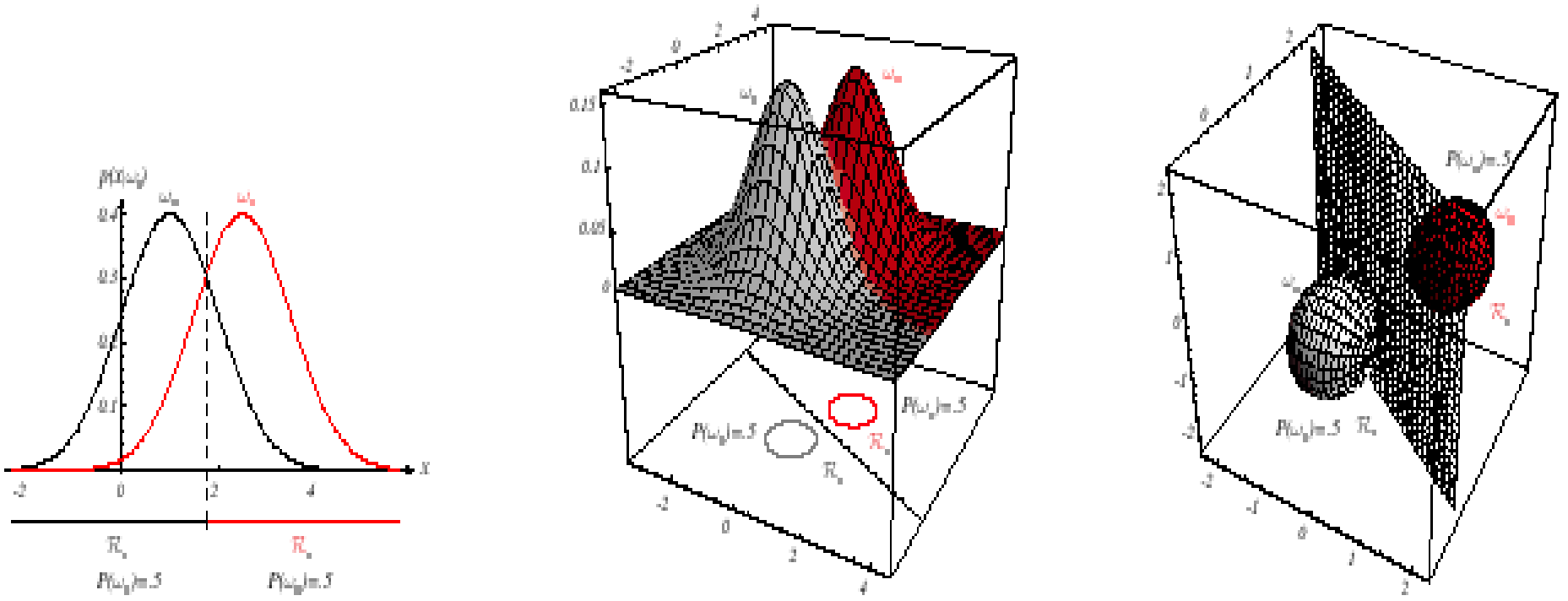


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

– El hiperplano que separa $\mathcal{R}_i$ y $\mathcal{R}_j$ *es ortogonal a*

w=(μ_i -μ_j) y pasa por

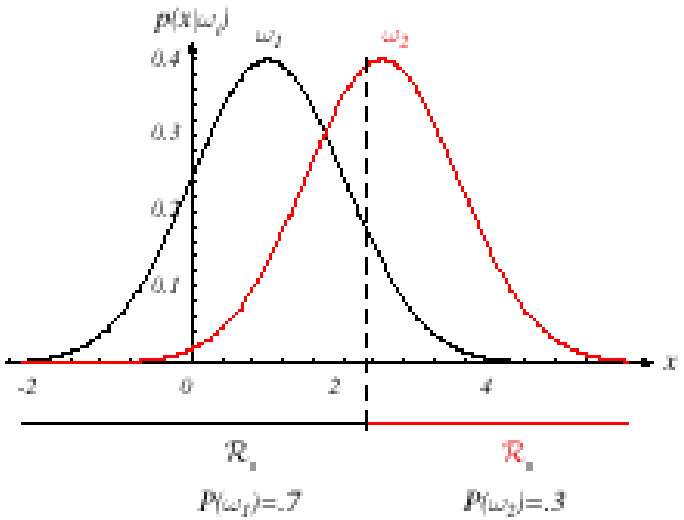
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

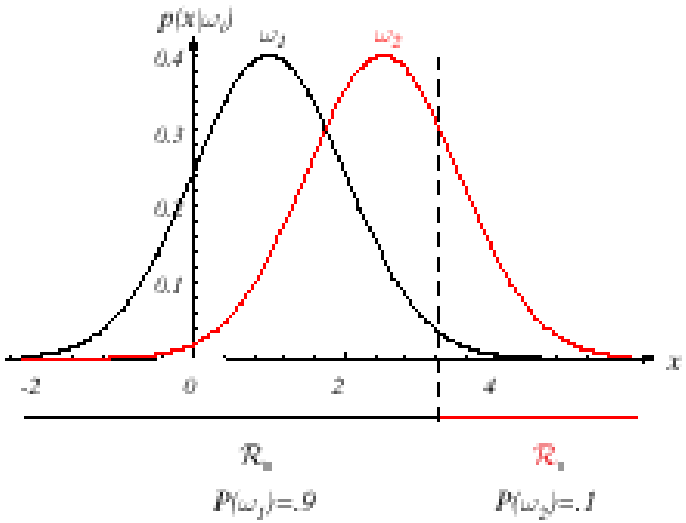
– Es siempre ortogonal a la recta que conecta las medias

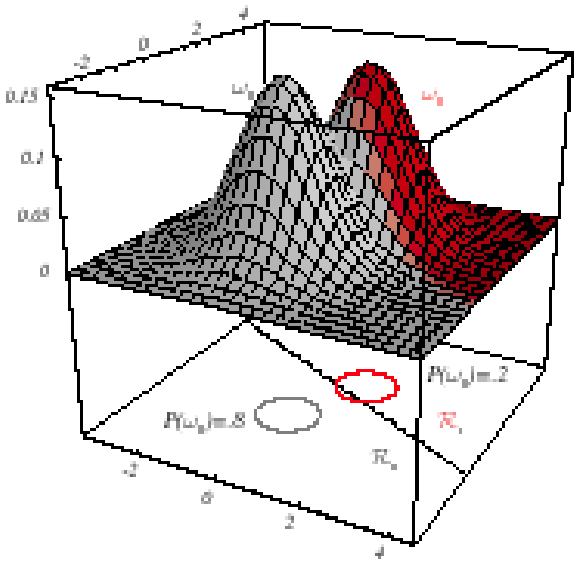
– Si $P(\omega_i) = P(\omega_j)$ entonces $x_0 = \frac{1}{2}(\mu_i + \mu_j)$

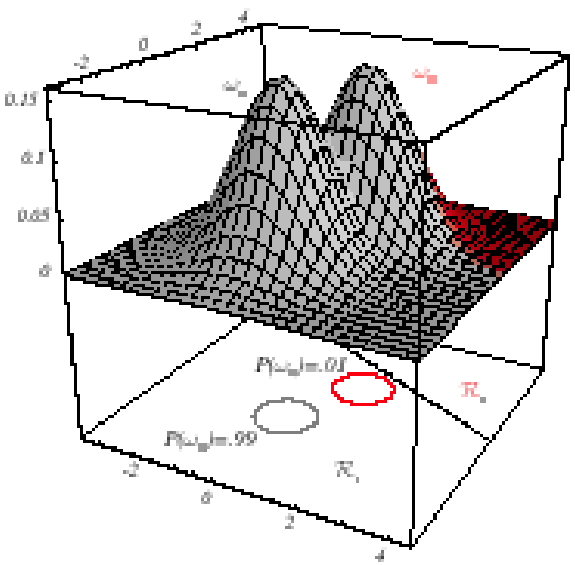
y el hiperplano pasa por la linea bisectriz entre las medias

- Si las probabilidades a priori $P(w_i)$ son las mismas para las c clases, entonces el termino de $\ln P(w_i)$ es otra constante sin importancia y puede ser ignorada.
- Entonces, la decisión optima clasifica a un vector x en la categoría de la media mas cercana
- Si $P(w_i)$ es bien distinto a $P(w_j)$ entonces el hiperplano que separa ya no se encuentra entre las medias.


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


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


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99

$$P(w_i) \neq P(w_j)$$

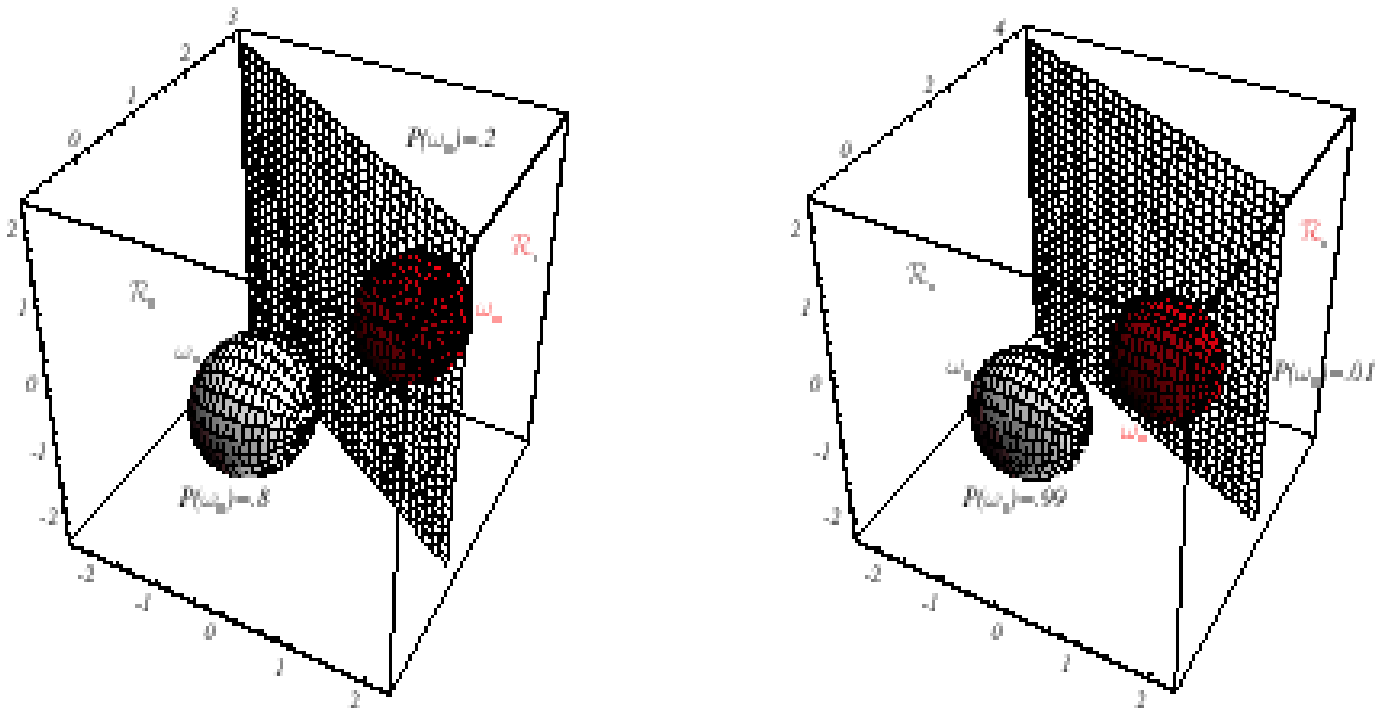


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Caso 2: $\Sigma_i = \Sigma$

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

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

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

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

$$= w_i^t x + w_{io}$$

Caso 2: $\Sigma_i = \Sigma$

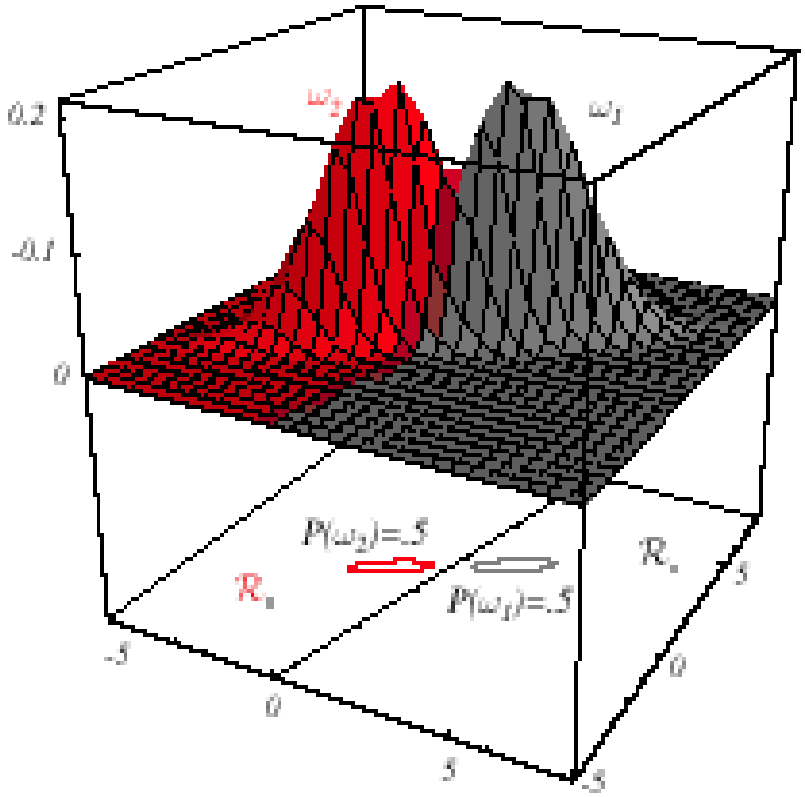
- Como los discriminantes tambien resultan lineales, si $\mathcal{R}_i$ y $\mathcal{R}_j$ son contiguas, el borde que separa tiene la ecuacion

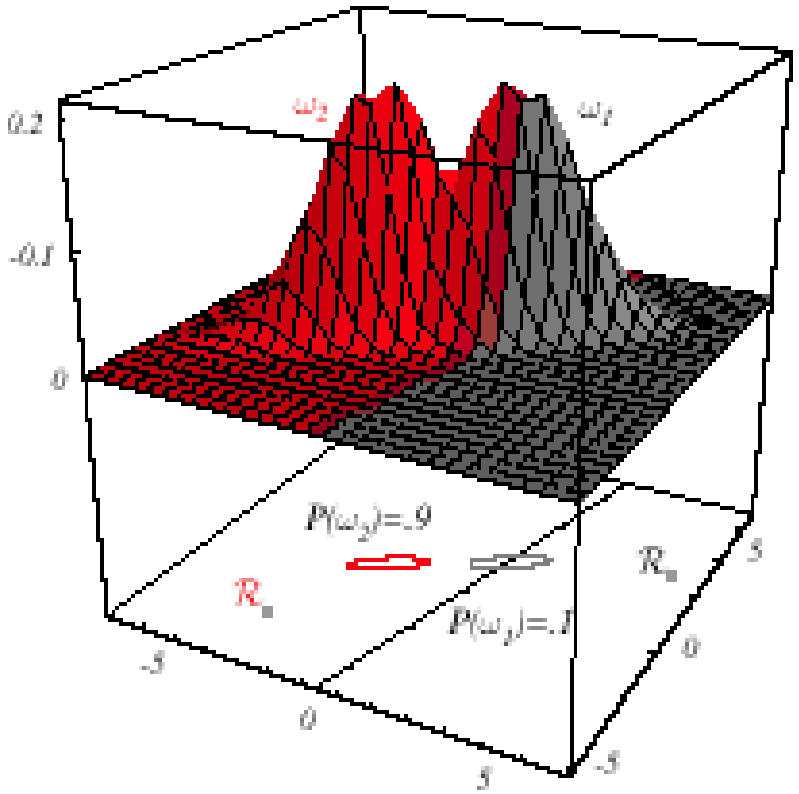
$$w^t(x - x_0) = 0$$

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

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

- El hiperplano que separa $\mathcal{R}_j$ generalmente no es ortogonal a la linea que une las medias


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

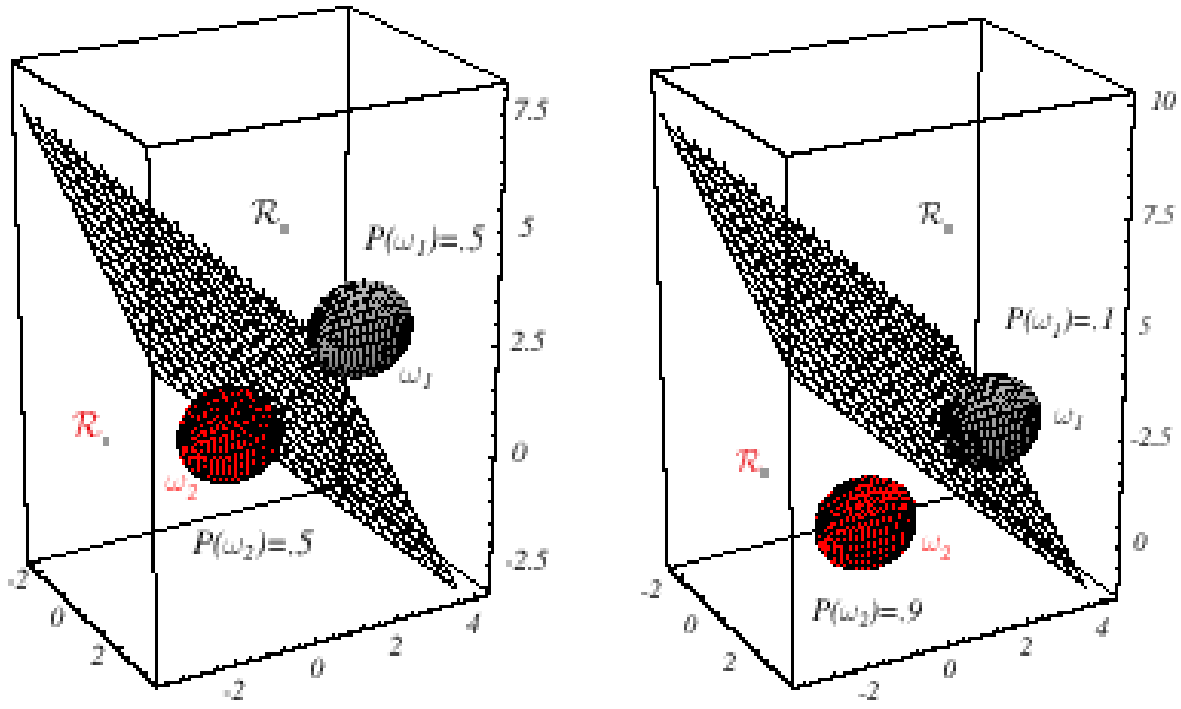


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Caso Σ_i arbitrario

- Las matrices de covarianza son todas diferentes, por lo cual solo puede cancelarse el termino (d/2)ln2π

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

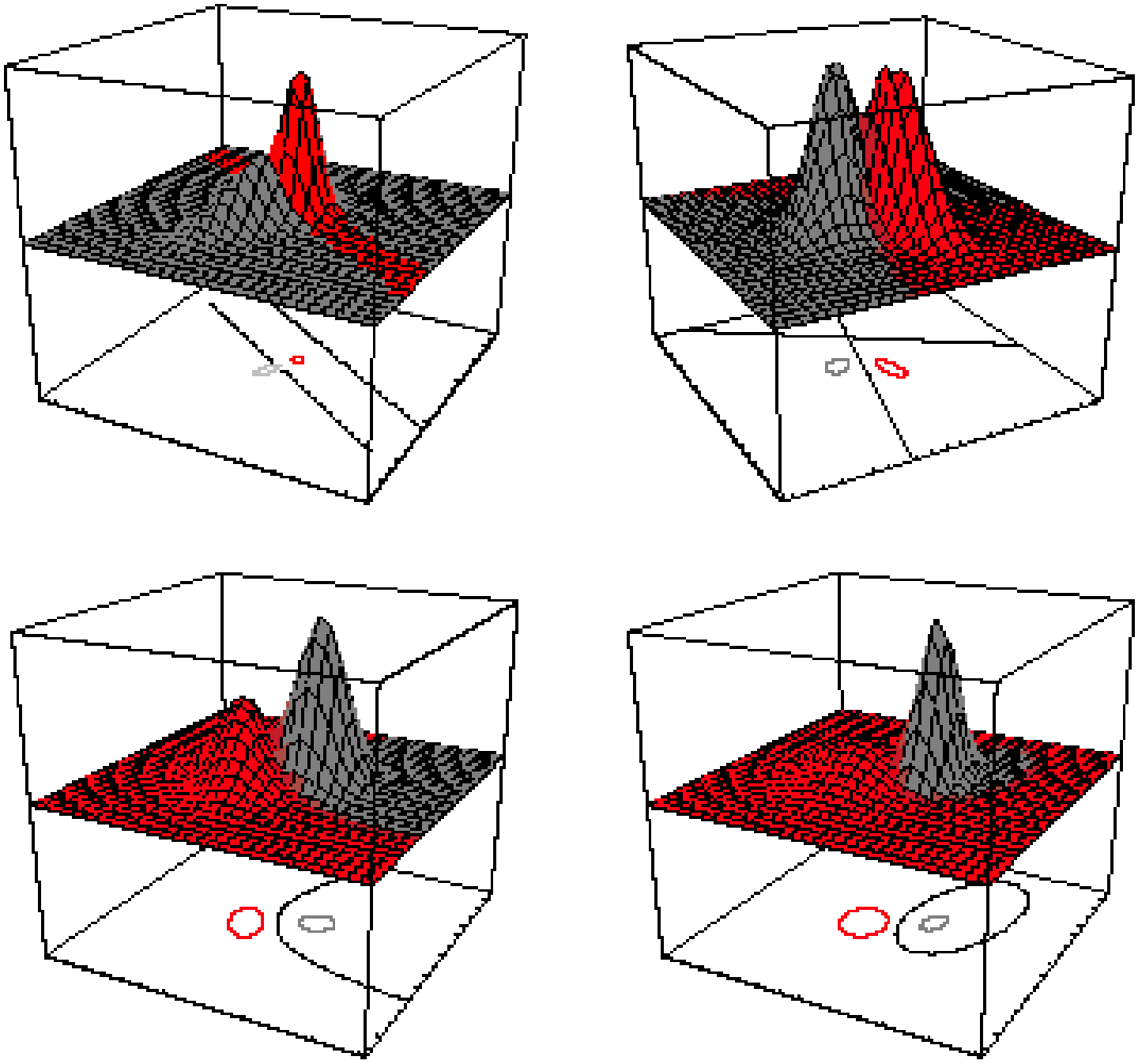
donde :

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

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

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

- Las superficies de decision son Hipercuadricas: hiperplanos, pares de hiperplanos, hiperesferas, hiperelipsoides, hiperparaboloides, hiperhiperboloides)



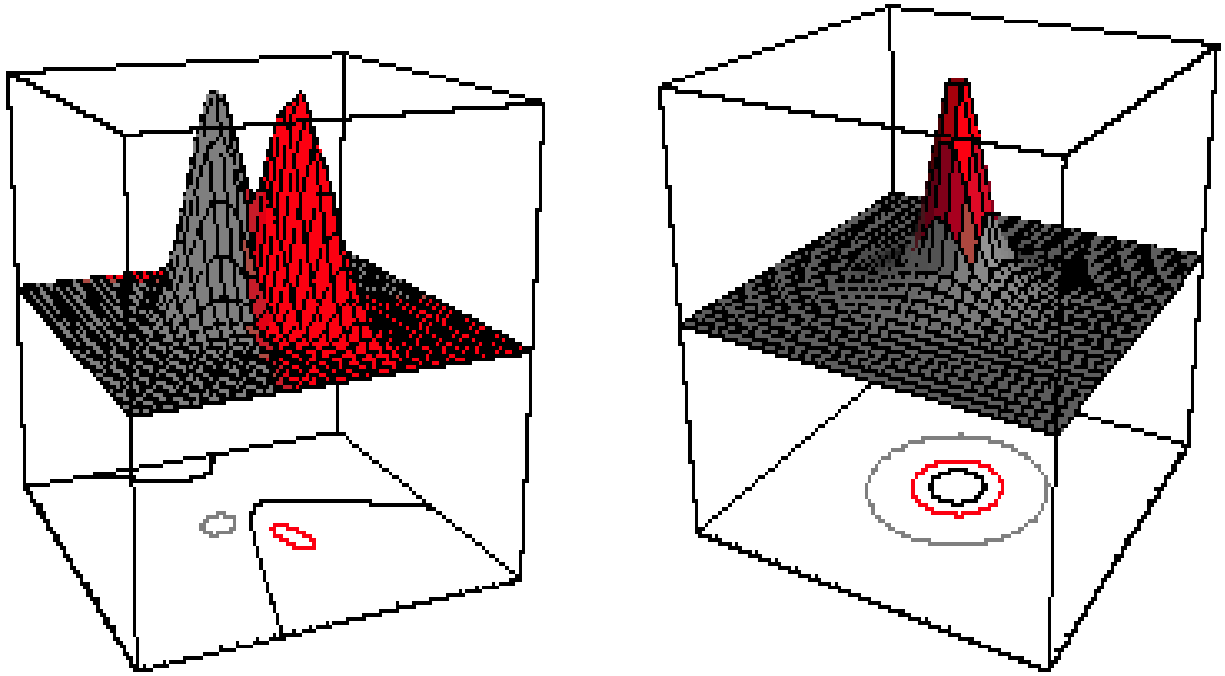


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Regiones de decision dos poblaciones Gaussianas: Ejercicio

- Supongamos concidos los parametros de las distribuciones

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \quad \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

- Iguales probabilidades a priori
- El borde de decision es

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

Errores

- Consideremos el clasificador dicotomico, donde hay solo dos estados naturales y dos regiones que caracterizan al clasificador.

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

Errores

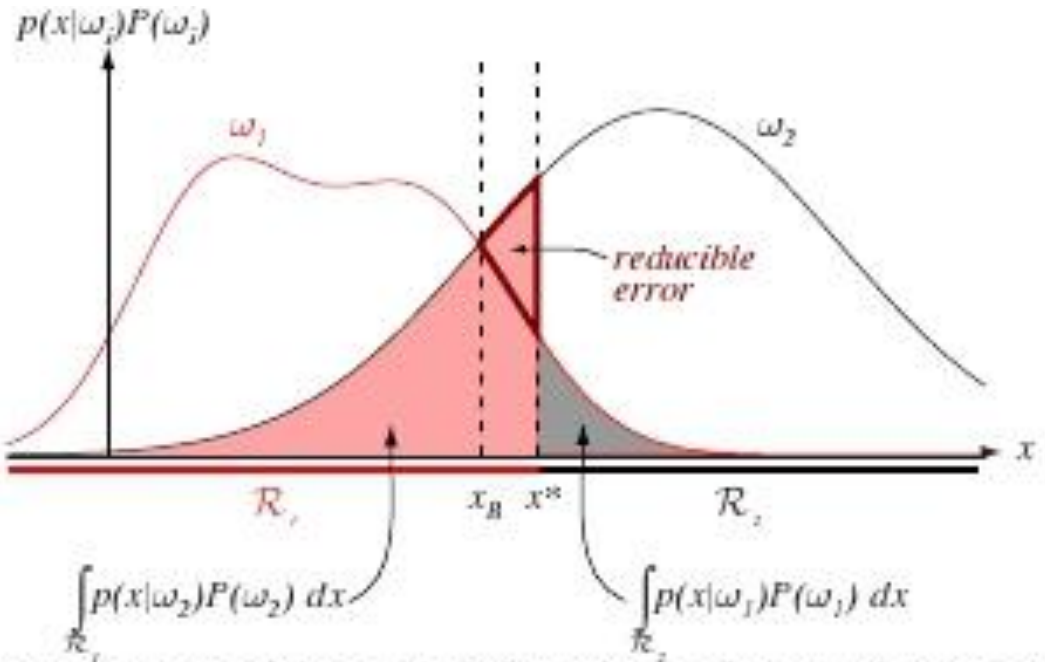


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary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

Errores

- En el caso de multicategoria, hay muchas mas formas de equivocarse que de acertar, por lo cual es mejor calcular la probabilidad de clasificar correctamente

$$P(correcto) = \sum_{j=1}^{c} P(x \in R_j, \omega_j)$$

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

$$= \sum_{j=1}^{c} \int_{R_j} p(x \mid \omega_j) P(\omega_j) dx$$

Cotas de error para densidades normales

- Desigualdad

$$\min[a,b] \le a^{\beta} b^{1-\beta} \quad 0 \le \beta \le 1 \quad a,b > 0$$

- Cota de Chernoff

$$P(error) = \int P(error \mid x)\, p(x)\, dx$$

$$P(error \mid x) = \min(P(w_1 \mid x), P(w_2 \mid x)) \le P(w_1 \mid x)^{\beta} P(w_2 \mid x)^{1-\beta}$$

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

- La integral es sobre todo el espacio, no necesitamos usar las particiones de la regla

Cota de Bhattacharyya β=1/2

$$P(error) \le \sqrt{P(\omega_1)P(\omega_2)} \int_R \sqrt{p(x|\omega_1)p(x|\omega_2)}dx = \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)}$$

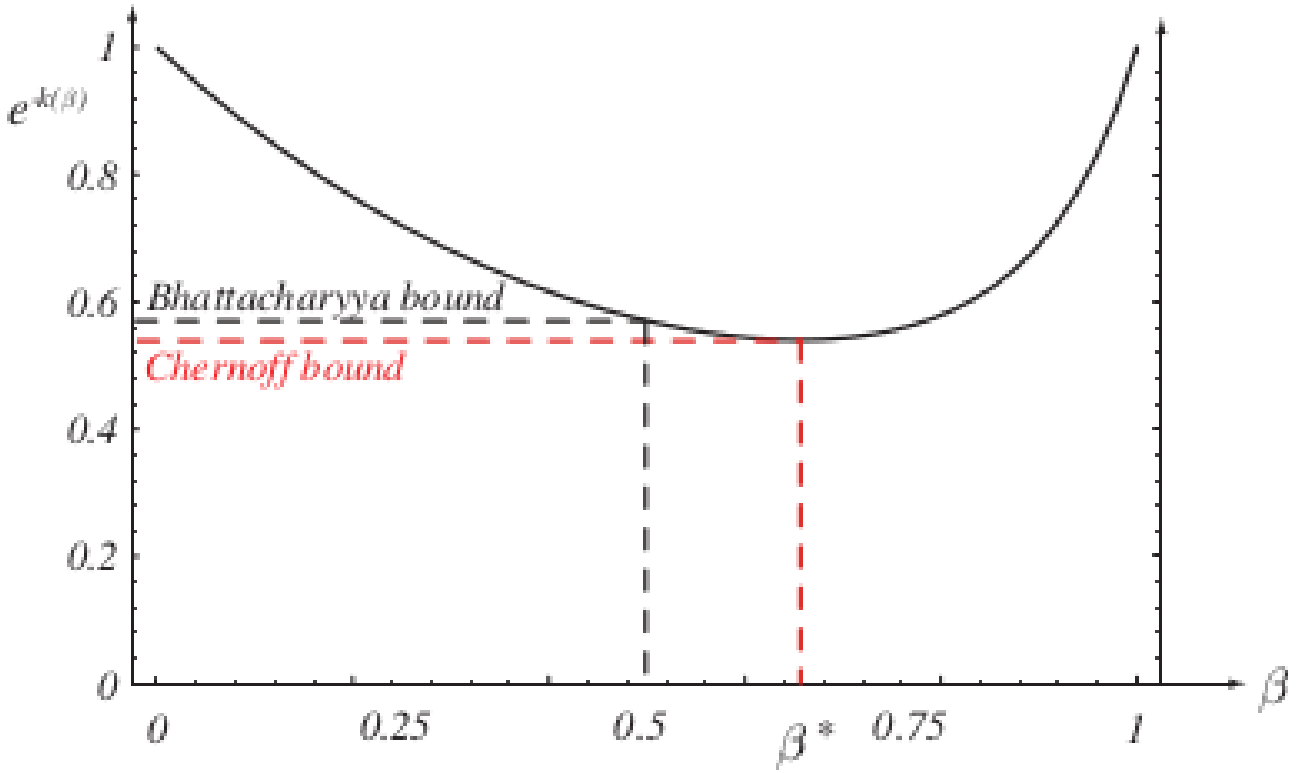


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Ejercicio:

- Calcular las cotas para las gaussianas del ejemplo 1.
- Calcular las cotas suponiendo distribuciones Gaussianas con los datos de Iris en matlab.

ROC curves

- Teoria de deteccion de señales y caracteristicas de operacion
  - Supongamos que queremos detectar un pulso como un parpadeo de luz o una señal de radar debil.
  - El modelo es:
    - Hay una señal interna en el detector que tiene media μ_1 cuando la señal externa no esta presente, y media μ_2 cuando esta presente.
    - Hay ruido blanco fuera de las señales, por lo cual se las modela como variables aleatorias.
    - Las distribuciones son normales con igual varianza

$$p(x \mid \omega_i) \sim N(\mu_i, \sigma^2)$$

Señal Detectada

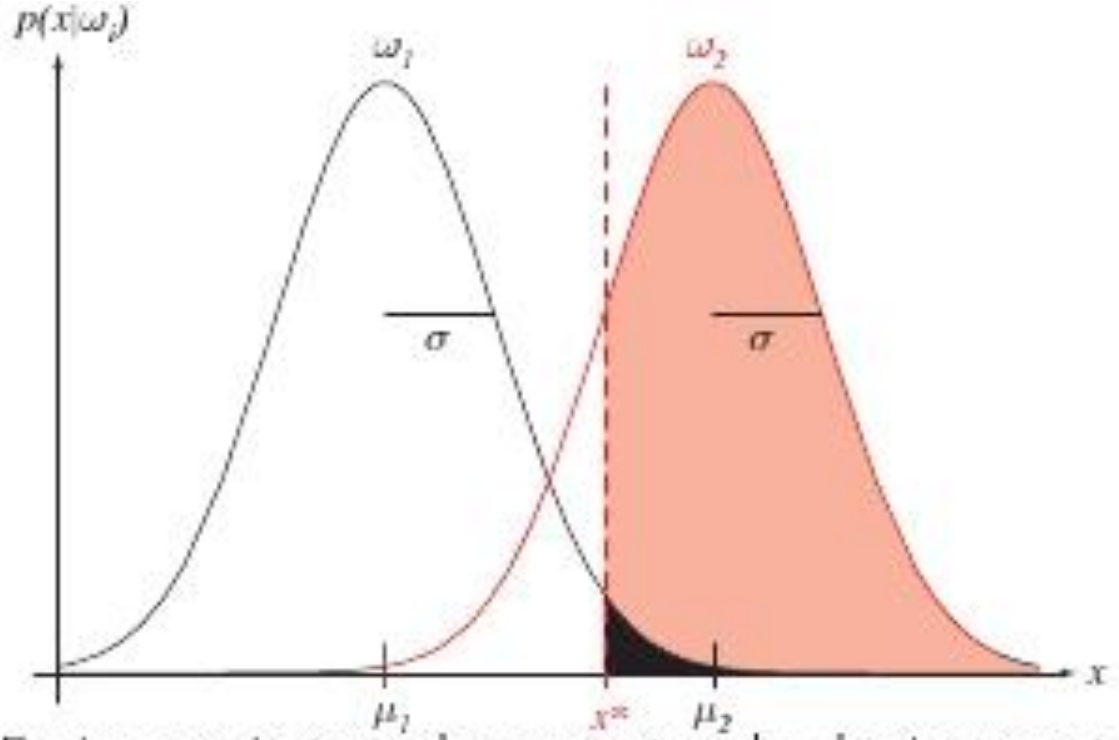


FIGURE 2.19. During any instant when no external pulse is present, the probability density for an internal signal is normal, that is, $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine the probability of a hit (the pink area under the ω_2 curve, above x^*) and of a false alarm (the black area under the ω_1 curve, above x^*). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminabilidad

- El grafico nos dice que el problema de encontrar el umbral de deteccion x* es inherente a cuan distantes estan las normales
- Esto se puede medir con

$$d' = \frac{|\mu_1 - \mu_2|}{\sigma}$$

- Si d'es grande, las normales estan separadas y la discriminabilidad es grande
- Si tenemos acceso a una muestra de entrenamiento podemos estimar d' , μ_1, y μ_2

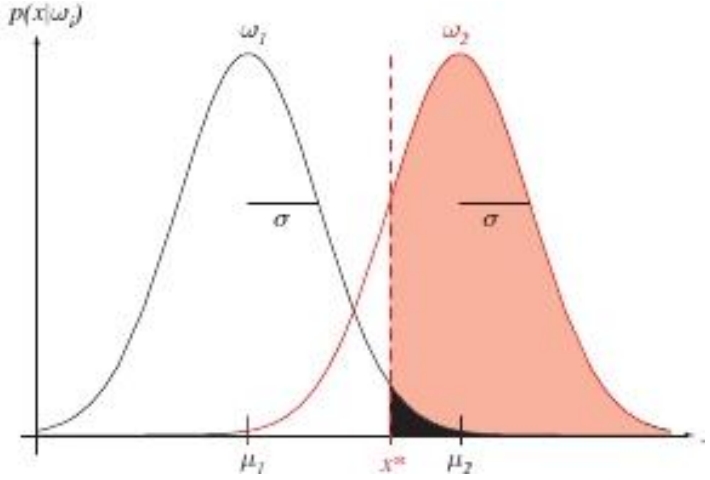
Notacion

- $P(x>x^* | x \epsilon \omega_2)$ es un acierto, la probabilidad de que la señal interna esta por arriba del valor x* dado que la señal esta presente.
- $P(x>x^* | x \epsilon \omega_1)$ es una falsa alarma, la probabilidad de que la señal interna esta por arriba del valor x* a pesar de que no hay señal externa presente.
- $P(x<x^* | x \epsilon \omega_2)$ es un falso rechazo, la probabilidad de que la señal interna esta por debajo del valor x* dado que la señal esta presente.
- $P(x<x^* | x \epsilon \omega_1)$ es un rechazo correcto, la probabilidad de que la señal interna esta por debajo de x* dado que la señal externa no esta presente

ROC curve

- Si tenemos un gran numero de ensayos (con x* fijo, aunque desconocido), podemos determinar estas probabilidades experimentalmente, en particular las tasas de aciertos y falsas alarmas
- Podemos realizar un grafico representando estas tasas Si las densidades están fijas pero cambia el umbral x*, entonces nuestras tasas de aciertos y falsas alarmas cambian también.

ROC curve



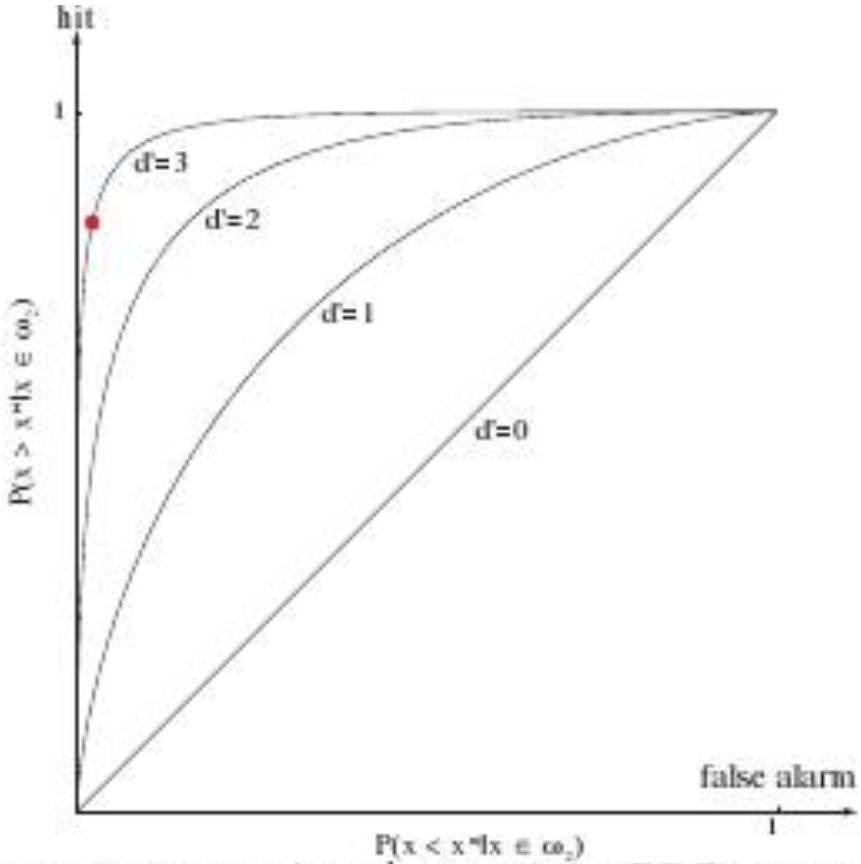


FIGURE 2.20. In a receiver operating characteristic (ROC) curve, the abscissa is the probability of false alarm, $P(x > x^*|x \in \omega_1)$, and the ordinate is the probability of hit, $P(x > x^*|x \in \omega_2)$. From the measured hit and false alarm rates (here corresponding to x^* in Fig. 2.19 and shown as the red dot), we can deduce that $d' = 3$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Beneficios

- El gran beneficio del marco de detección de señales es que podemos distinguir operacionalmente entre discriminabilidad y sesgo de decisión. El primero es inherente al problema, mientras que el segundo depende de la regla, o la función de perdida usada.
- Por cada par de tasas de aciertos y falsa alarma pasa una sola curva ROC, y puede deducirse de ese punto la discriminabilidad.
- Mas aun, en el caso Gaussiano, se puede calcular el error de Bayes.
- Si el error actual difiere de la tasa de Bayes inferida, se puede mover el x*.

Caracteristicas discretas

- Componentes de X son binarias o a valores enteros, X puede tomar uno de m valores discretos $v_1, v_2, \ldots, v_m$ por lo cual densidades continuas se transforman en densidades puntuales, e integrales en sumas.
- La definicion de riesgo condicional no se cambia y la regla de decision de Bayes resulta la misma, seleccionar lla accion tal que el riesgo sea minimo.
- La regla basica de minimizar la tasa de error maximizando la probabilidad a posteriori se mantiene asi, como las funciones discriminantes, reemplazando las densidad es por probabilidades puntuales.

Caracteristicas binarias independientes, dos categorias

- Sea $x = [x_1, x_2, \ldots, x_d]^t$ donde cada x_i es 0 o 1, con probabilidades:

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

- Asumiendo independencia condicional se puede escribir

$$P(x \mid w_1) = \prod_{i=1}^{d} p_i^{x_i} (1 - p_i)^{1-x_i} \qquad P(x \mid w_2) = \prod_{i=1}^{d} q_i^{x_i} (1 - q_i)^{1-x_i}$$

$$\frac{P(x \mid w_1)}{P(x \mid w_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i} \right)^{x_i} \left(\frac{1 - p_i}{1 - q_i} \right)^{1-x_i}$$

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

- La funcion discriminante es lineal en x_i :

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

$$donde:$$

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

$$y:$$

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

$$decide\ \omega_1 \text{ si } g(x) > 0 \text{ y } \omega_2 \text{ si } g(x) \leq 0$$

Estudiar casos especiales

- Caracteristicas no muestradas
- Caracteristicas ruidosas
- Hacer ejercicios 49 y 47 despues de deducir las reglas.