Capitulo 2
Teoría de Decisión Bayesiana
(Secciones 2-6,2-9)

FUNCIONES DISCRIMINANTES PARA LA DENSIDAD NORMAL

ERRORES Y COTAS DE ERRORES

Regla discriminante lineal de Fisher

- Sea la variable $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ y dos estados naturales w_1 w_2.
- Supongamos conocer

$$E_{w_1}(X) = \mu_1 \quad E_{w_2}(X) = \mu_2 \quad V_{w_1}(X) = V_{w_2}(X) = \sum.$$

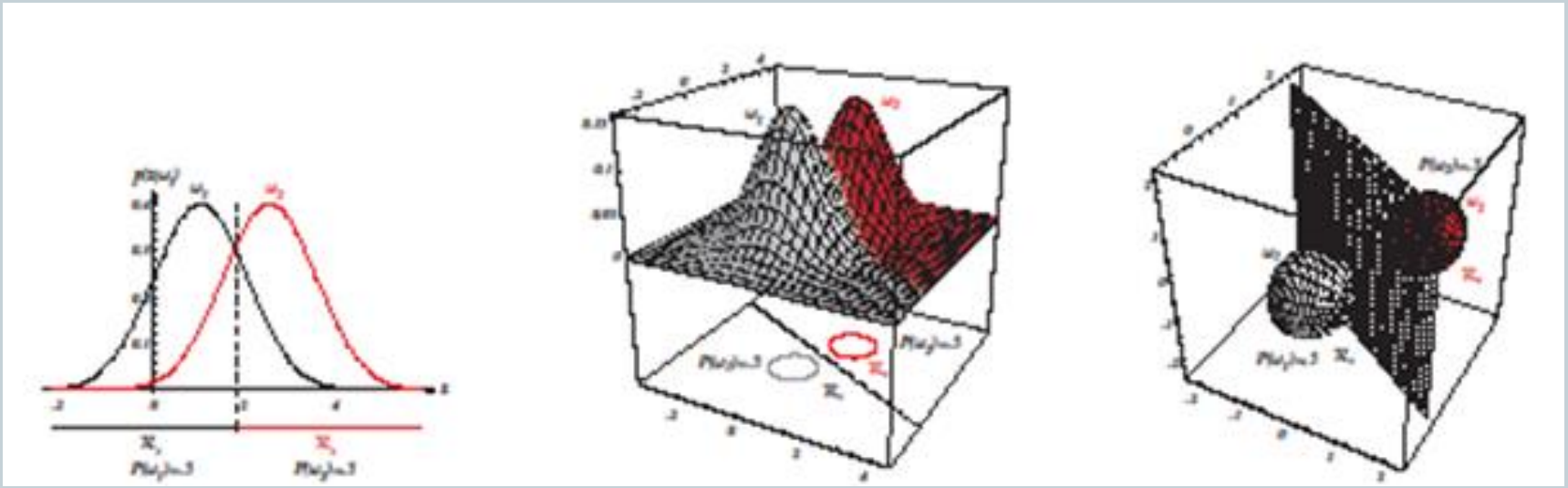
- Fisher estudió el problema de definir la combinación lineal de la forma

$$Y = l'X = l_1X_1 + l_2X_2 + \cdots + l_pX_p$$

que sea óptima para clasificar una observación en alguna de las dos poblaciones.

Regla discriminante lineal de Fisher

Hay que buscar *l* que optimice la separación entre las dos poblaciones, esto puede lograrse maximizando la separación entre las medias:



Regla discriminante lineal de Fisher

- Definiendo $Y = l'X = l_1 X_1 + l_2 X_2 + \cdots + l_p X_p$

$$E_{w_1}(Y) = E_{w_1}(l'X) = l'\mu_1 = \mu_{Y1}$$
$$E_{w_2}(Y) = E_{w_2}(l'X) = l'\mu_2 = \mu_{Y2}$$
$$V_{w_1}(Y) = V_{w_1}(l'X) = l'\sum l = \sigma_Y^2 = V_{\pi_2}(l'X) = V_{\pi_2}(Y)$$

- La regla de Fisher busca l tal que

$$\max_{l \in \Re^p} (\mu_{Y1} - \mu_{Y2})^2 = \max_{l \in \Re^p} (l'\mu_1 - l'\mu_2)^2$$

Regla discriminante lineal de Fisher

Problema: Si se maximiza sin restricciones, el máximo puede no ser finito: se maximiza dividiendo por la varianza común σ_Y^2

$$\max_{l\in\Re^p} \frac{(\mu_{Y1}-\mu_{Y2})^2}{\sigma_Y^2} = \max_{l\in\Re^p} \frac{(l'\mu_1 - l'\mu_2)^2}{l'\Sigma l}$$

$$\max_{l\in\Re^p} \frac{l'(\mu_1-\mu_2)'(\mu_1-\mu_2)l}{l'\Sigma l}$$

La solución que se obtiene es:

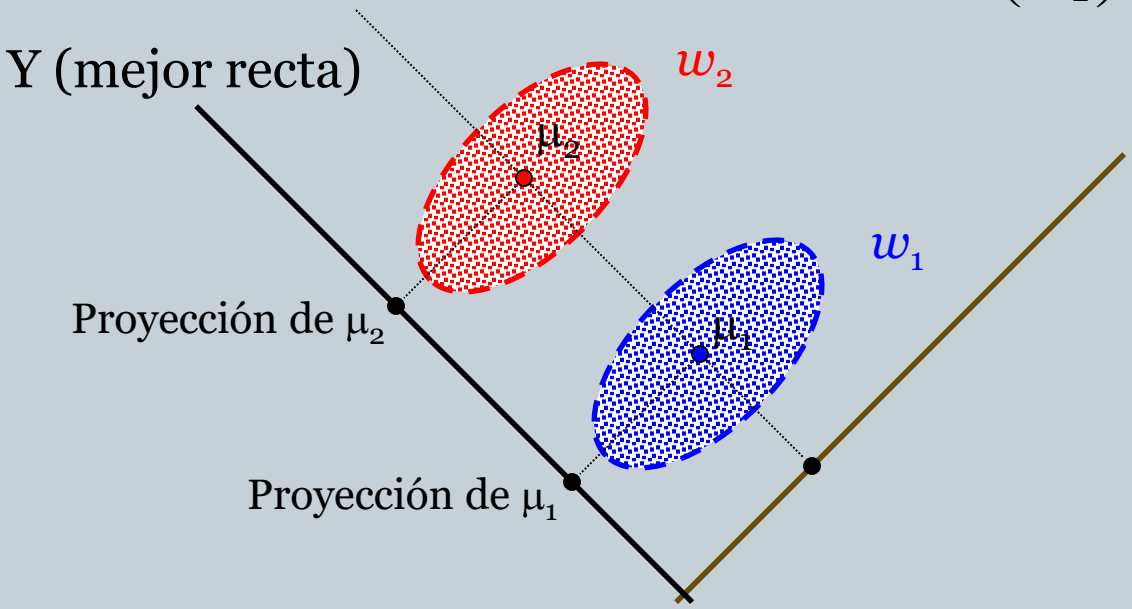
$$l = \Sigma^{-1}(\mu_1-\mu_2)$$

$$Y = (\mu_1-\mu_2)'\Sigma^{-1}X$$

Función discriminante lineal de Fisher

Regla discriminante lineal de Fisher

- En el caso de dos coordenadas $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ $l = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix}$

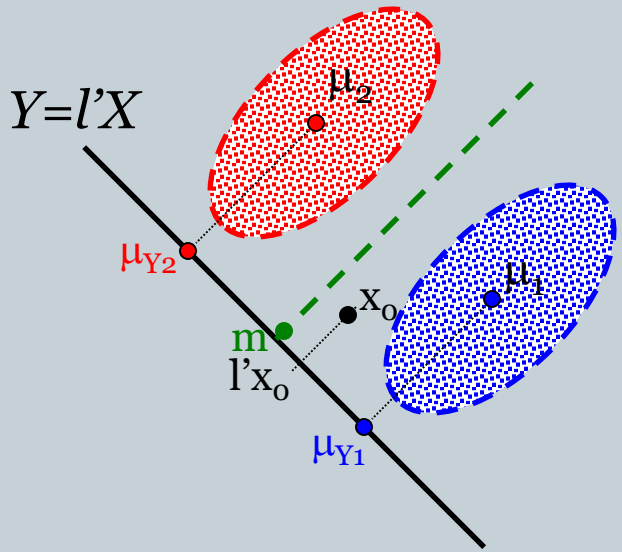


$$Y = l'X = l_1X_1 + l_2X_2$$

l_1 y l_2 determinan la recta

Regla discriminante lineal de Fisher

El *punto medio* es: $m = \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$



Dada una nueva observación x_o:

- Asignar x_o a π_1 si

$$(\mu_1 - \mu_2)'\Sigma^{-1}x_0 - m \geq 0$$

- Asignar x_o a π_2 si

$$(\mu_1 - \mu_2)'\Sigma^{-1}x_0 - m < 0$$

Regla discriminante lineal de Fisher

- Dadas dos estados naturales w_1 y w_2, se tienen las siguientes matrices de datos:

$$X^{(1)} = \begin{pmatrix} X_{11}^{(1)} & X_{12}^{(1)} & \cdots & X_{1p}^{(1)} \\ X_{21}^{(1)} & X_{22}^{(1)} & \cdots & X_{2p}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_1 1}^{(1)} & X_{n_1 2}^{(1)} & \cdots & X_{n_1 p}^{(1)} \end{pmatrix} \quad X^{(2)} = \begin{pmatrix} X_{11}^{(2)} & X_{12}^{(2)} & \cdots & X_{1p}^{(2)} \\ X_{21}^{(2)} & X_{22}^{(2)} & \cdots & X_{2p}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_2 1}^{(2)} & X_{n_2 2}^{(2)} & \cdots & X_{n_2 p}^{(2)} \end{pmatrix}$$

- Las medias y varianza estimadas son

$$\overline{X}_1, \overline{X}_2, \quad S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}.$$

Regla discriminante lineal de Fisher

- La regla lineal estimada

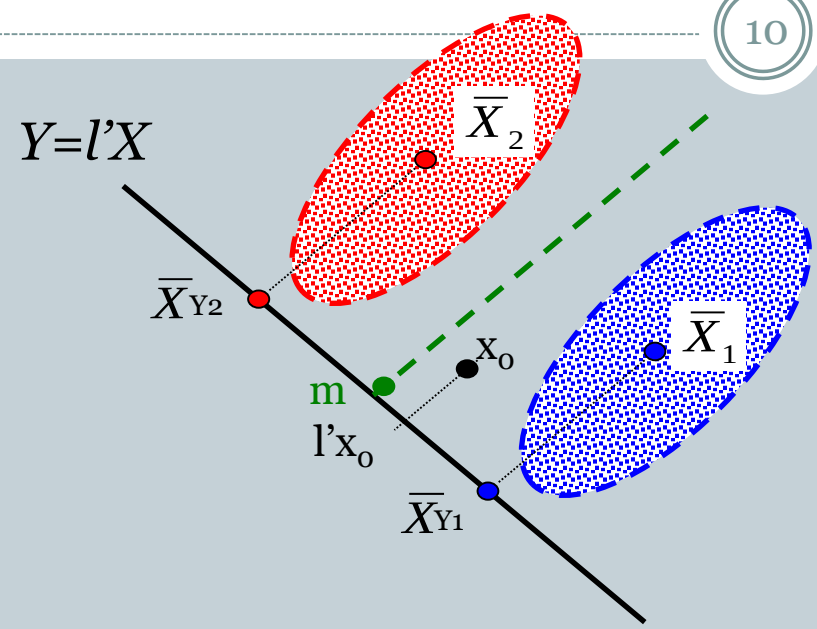
$$Y = \hat{l}' X = (\bar{X}_1 - \bar{X}_2)' S_p^{-1} X$$

es la mejor entre las lineales para clasificar entre las dos poblaciones, con el criterio de Fisher.

- El *punto medio* es:

$$\hat{m} = \frac{1}{2}(\bar{X}_1 - \bar{X}_2)' S_p^{-1} (\bar{X}_1 + \bar{X}_2).$$

Regla discriminante lineal de Fisher



- Dada una nueva observación x_o, la regla de Fisher es
  - Asignar x_o a π_1 si $(\overline{X}_1 - \overline{X}_2)' S_p^{-1} x_0 - \hat{m} \geq 0$
  - Asignar x_o a π_2 en otro caso

Optimalidad de la regla

- La regla discriminante de Fisher no considera riesgos ni errores.
- Se construye optimizando una clase especial de clasificadores.
- Genera una regla simple.
- En casos particulares puede verse que el error de la regla de Fisher alcanza el riesgo de Bayes, por lo cual es el clasificador óptimo.
- En general no se cumple esa relación.

Funciones discriminantes para la normal

- Vimos que la clasificación con tasa de error mínima puede alcanzarse con la función de discriminación

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

- Caso normal multivariada

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

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

Caso Σ_i diferentes por categoría

- Las matrices de covarianza son diferentes por categoría

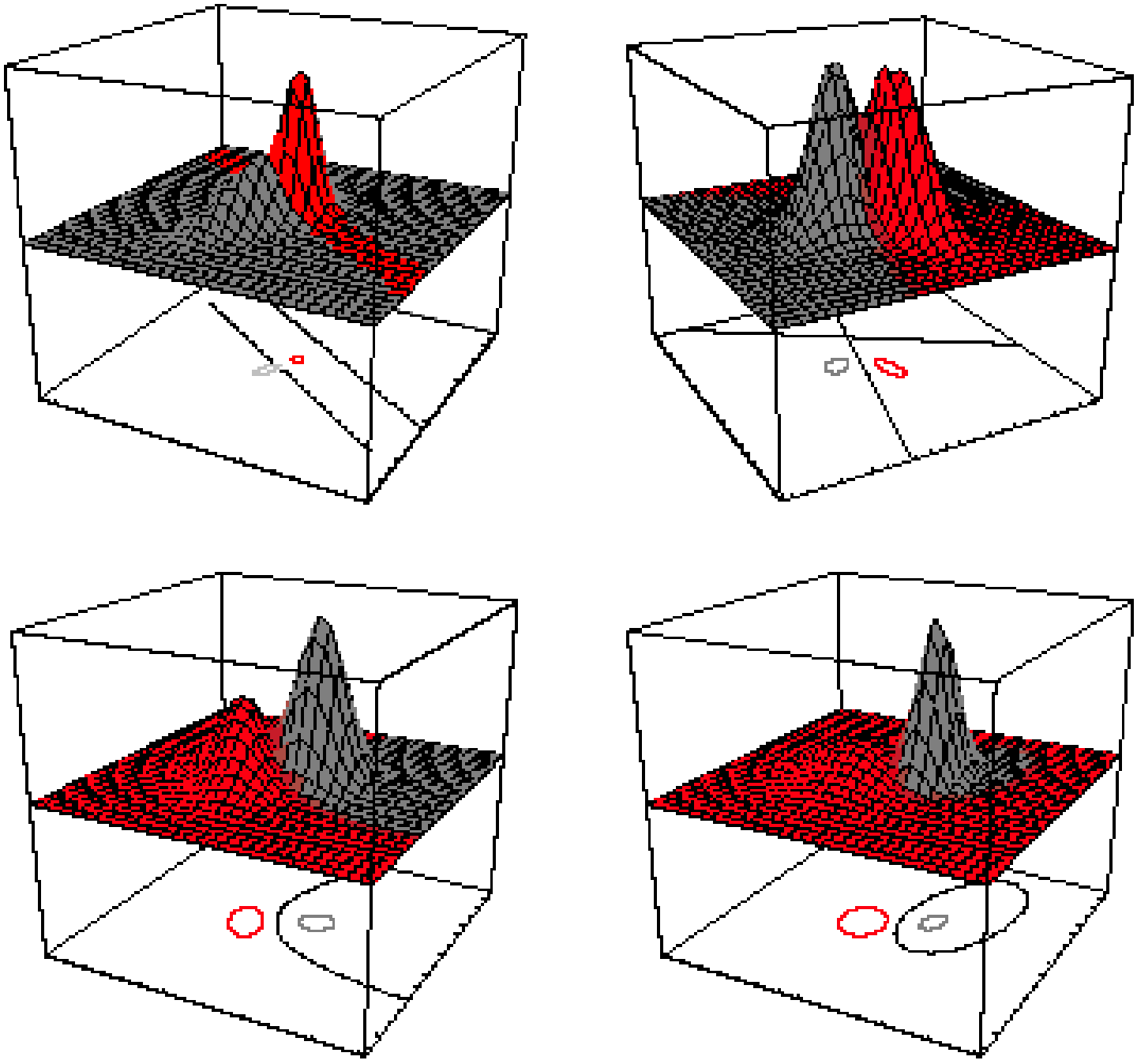
$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

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

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

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

- (Hipercuadricas las cuales son: hiperplanos, pares de hiperplanos, hiperesferas, hiperelipsoides, hiperparaboloides, hiperhiperboloides)



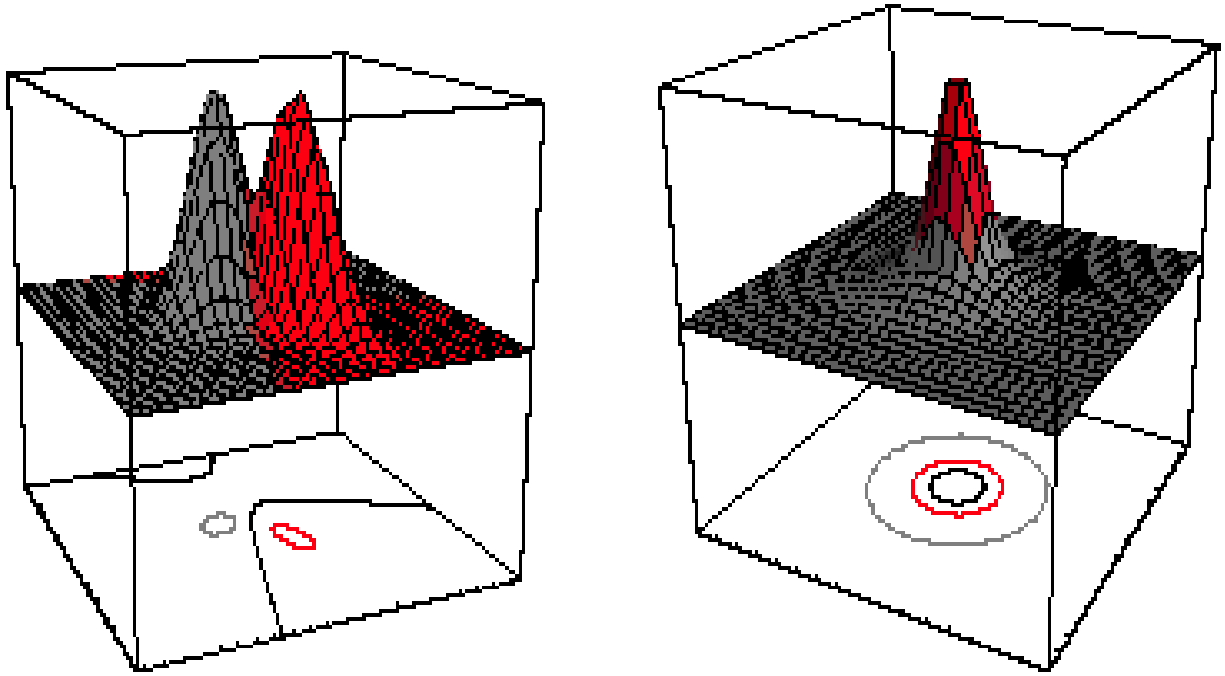


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

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

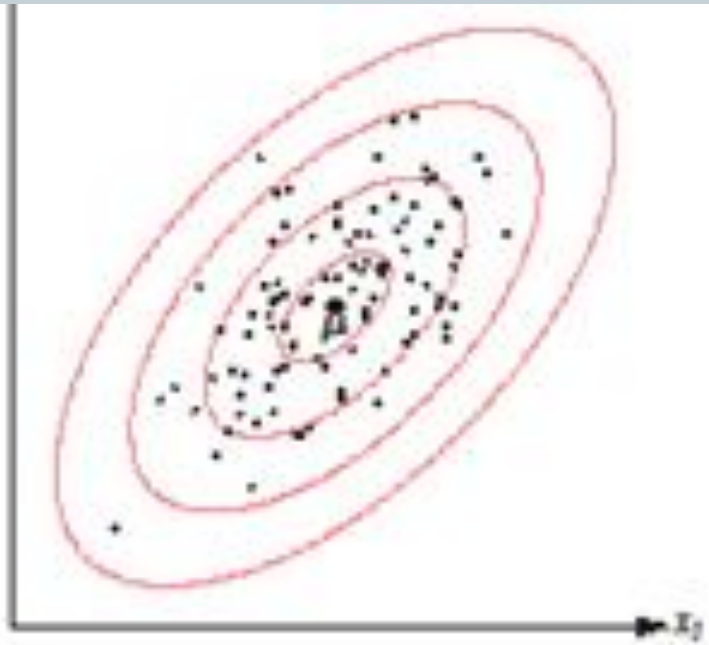


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

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

- Las matrices de covarianza son iguales

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

$$W_i = -\frac{1}{2\sigma^2} I$$

$$w_i = \frac{1}{\sigma^2} \mu_i$$

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

- El término cuadrático se repite en todas las g_i

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

- Sacando el término cuadrático

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

donde :

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

(w_{i0} se llama umbral para la ***i*** th categoria)

- Función discriminante es lineal en X

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

- Un clasificador que usa funciones discriminantes lineales se llama " linear machine"

- Las superficies de decisión de una maquina lineal son piezas de hiperplanos definidos por

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

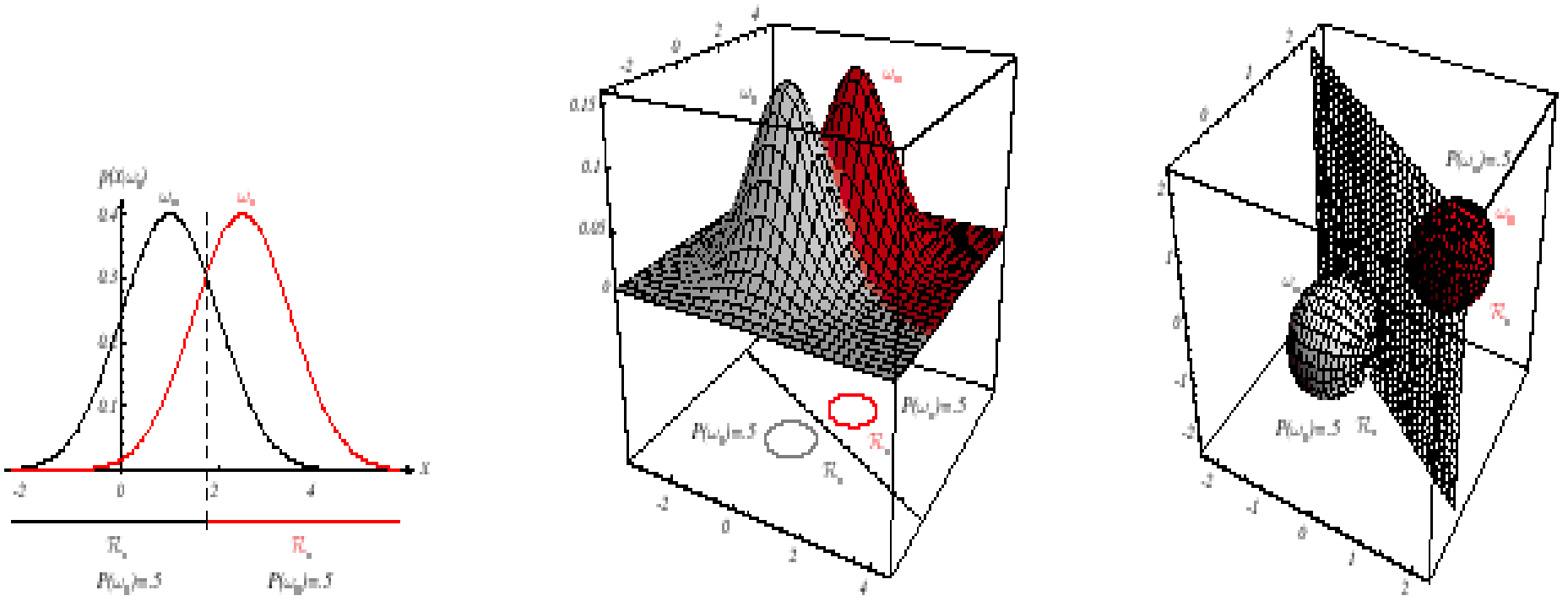


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

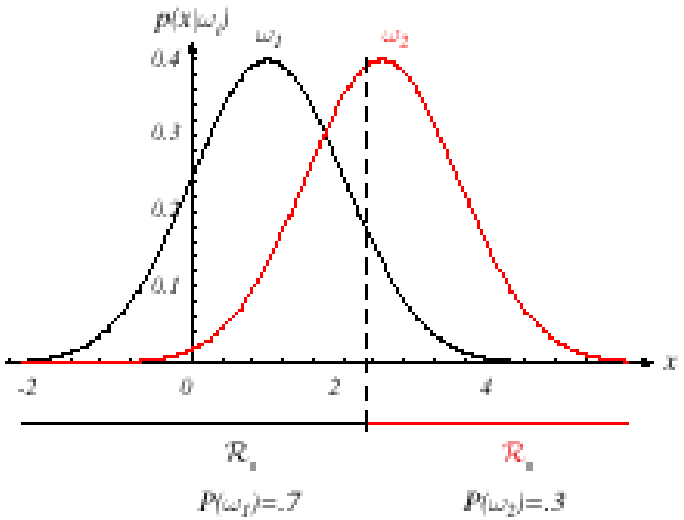
Caso esférico: $\Sigma_i = \sigma^2.I$

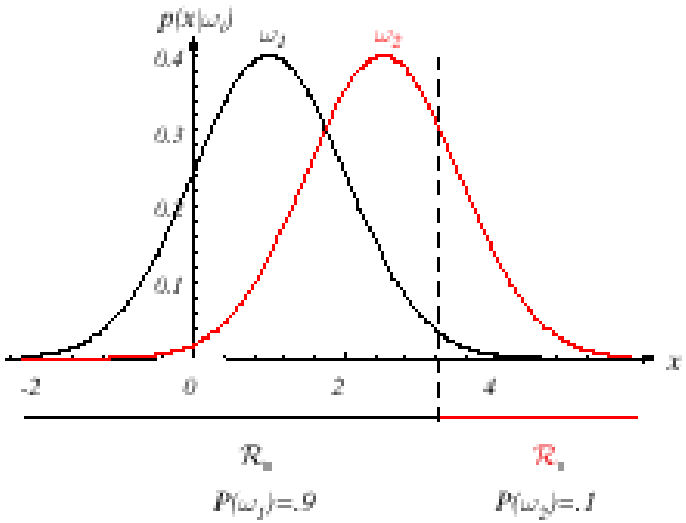
El hiperplano que separa $\mathcal{R}_i$ y $\mathcal{R}_j$

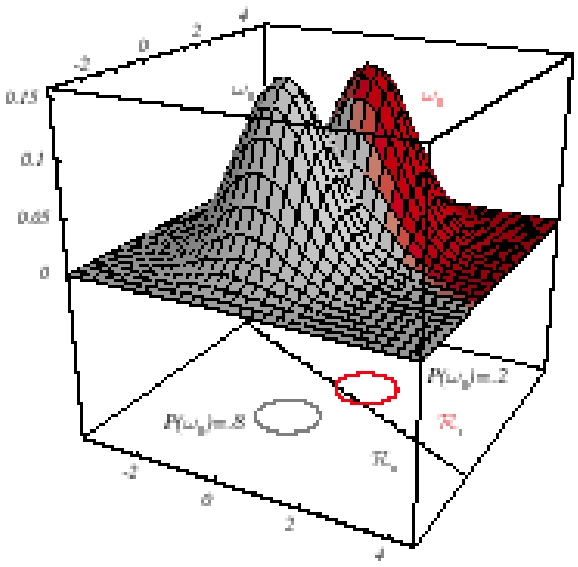
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

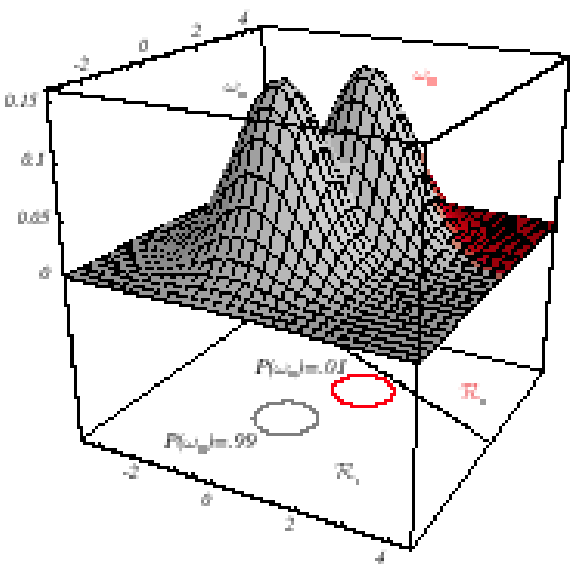
es siempre ortogonal a la recta que conecta las medias!

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


p(x|ω_j)
ω_1
ω_2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω_1)=.7
P(ω_2)=.3


p(x|ω_j)
ω_1
ω_2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω_1)=.9
P(ω_2)=.1


0.15
0.1
0.05
0
P(ω_2)=.2
P(ω_1)=.8


0.15
0.1
0.05
0
P(ω_2)=.01
P(ω_1)=.99

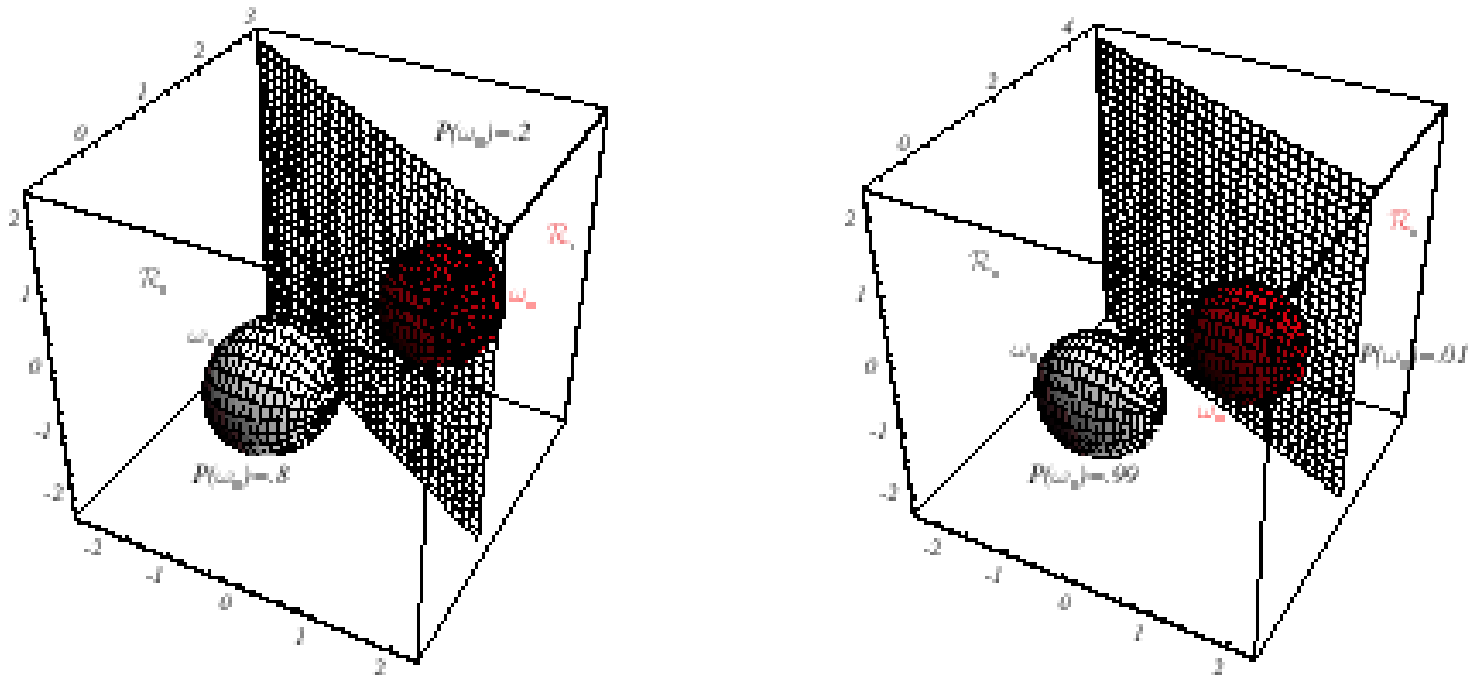


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Caso $\Sigma_i = \Sigma$

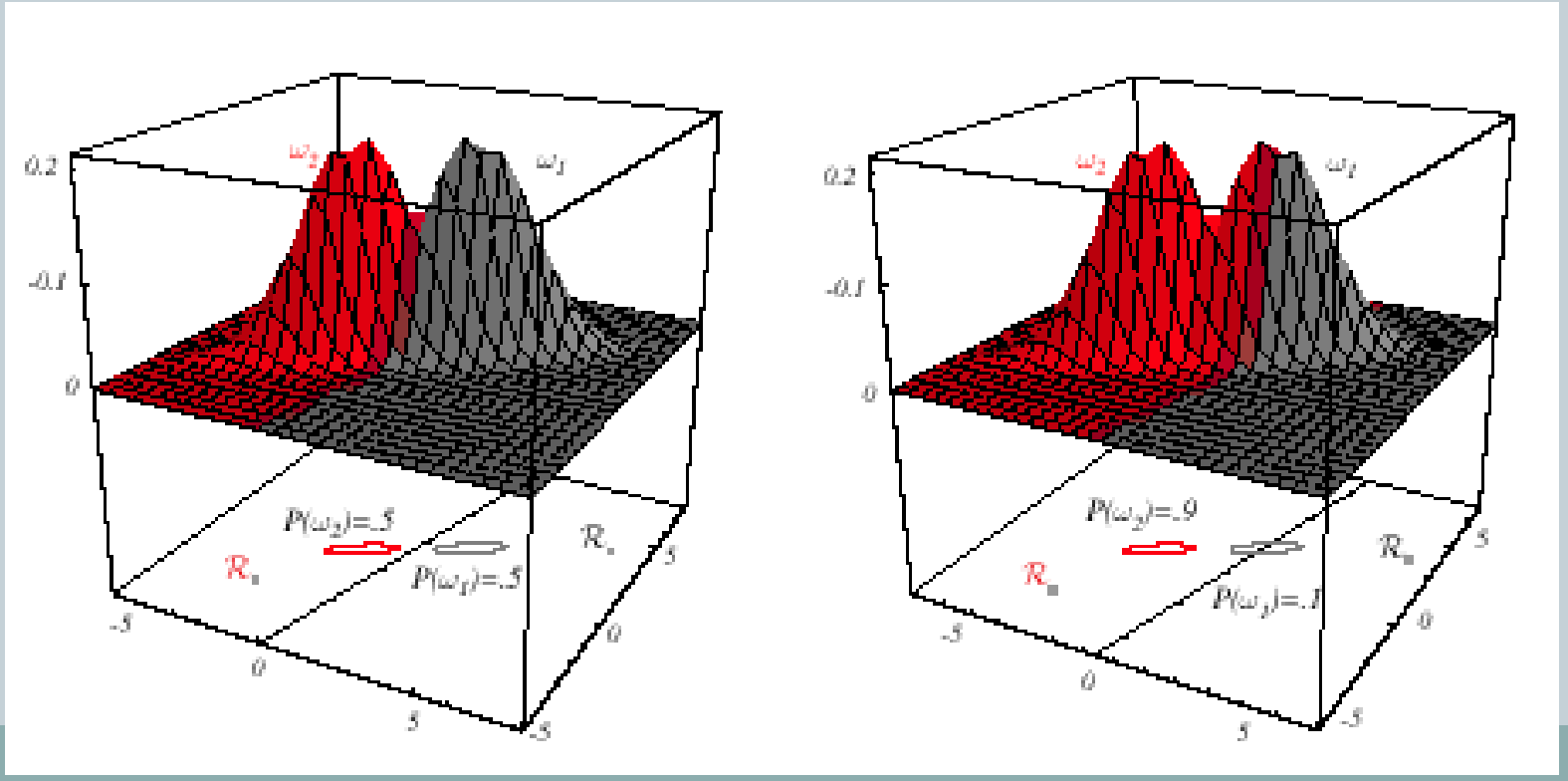
- En este caso las covarianzas de todas las clases son idénticas pero no arbitrarias
- Hiperplano que separa $\mathcal{R}_i$ y $\mathcal{R}_j$

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

- Este hiperplano generalmente no es ortogonal a la línea que une las medias

Probabilidades a priori cambian la regla

- Con probabilidades a priori iguales regla lineal ortogonal a la línea que une las medias
- Con probabilidades a priori diferentes no



Probabilidades a priori cambian la regla

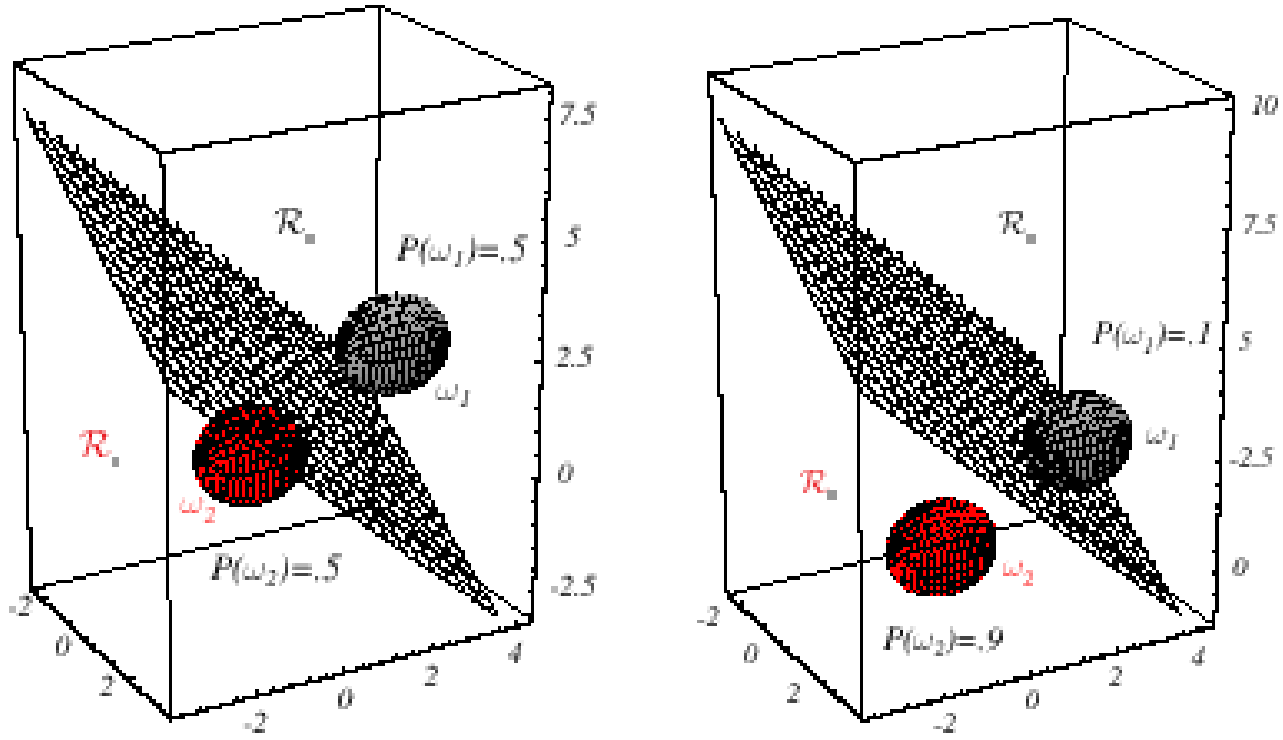


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Fuentes de error

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

$$P(error) = P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2)$$

$$= P(x \in R_2 \mid \omega_1)P(\omega_1) + P(x \in R_1 \mid \omega_2)P(\omega_2)$$

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

Fuentes de error

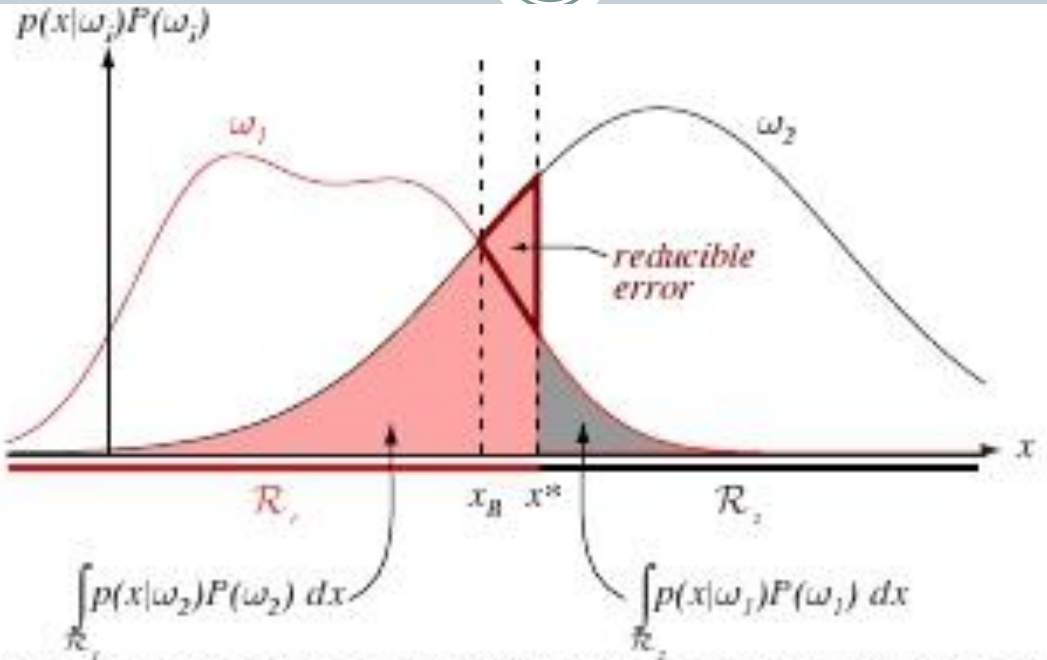


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary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

Fuentes de error

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

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

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

- La regla que maximiza esta cantidad elige la región que maximiza el integrando para todo x.

Cotas de error

- El calculo del error en muchos casos es muy difícil, pero para el problema dicotómico, se puede dar una cota superior al error usando la identidad

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

- Cota de Chernoff

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

- Si las densidades condicionales son normales, este integrando puede ser evaluado

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

Cotas de error: Normales

- La función $k(\beta)$ es

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

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

- La cota de Chernoff para *P(error)* se encuentra analítica o numéricamente buscando el valor de β que minimiza

$$p^{\beta}(x|\omega_1)p^{1-\beta}(x|\omega_2)e^{-k(\beta)}$$

- reemplazando ese valor en la cota del error.

Cotas de error: caso normal

- El error se acota minimizando en el espacio β *unidimensional*.
- La dependencia de β provoca que la cota sea mala para β cercanos a cero y uno
- Tomando $\beta=1/2$ se reduce el problema computacional generando la llamada Cota de Bhattacharyya

$$P(error) \leq \sqrt{P(\omega_1)P(\omega_2)} \int_R \sqrt{p(x|\omega_1)p(x|\omega_2)}dx$$

$$= \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)}$$

- En el caso no normal, las cotas pueden ser no informativas

Cotas de error para densidades normales

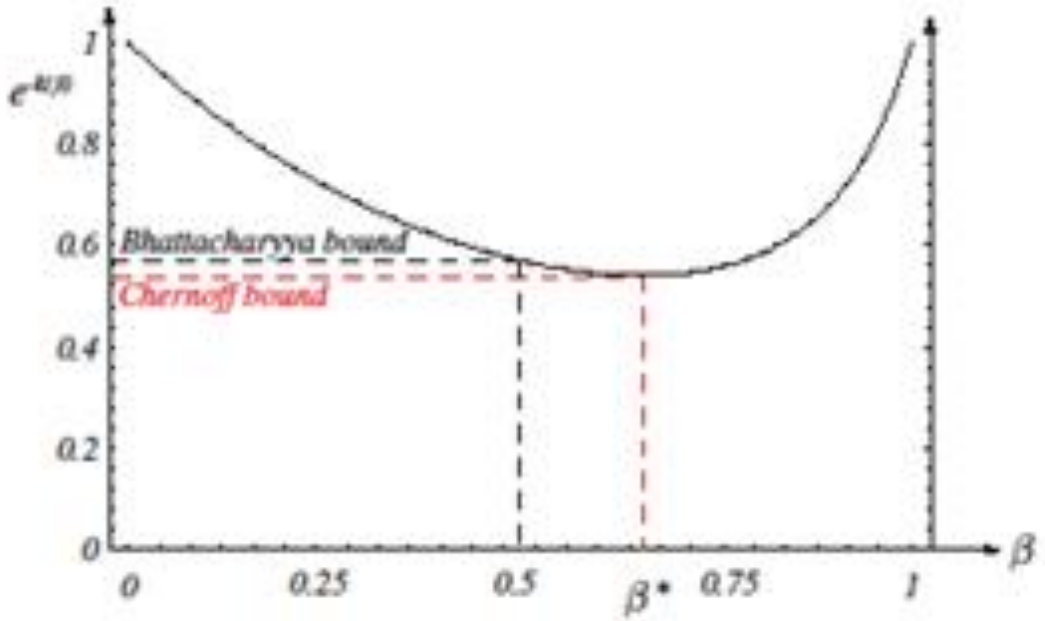


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Teoría de detección de señales y características de operación

- Supongamos que queremos detectar un pulso como un parpadeo de luz o una señal de radar débil.
- 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

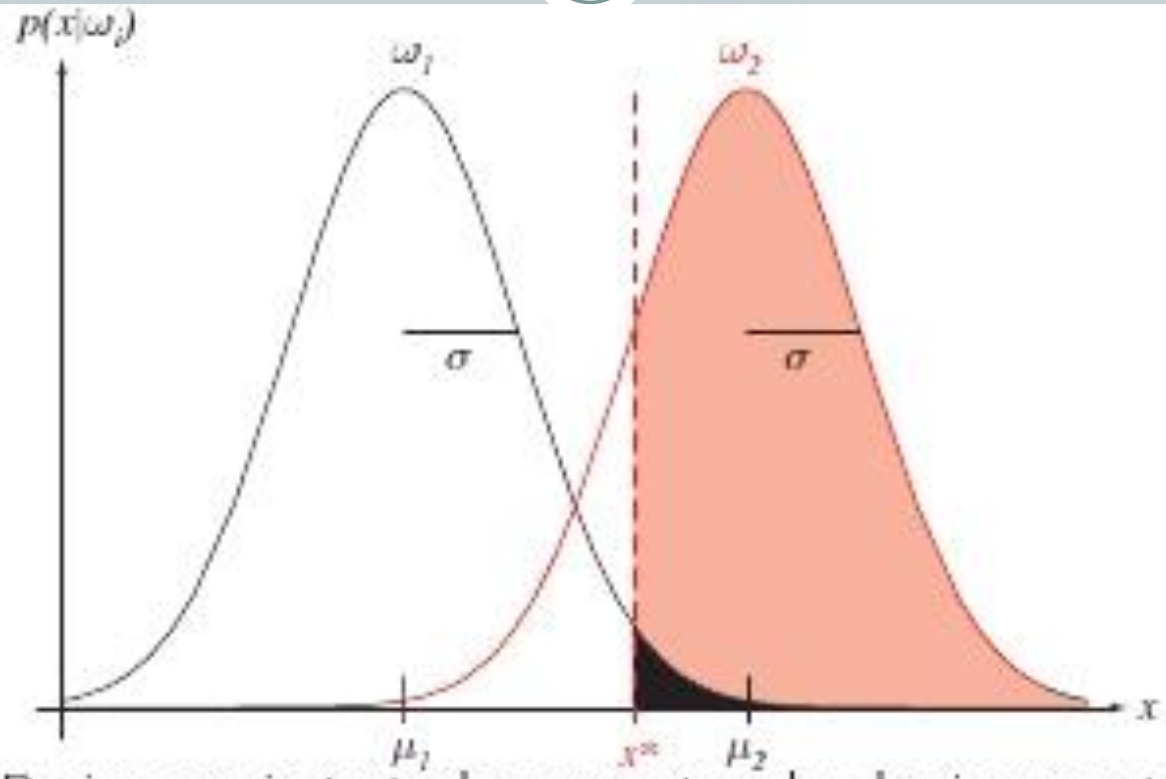


FIGURE 2.19. During any instant when no external pulse is present, the probability density for an internal signal is normal, that is, $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine 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 detección x* es inherente a cuan distantes están las normales.
- Esto se puede medir con $$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$
- Si *d* es grande, las normales están separadas y la discriminación es grande
- Si tenemos acceso a una muestra de entrenamiento podemos estimar d , μ_1, y μ_2 .

Notación

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

Curva ROC (recepción-operación)

- Una curva ROC se arma graficando la probabilidad de acierto en función de la probabilidad de falsa alarma cuando las densidades normales con igual varianza están fijas y la regla cambia.
- La regla está caracterizada por el valor de umbral x*.
- La abscisa es $P(x>x^* | x \epsilon \omega_1)$ y la ordenada es $P(x>x^* | x \epsilon \omega_2)$ para cada valor de x*.
- Las probabilidades pueden estimarse con tasas de acierto y falsa alarma sobre una database grande.

ROC curve

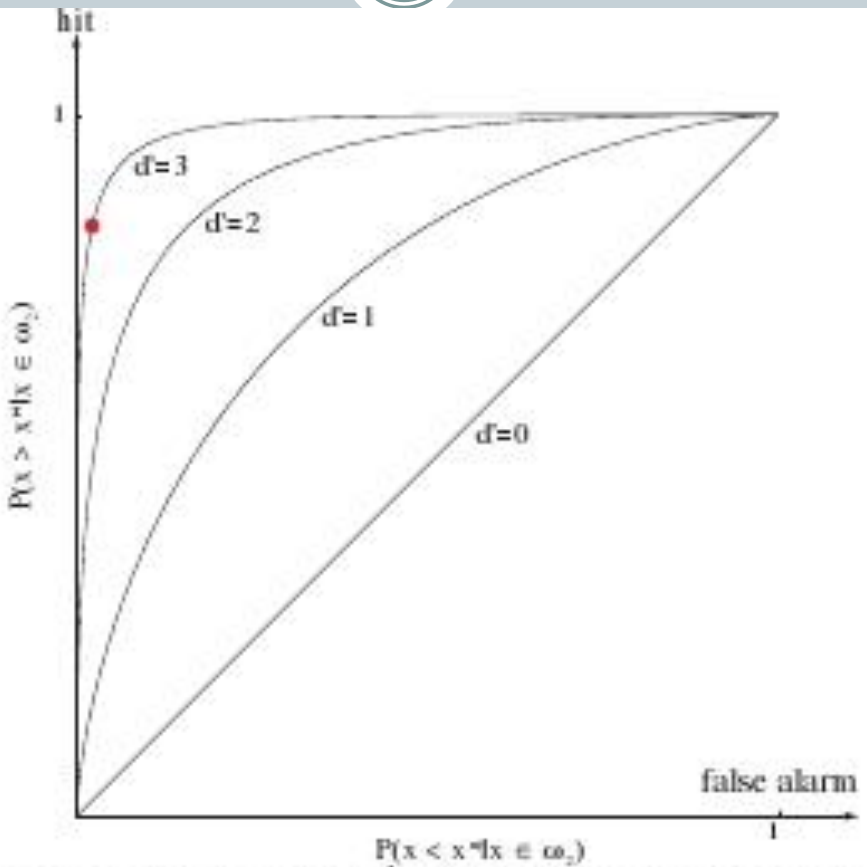


FIGURE 2.20. In a receiver operating characteristic (ROC) curve, the abscissa is the probability of false alarm, $P(x > x^*|x \in \omega_1)$, and the ordinate is the probability of hit, $P(x > x^*|x \in \omega_2)$. From the measured hit and false alarm rates (here corresponding to x^* in Fig. 2.19 and 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.

Distribuciones generales

- En la práctica, si no se considera optimalidad se puede armar una curva de características de operación.
- Esta curva se construye cuando se puede variar un solo parámetro en la regla de decisión y graficar las tasas de falsa alarma y acierto resultantes.
- Esto es útil cuando se necesita cambiar la función de pérdida λ_{ij} , dado que puede deducirse de la curva el parámetro óptimo para el nuevo riesgo

Características discretas

- Componentes de x son binarias o a valores enteros, x puede tomar uno de m valores discretos

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

- Caso de características binarias independientes en el problema de dos categorías
- Sea $x = [x_1, x_2, \ldots, x_d]^t$ donde cada x_i es 0 ó 1, con probabilidades:

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

Características discretas

- Componentes de x son binarias o a valores enteros, x puede tomar uno de m valores discretos

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

- En estos casos las integrales deben ser reemplazadas por sumas sobre todos los valores posibles de la variable x
- La definicion de riesgo de Bayes queda sin cambios, y la regla de Bayes sigue siendo al regla α que minimiza el riesgo de Bayes
- La regla simple de minimizar el error mediante la maximización de la probabilidad a posteriori, reemplazando en las fórmulas las densidades continuas por las probabilidades discretas y las densidades discretas

Características discretas

- Si se consideran características binarias independientes en el problema de dos categorías
- Sea $x = [x_1, x_2, \ldots, x_d]^t$ donde cada x_i es 0 ó 1, con probabilidades $p_i = P(x_i = 1 \mid \omega_1)$ $q_i = P(x_i = 1 \mid \omega_2)$

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

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

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

Características discretas

- *Aplicando la fórmula discriminante*

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

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

donde :

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

decide ω_1 si $g(x) > 0$ y ω_2 si $g(x) \leq 0$