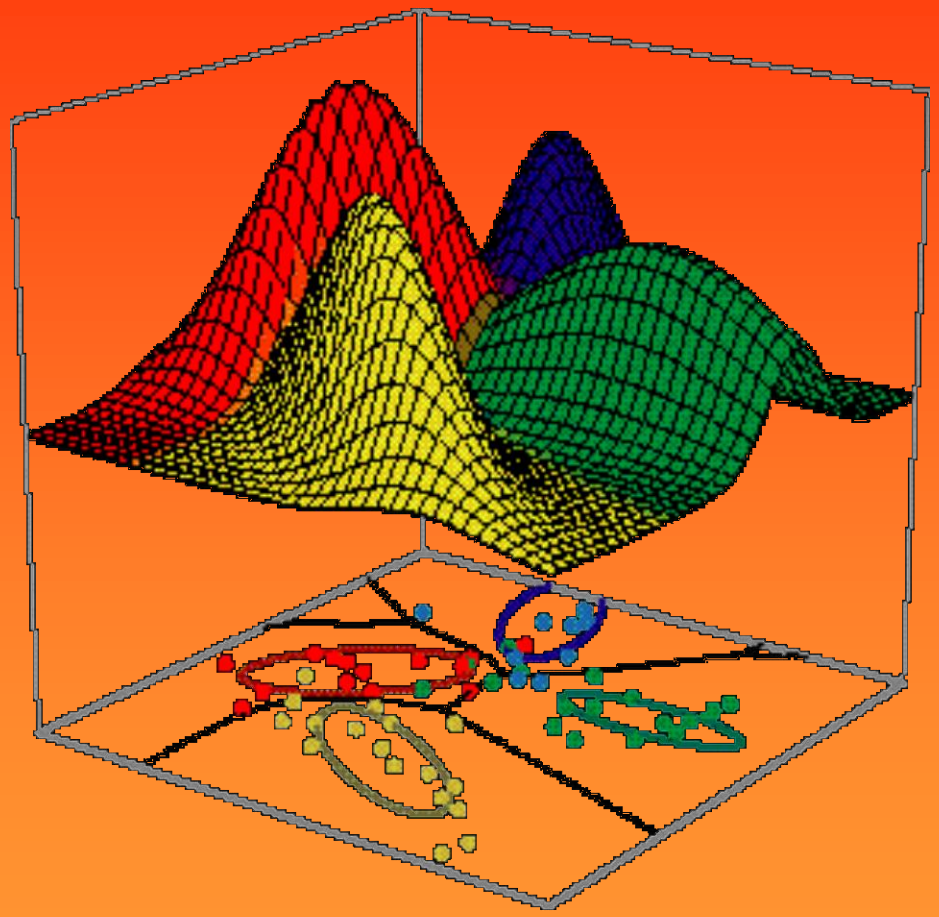


Pattern Classification

Capitulo 2 :
Teoria de Decision Bayesiana
(Secciones 2.3-2.5)

Clasificacion con minima tasa de error

Clasificadores, Funciones discriminantes y superficies de decision

La densidad Normal

Clasificador con minima tasa de error

- Recordemos que las acciones son decisiones sobre las clases Si la accion α_i se toma y el estado verdadero es ω_j entonces la decision es correcta si $i = j$ e incorrecta cuando $i \neq j$
- *Teoria de la decision pide que se minimize el riesgo total, que es la esperanza de la funcion de perdida*

$$R = \int_{-\infty}^{\infty} R(\alpha(x) \mid x)\, p(x) dx$$

- *R se minimiza si $R(\alpha_i \mid x)$ se achica para todo i:1,…,c*

- Funcion de perdida cero-uno:

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

Por lo cual el riesgo condicional es :

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

$$= \sum_{j \neq 1} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

- Minimizar el riesgo requiere entonces maximizar $P(\omega_i \mid x)$ *para todo i:1,…,c*

- La regla de Bayes para la funcion de perdida cero uno
  - asigna x a ω_i *si* $x \in R_{i,0}$ *con*

$$R_{i,0} = \left\{ x \mid p(\omega_i \mid x) > p(\omega_j \mid x),\ j \neq i \right\}$$

- *Esta es la regla que surge de minimizar el error total de mala clasificacion*

$$P(error) = \sum_{i=1}^{c} P(error \mid \omega_i)\, p(\omega_i) = 1 - \sum_{i=1}^{c} P(decidir\ \ \omega_i \mid \omega_i)\, p(\omega_i)$$

$$= 1 - \sum_{i=1}^{c} [\int_{R_{i,0}} p(x \mid \omega_i) dx]\, p(\omega_i) = 1 - \int_{R} \sum_{i=1}^{c} I_{R_{i,0}}(x) P(\omega_i \mid x)\, p(x) dx$$

$$\leq 1 - \int_{R} \sum_{i=1}^{c} I_{R_i}(x) P(\omega_i \mid x)\, p(x) dx \quad pues \quad P(\omega_i \mid x)\ es\ \text{maxima}\ \text{en}\ R_{i,0}$$

Regiones de decision λ general en dos poblaciones

$$Sea \quad \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}.\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda, \qquad \text{se decide por } \omega_1 \text{ si: } \frac{P(x\mid\omega_1)}{P(x\mid\omega_2)}>\theta_\lambda$$

- Si λ es la funcion de perdida cero-uno

$$si \quad \lambda=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \quad entonces \quad \theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}=\theta_a$$

- Si λ duplica el costo de misclasificar la clase 2

$$si \quad \lambda=\begin{pmatrix}0 & 2\\ 1 & 0\end{pmatrix} entonces \quad \theta_\lambda=\frac{2P(\omega_2)}{P(\omega_1)}=\theta_b$$

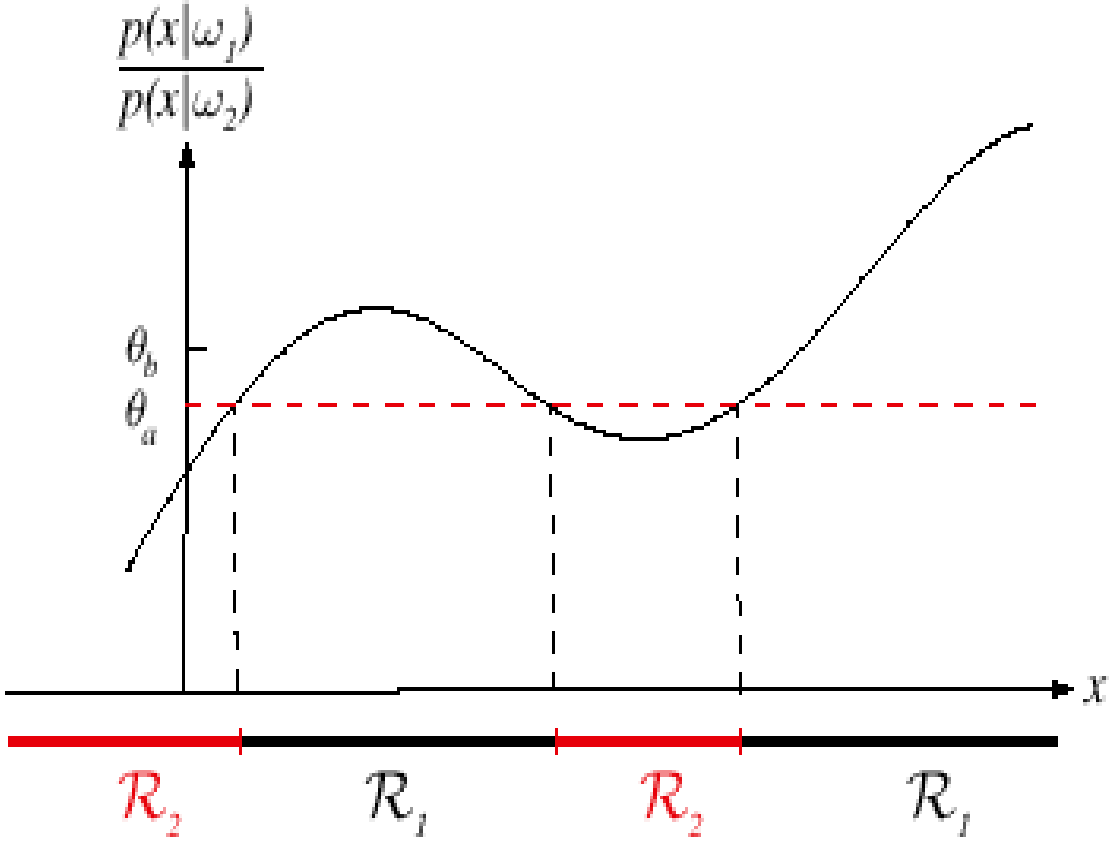


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Otros Riesgos

- Minimax,
  - maximiza el riesgo sobre el conjunto de probabilidades a priori, y luego selecciona la regla que minimiza ese riesgo (sobredimensionado)
  - Tiene sentido cuando los estados naturales estan movidos por un oponente que pretende hacer el mayor daño posible, por lo cual se prepara para minimizar esos daños
- Neyman-Pearson
  - Minimiza el riesgo total sujeto a una restriccion para alguna clase

$$\int R(\alpha_i, | \, x) dx < \text{constante}$$

Clasificadores, Funciones Discriminantes y Superficies de Decision

Supongamos tener un conjunto de funciones discriminantes

$$g_i(x),\ i = 1,\ldots,\ c$$

El clasificador asigna el vector de caracteristicas x a la clase ω_i si:

$$g_i(x) > g_j(x)\ \ \forall j \neq I$$

Por lo cual el clasificador puede verse como una red o maquina que calcula c funciones discriminantes y selecciona una categoria correspondiente al mayor discriminante.

Los clasificadores bayesianos pueden ser representados de esta forma, para los distintos riesgos y funciones de perdida.

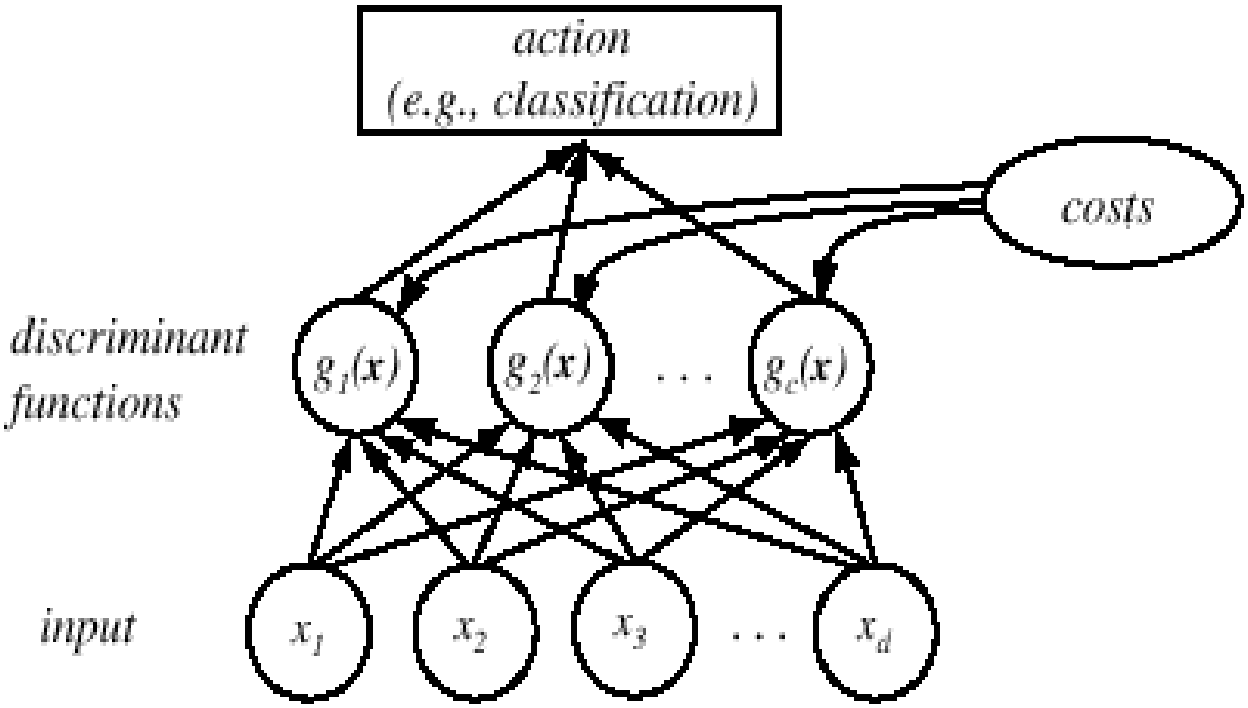


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction 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.

Caso general con riesgo R,

$$g_i(x) = - R(\alpha_i \mid x)$$

Aqui el maximo discriminante corresponde al minimo riesgo.

Caso de tasa de error minima, tomamos

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

Aqui la maxima discriminacion corresponde al maximo a posteriori.

Las funciones discriminantes no son unicas

- Las siguientes funciones son equivalentes

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

$$g_i^b(x) = p(x \mid \omega_i)P(\omega_i)$$

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

- Decisiones equivalentes parten el espacio de caracteristicas de forma igual.

- Se definen entonces las reglas de acuerdo a las funciones discriminantes que impliquen la menor cantidad de operaciones computacionales.

- Las regiones de decision definidas son $\mathcal{R}_1, \ldots, \mathcal{R}_c$

$$R_i = \left\{ x \mid g_i(x) > g_j(x),\ j \neq i \right\}$$

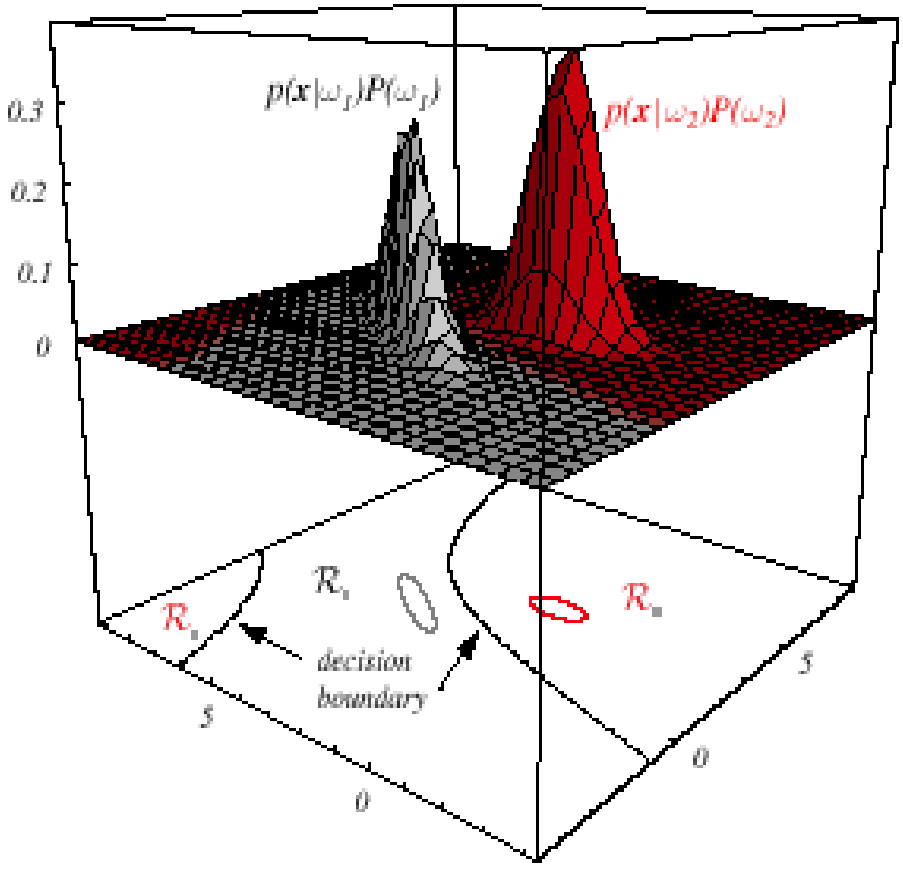


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Caso de dos categorias

– Un clasificador "dicotomico" tiene dos funciones discriminantes g_1 y g_2 ,*pero se usa una usualmente*

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

Decide ω_1 *si* $g(x) > 0$; en otro caso se decide por ω_2

$$g^a(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$

$$g^b(x) = \ln \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Densidad Normal

- Univariada
  - Densidad analiticamente manejable
  - Densidad continua
  - Muchos procesos son asintoticamente gausianos
  - Caracteres escritos a mano, sonidos del habla se modelan como prototipos corruptos por un proceso aleatorio

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

Donde:

μ = esperanza de x

σ^2 = varianza

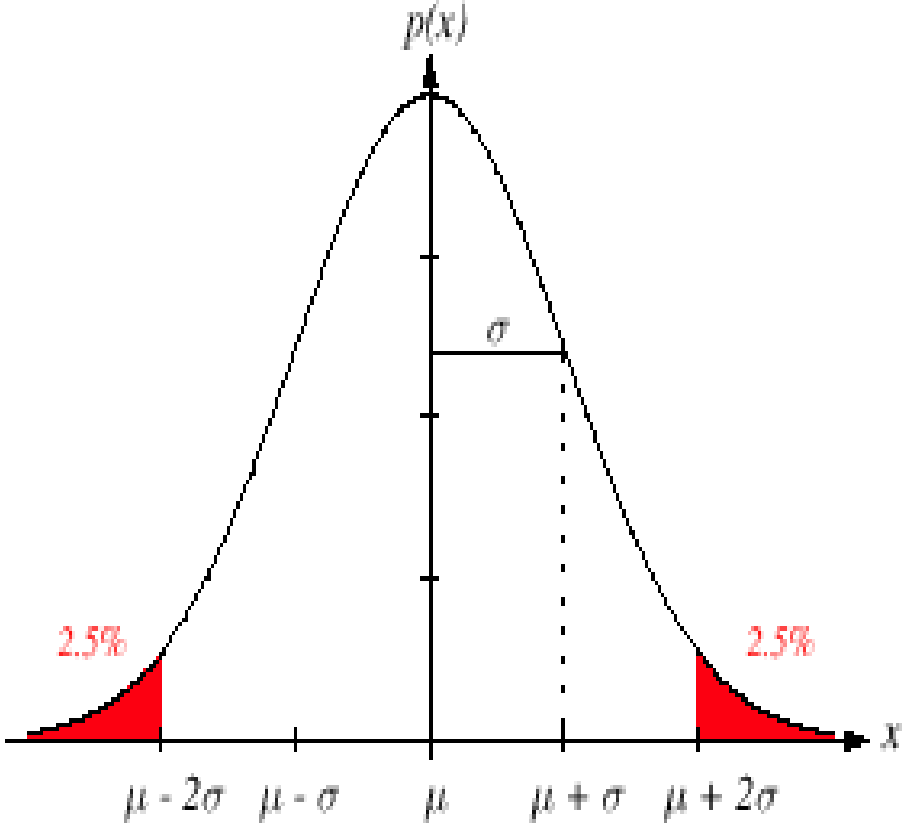


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

Densidad Multivariada

La densidad Multivariada en d dimensiones es:

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

donde:

- $x = (x_1, x_2, \ldots, x_d)^t$
- $\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ vector de medias
- Σ matriz de varianza covarianza
- $|\Sigma|$ y Σ^{-1} son el determinante y su inversa

Definiciones

$$\mu = E(x) = \int x p(x) dx$$

$$\Sigma = E[(x-\mu)(x-\mu)'] = \int (x-\mu)(x-\mu)' \, p(x) dx$$

- Distancia de Mahalanobis

$$d(x,y) = (x-\mu)'\Sigma^{-1}(x-\mu)$$

- Contornos de densidad constante son hiperelipsoides con distancia de mahalanobis constante.
- Proyecciones de normales son normales

$$X \sim N(\mu,\Sigma) \quad entonces \quad AX \sim N(A\mu, A'\Sigma A)$$

$$l'X \sim N(l'\mu, l'\Sigma l) \quad \text{univariada}$$

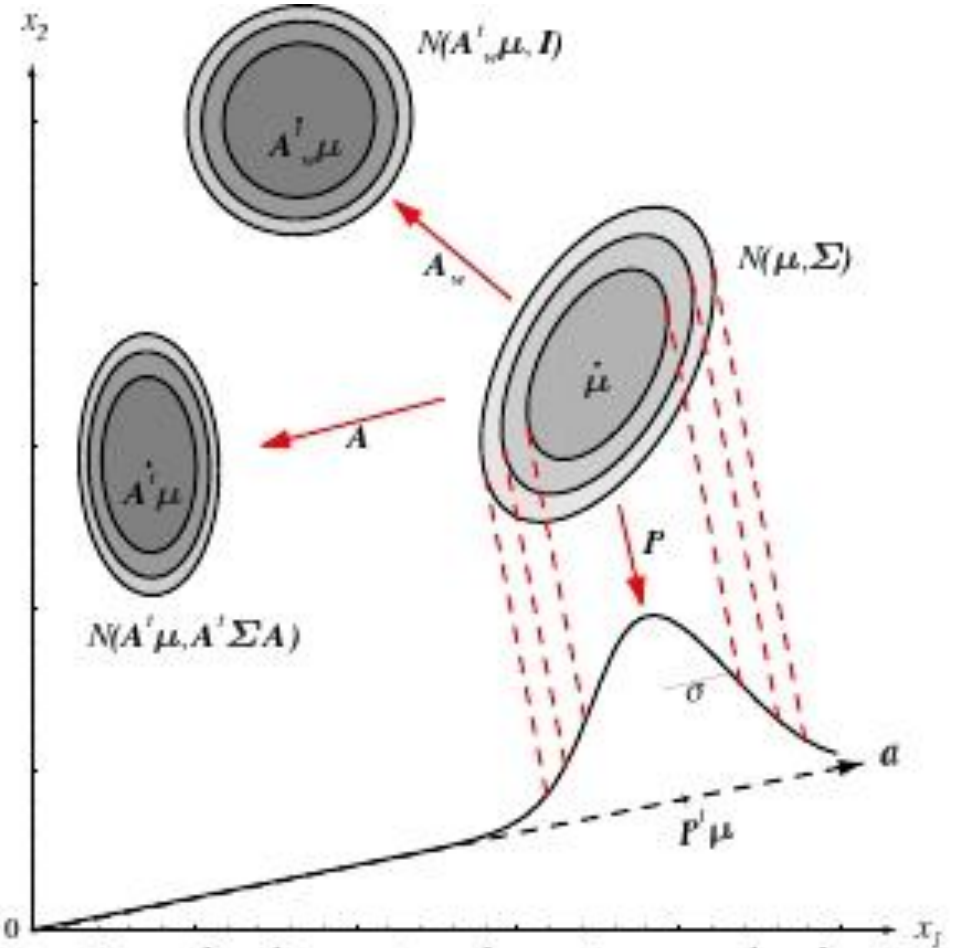


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, $\mathbf{A}$, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. Another linear transformation—a projection $\mathbf{P}$ onto a line defined by vector $\mathbf{a}$—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.

Regla discriminante lineal de Fisher

Sea la variable $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ y dos poblaciones π_1 y π_2

Sean $E_{\pi_1}(X) = \mu_1 \quad E_{\pi_2}(X) = \mu_2$

$V_{\pi_1}(X) = V_{\pi_2}(X) = \sum.$

Se busca una 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

- Se tiene que

$$E_{\pi_1}(Y) = E_{\pi_1}(l'X) = l'\mu_1 = \mu_{Y1}$$

$$E_{\pi_2}(Y) = E_{\pi_2}(l'X) = l'\mu_2 = \mu_{Y2}$$

$$V_{\pi_1}(Y) = V_{\pi_1}(l'X) = l'\sum l = \sigma_Y^2 = V_{\pi_2}(l'X) = V_{\pi_2}(Y)$$

Regla discriminante lineal de Fisher

Hay que buscar *l* que optimice la separación entre las dos poblaciones:

Se maximiza la separación entre las medias:

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

Regla discriminante lineal de Fisher

Si se maximiza sin restricciones, el máximo puede no ser finito: se maximiza dividiendo por la varianza

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

- La solución que se obtiene es:

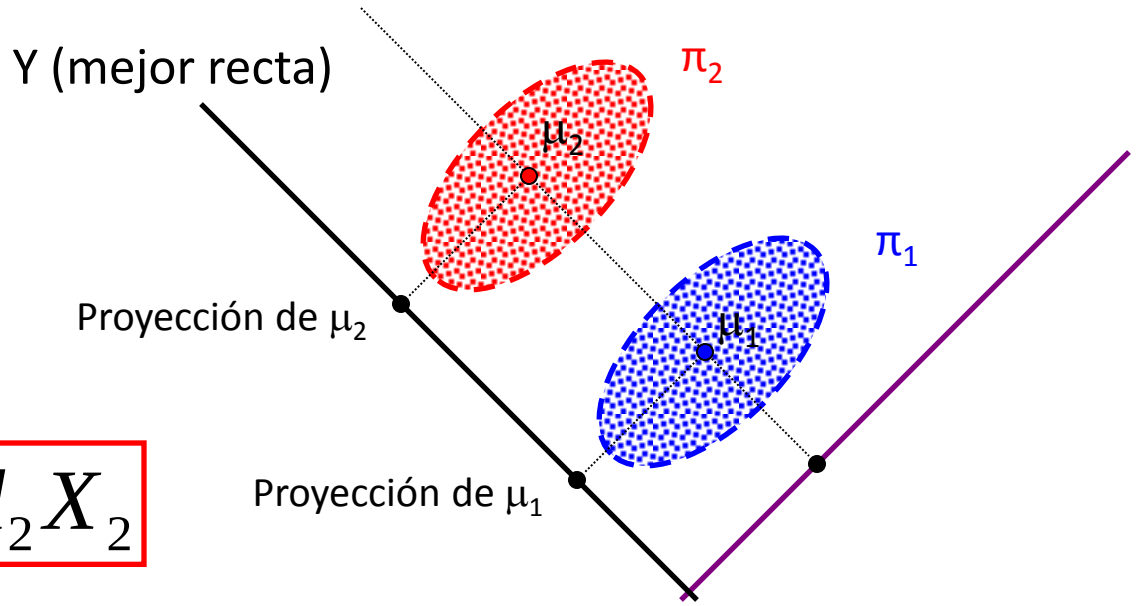
$$Y = (\mu_1 - \mu_2)' \Sigma^{-1} X$$

Función discriminante lineal de Fisher

- σ_Y^2 es comun a las dos poblaciones

Regla discriminante lineal de Fisher

- En el caso en que $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, se tiene:

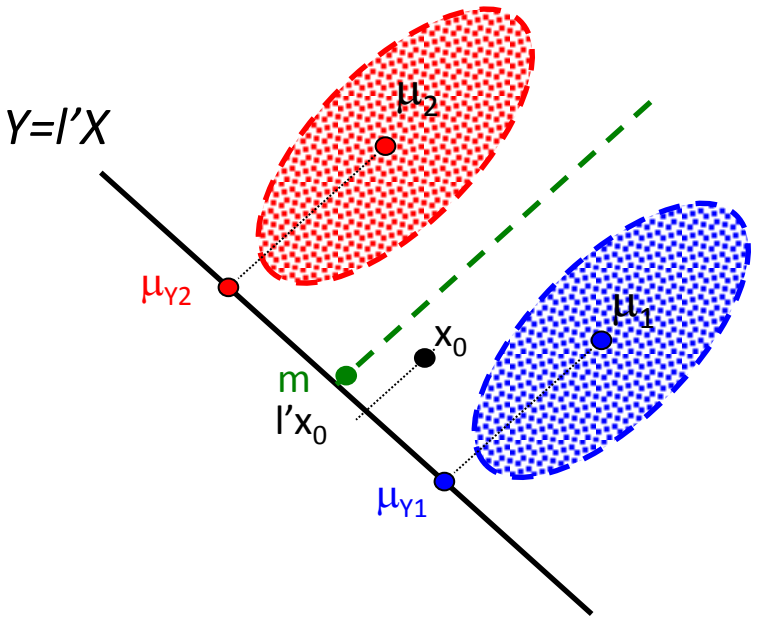


$$Y = l'X = l_1 X_1 + l_2 X_2$$

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_0:

- Asignar x_0 a π_1 si

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

- Asignar x_0 a π_2 si

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

Regla discriminante lineal de Fisher: *Versión muestral*

Dadas dos poblaciones π_1 y π_2, se tienen las siguientes matrices de datos:

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

- Y sean $\bar{X}_1, \bar{X}_2$ las medias muestrales

$$\text{y } S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}.$$

Regla discriminante lineal de Fisher: *Versión muestral*

- La funcion discriminante lineal de Fisher es:

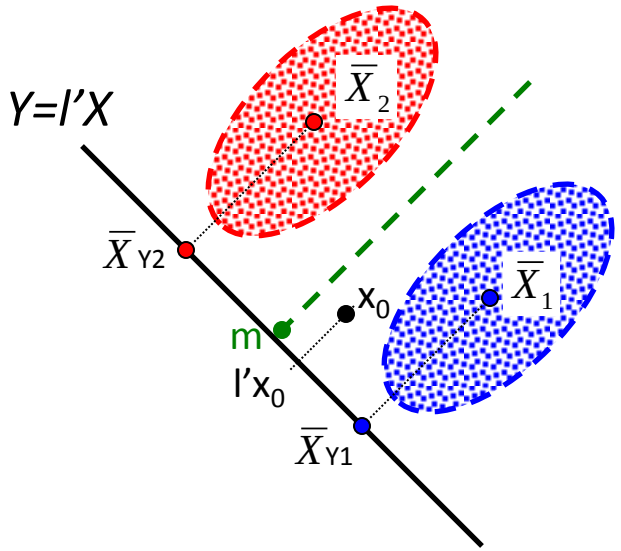
$$Y = \hat{l}' X = (\overline{X}_1 - \overline{X}_2)' S_p^{-1} X$$

que es óptima para clasificar entre las dos poblaciones

- El *punto medio* es:

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

Regla discriminante lineal de Fisher: *Versión muestral*



- Dada una nueva observación x_0, la regla de clasificación sería:
- Asignar x_0 a π_1 si $$(\bar{X}_1 - \bar{X}_2)' S_p^{-1} x_0 - \hat{m} \geq 0$$
- Asignar x_0 a π_2 si $$(\bar{X}_1 - \bar{X}_2)' S_p^{-1} x_0 - \hat{m} < 0$$

Ejercicio

- Sobre el grupo de datos Iris de matlab, estimar las distribuciones normales por clase
- Seleccionar dos clases
- Dividir en dos grupos en forma aleatoria dichas clases y generar funciones discriminantes de Fisher con el grupo de datos de entrenamiento
- Dar la tasa de error aparente usando el grupo de test