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

Clasificación con mínima tasa de error

Clasificadores, Funciones discriminantes y superficies de decisión

La densidad Normal

Clasificador con mínima tasa de error

- Recordemos que las acciones son decisiones sobre las clases
- Si la acción α_i se toma y el estado verdadero es ω_j entonces la decisión es correcta si $i = j$ e incorrecta cuando $i \neq j$
- *Teoría de la decisión pide que se minimice el riesgo total, que es la esperanza de la función de perdida con respecto a las clases y a la verosimilitud de la característica medida*

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

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

Función de pérdida cero-uno:

- Función de pérdida 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* $i:1,\ldots,c$

Función de pérdida cero-uno:

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

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

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

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

Función de pérdida cero-uno:

- *Reemplazando …*

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

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

Función de pérdida cero-uno:

.

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

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

$$\leq 1 - \int_{R} \sum_{i=1}^{c} I_{R_i}(x) P(\omega_i \mid x) \, p(x) dx$$

pues $P(\omega_i \mid x)$ *es* maxima en $R_{i,0}$

Regiones de decisión: dos clases

- Regiones de decisión λ general

$$Sea \;\; \theta_\lambda = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)},$$

$$\text{se decide por } \omega_1 \;\; \text{si}: \;\; \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \theta_\lambda$$

Regiones de decisión: dos clases

- Si λ es la función de pérdida 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 mal clasificar la clase 2

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

Regiones de decisión: dos clases

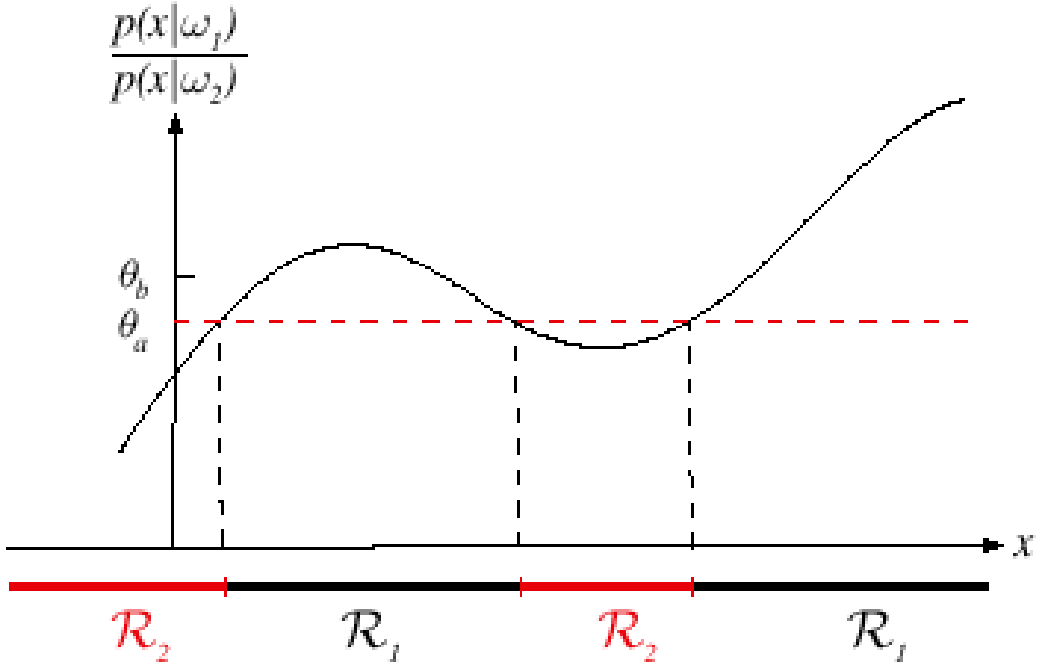

FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, 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*. Copyright © 2001 by John Wiley & Sons, Inc.

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 están 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 restricción para alguna clase

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

Riesgo minimax: Dos categorías

- Recordemos que el riesgo total para la regla con regiones de decisión R_1 es

$$R^* = \int R(\alpha^*(x) \mid x)\, p(x)\, dx$$

$$= \int_{R_1} \lambda_{11} P(x \mid \omega_1) P(\omega_1) + \lambda_{12} P(x \mid \omega_2) P(\omega_2)\, dx$$

$$+ \int_{R_2} \lambda_{21} P(x \mid \omega_1) P(\omega_1) + \lambda_{22} P(x \mid \omega_2) P(\omega_2)\, dx$$

Riesgo minimax: Dos categorías

$$R^* = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} P(x \mid \omega_2) dx$$

$$+ P(\omega_1) \left[(\lambda_{12} - \lambda_{22}) - (\lambda_{21} - \lambda_{11}) \int_{R_2} P(x \mid \omega_1) dx - (\lambda_{12} - \lambda_{22}) \int_{R_1} P(x \mid \omega_2) dx \right]$$

- Esta ecuación muestra que una vez que las regiones se fijan (i.e., *R1 y R2*), el riesgo total es lineal $P(\omega_1)$.
- Si podemos encontrar un borde de región tal que anule la constante que acompaña a $P(\omega_1)$, entonces el riesgo resultante es independiente de las probabilidades a priori

Riesgo minimax: Dos categorías

- Riesgo minimax

$$R^* = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} P(x \mid \omega_2) dx$$

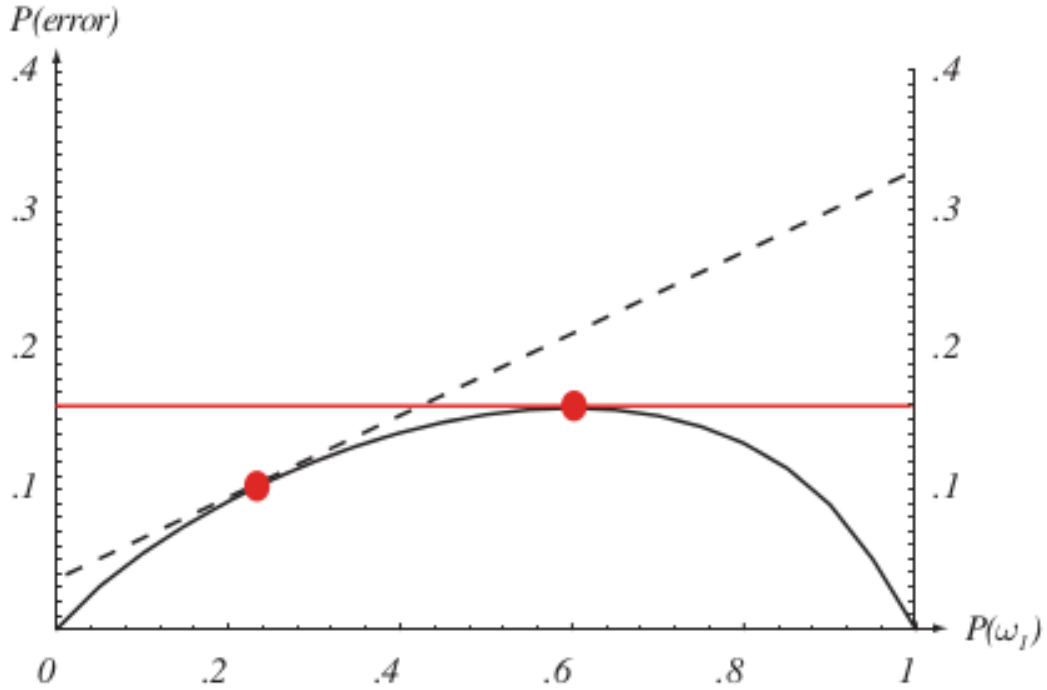

FIGURE 2.4. The curve at the bottom shows the minimum (Bayes) error as a function of prior probability $P(\omega_1)$ in a two-category classification problem of fixed distributions. For each value of the priors (e.g., $P(\omega_1) = 0.25$) there is a corresponding optimal decision boundary and associated Bayes error rate. For any (fixed) such boundary, if the priors are then changed, the probability of error will change as a linear function of $P(\omega_1)$ (shown by the dashed line). The maximum such error will occur at an extreme value of the prior, here at $P(\omega_1) = 1$. To minimize the maximum of such error, we should design our decision boundary for the maximum Bayes error (here $P(\omega_1) = 0.6$), and thus the error will not change as a function of prior, as shown by the solid red horizontal line. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Riesgo de Neyman Pearson

- En algunos problemas, se busca minimizar el riesgo total sujeto a una restricción para un *i* **particular**

$$\int R(\alpha_i \mid x)dx < cte$$

- Usualmente se ajustar las regiones de decisión numéricamente, pero para algunos casos, como el Gaussiano, es posible encontrar resultados analíticos

Clasificadores, Funciones Discriminantes y Superficies de Decisión

- Hay muchas formas de representar un clasificador
  - reglas de decisión $\alpha_i(x)$ $i = 1,\ldots, c$
    - *clasifico a x como* w_i *si* $R(\alpha_i(x)|x)$ *es mínima*
  - regiones o superficies de decisión R_i $i = 1,\ldots, c$
    - *clasifico a x como* w_i *si* $x \epsilon R_i$
  - funciones discriminantes $g_i(x)$, $i = 1,\ldots, c$
    - *clasifico a x como* w_i *si* $g_i(x) > g_j(x)$,

Funciones Discriminantes

- Supongamos tener un conjunto de funciones discriminantes $g_i(x)$, $i = 1,\ldots, c$, tal que el clasificador asigna el vector de características x a la clase ω_i si:

$$g_i(x) > g_j(x) \quad \forall j \neq i$$

- Por lo cual el clasificador puede verse como una red o maquina que calcula c funciones discriminantes y selecciona una categoría correspondiente al mayor discriminante.
- Los clasificadores bayesianos pueden ser representados de esta forma, para los distintos riesgos y funciones de perdida.

Clasificación

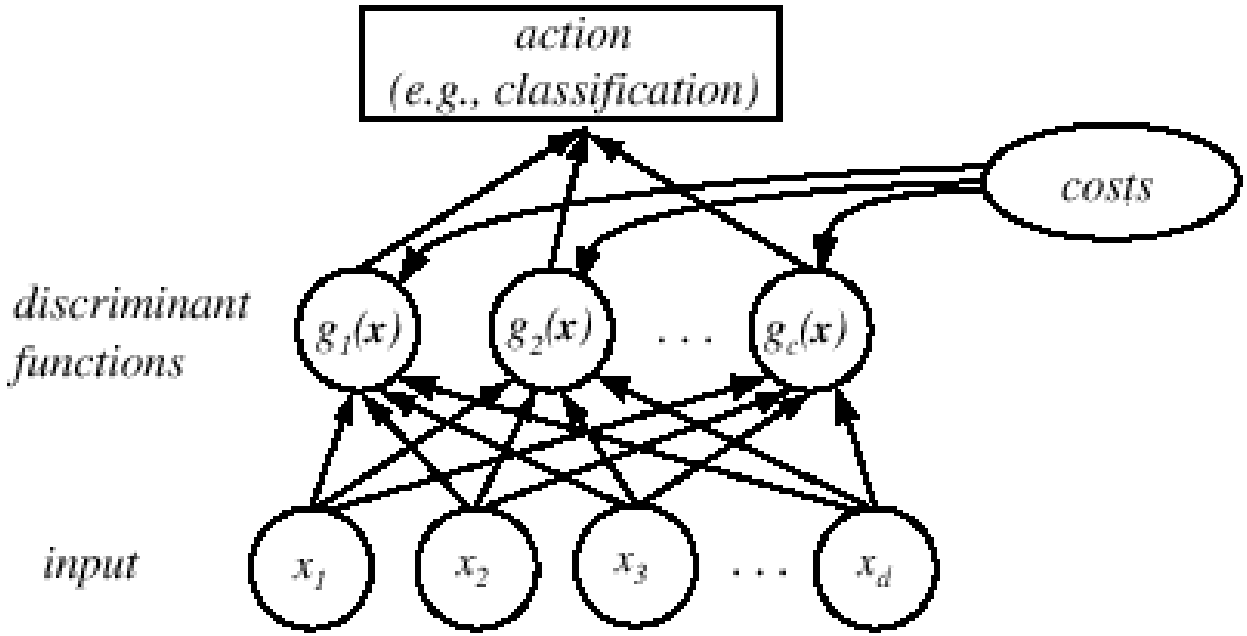

FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction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

Ejemplos

- *Caso general con riesgo R,*

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

  - Aquí el máximo discriminante corresponde al mínimo riesgo.
- *Caso de tasa de error mínima,*
  - Es el caso de la pérdida cero-uno

$$g_i(x) = - R(\alpha_i \mid x) = -(1 - P(\omega_i \mid x))$$

  - Aquí la máxima discriminación corresponde al máximo a posteriori.

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

Funciones discriminantes

- Las funciones discriminantes no son únicas

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

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

$$g_i^{**}(x) = \ln(p(x \mid \omega_i)) + \ln(P(\omega_i))$$

- Son todas equivalentes, pues generan la misma regla

Reglas equivalentes

- Decisiones equivalentes parten el espacio de características de forma igual.
- Se definen entonces las reglas de acuerdo a las funciones discriminantes que impliquen la menor cantidad de operaciones
- Las regiones de decisión definidas son $\mathcal{R}_1, \ldots, \mathcal{R}_c$

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

Ejemplo gaussiano

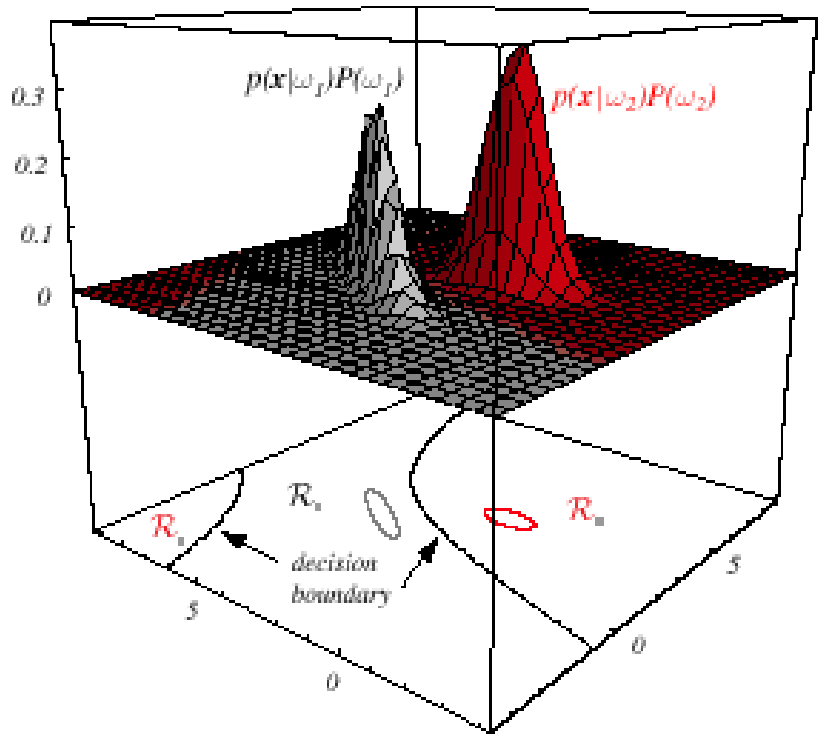

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

Caso de dos categorías

- Un clasificador "dicotómico" tiene dos funciones discriminantes g_1 y g_2

$$g_1(x) = P(\omega_1 \mid x)$$
$$g_2(x) = P(\omega_2 \mid x)$$

Decide ω_1 *si*

$$g_1(x) > g_2(x)$$
$$P(\omega_1 \mid x) > P(\omega_2 \mid x)$$

en otro caso se decide por ω_2

Caso de dos categorías

- Para este caso es común definir una única función discriminante

$$g(x) = g_1(x) - g_2(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$

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

- Otra función discriminante muy usada es

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

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

Densidad Normal Univariada

- Densidad continua y analíticamente manejable.
- Muchos patrones son asintóticamente Gaussianos si se modelan como prototipos corruptos por una gran cantidad de procesos aleatorios

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

Donde:

μ = esperanza de x

σ^2 = varianza

Densidad Normal Univariada

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

$$\sigma = \int (x-\mu)^2 p(x) dx$$

La entropía de una distribución es

$$H(p(x)) = -\int \ln(p(x)) p(x) dx$$

Puede verse que la entropía de la distribución Gaussiana es máxima sobre la clase de distribuciones con la misma media y varianza.

Densidad Normal Univariada

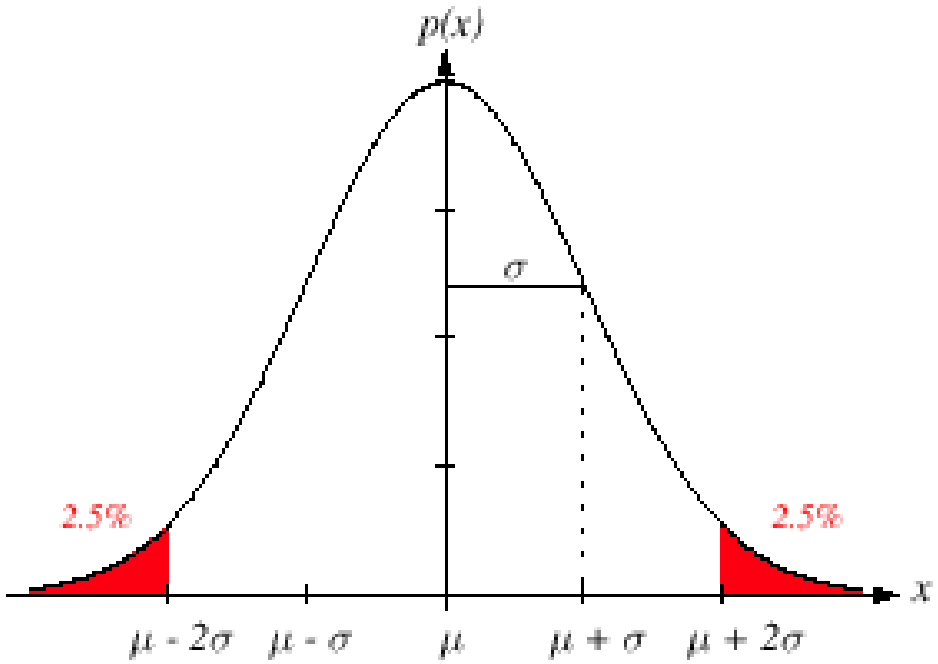

FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

Densidad Normal Multivariada

- La Gaussiana 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 respectivamente

Proyecciones

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

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

- Proyecciones de normales son normales

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

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

- Por lo cual se puede transformar una normal multivariada en una esférica, con $\Sigma = I$

$$A_w = \Phi\Lambda^{-1/2}, A_w'\Sigma A_w = I$$

$$A_w X \sim N(A_w\mu, A_w'\Sigma A_w)$$

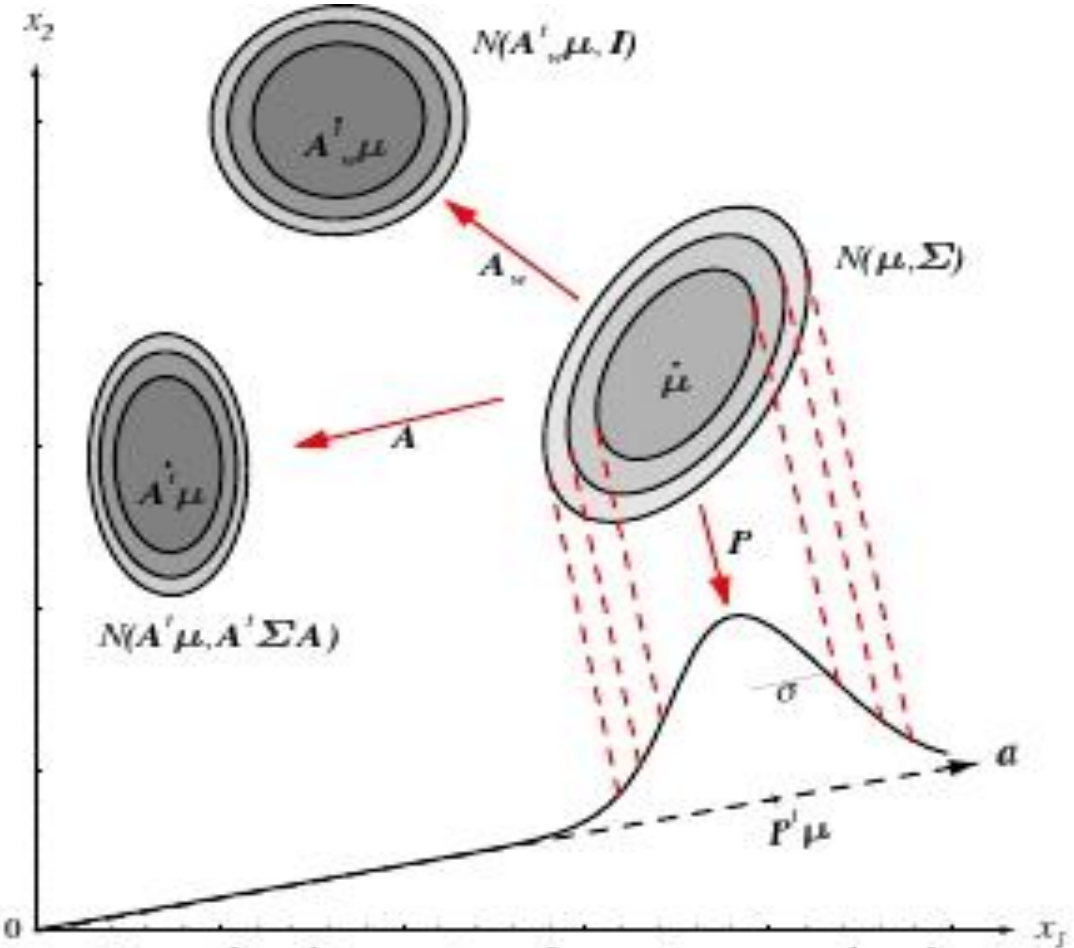

FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, **A**, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$.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

Distancia de Mahalanobis

- Distancia de Mahalanobis

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

- Contornos de densidad constante son hiper-elipsoides con distancia de Mahalanobis constante.

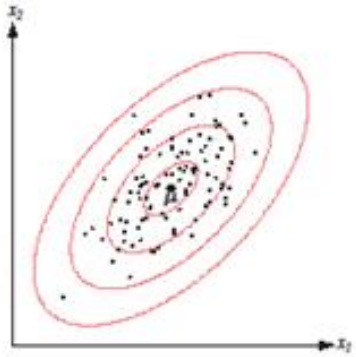

FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.