CIMPA-UCR

# Análisis Discriminante

¿Es posible predecir con antelación si un cliente que solicita un préstamo a un banco va a ser un cliente moroso?

¿Cuáles son los factores que influyen en el desarrollo de un infarto de miocardio? ¿Es posible predecir de antemano que un paciente corre un riesgo cierto de infarto?

¿Se puede predecir de antemano si un recluso que ha solicitado un permiso carcelario, huirá?

¿Se puede predecir si una empresa va a entrar en bancarrota?

¿Cuáles son las razones que llevan a un consumidor a preferir una determinada marca sobre otras existentes en el mercado?

¿Existe discriminación por razones de sexo o de raza en una empresa o en un colegio?

¿QUÉ TIENEN EN COMÚN TODOS ESTOS PROBLEMAS? ¿CÓMO RESOLVERLOS?

# Discriminación

## Técnicas:

- Análisis factorial discriminante (algebra lineal)
- Discriminación bayesiana (probabilidad)
- Discriminación cualitativa (puntaje)
- Segmentación

- Redes neuronales
- Generación de reglas
- Conjuntos aproximados

CIMPA-UCR

# Discriminación

$X$

$p$ variables explicativas

Datos nuevos

$y_1$

…

$y_k$

?

$\boldsymbol{y}$ Variable cualitativa

Ej.:

- Diagnósticos médicos
- Previsión meteorológica
- Segmentar el mercado
- Asignación de crédito bancario

## Propósito del análisis:

- Describir la clasificación
- Pronóstico de nuevos individuos

CIMPA-UCR

# Ejemplo: colestasis infantil

Problemas del hígado:

- Hepatitis: infección
- Mal formación de las vías hepáticas

Cura :

- Hepatitis: reposo
- Malformación: operar

Riesgo:

- Hepatitis + operación $\Rightarrow$ muerte!
- Malformación + reposo $\Rightarrow$ muerte!

Sin Análisis Discriminante: 50%

Usando Análisis Discriminante: 75%

CIMPA-UCR



# Ejemplo: colestasis infantil

| | | Diagnóstico | |
|---|---|---|---|
| | | Hepatitis (reposo) | Mal formación (operar) |
| Tiene | Hepatitis | OK | Muerte |
| | Mal formación | Muerte | OK |

Riesgo usando Análisis Discriminante: 25%

(vs. 50% si se tira la moneda)

# Análisis Discriminante: notaciones

$n$: número de individuos

$\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ los individuos, con pesos $p_i > 0, \quad \sum p_i = 1$

$x^1, x^2, \ldots, x^p$ variables cuantitativas explicativas

$y$ : variable cualitativa a explicar

$D = \operatorname{diag}(p_i)$: métrica de pesos en $F = R^n$ ( esp. de variables)

$M$ : métrica en $E = R^p$ ( esp. de individuos)

$N = (X, M, D)$ : nube de puntos

$$\bar{g} = \sum_{i=1}^{n} p_i \bar{x}_i = \left(\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^p\right) \in R^p : \quad \text{centro de gravedad}$$

C

# Análisis discriminante: notaciones

La variable $y$ tiene $k$ modalidades: $\rightarrow$ partición

Hay $k$ clases $c_1, c_2, \ldots, c_k$

$\overline{g}_l$ : centro de gravedad de $C_l$

$$\overline{g}_l = \frac{1}{\mu_l} \sum_{x_i \in C_l} p_i \overline{x}_i$$

$$\mu_l = \frac{1}{n} \sum_{x_i \in C_l} p_i : \text{ peso de } C_l$$

CI

# Análisis discriminante

## Descomposición de la varianza

Las $x^j$ son centradas $\Rightarrow \vec{g} = \vec{0}$

$V = X^t DX$ : matriz de varianzas - covarianzas

$W = \sum_{l=1}^{k} V_l$ con $V_l = \sum_{x_i \in C_l} p_i \left(\vec{x}_i - \vec{g}_l\right)\left(\vec{x}_i - \vec{g}_l\right)^t$ : matriz de varianzas - covarianzas internas de la clase $C_l$

$W$ : matriz de varianzas-covarianzas intra-clases

$B = \sum_{l=1}^{k} \mu_l \left(\vec{g}_l - \vec{g}\right)\left(\vec{g}_l - \vec{g}\right)^t$ : matriz de varianzas-covarianzas inter-clases

# Descomposición de la varianza

Relación de Huygens-Fisher: $V = W + B$

$$V = \sum_{i=1}^{n} p_i \left(\vec{x}_i - \vec{g}\right)\left(\vec{x}_i - \vec{g}\right)^t = \sum_{l=1}^{k} \sum_{x_i \in C_l} p_i \left(\vec{x}_i - \vec{g}_l + \vec{g}_l - \vec{g}\right)\left(\vec{x}_i - \vec{g}_l + \vec{g}_l - \vec{g}\right)^t$$

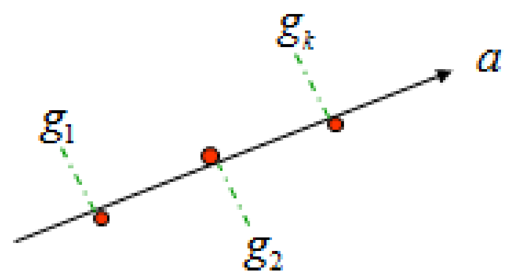
$$= \sum_{l=1}^{k} \sum_{x_i \in C_l} p_i \left(\vec{x}_i - \vec{g}_l\right)\left(\vec{x}_i - \vec{g}_l\right)^t = \sum_{l=1}^{k} V_l = W$$

$$+ \sum_{l=1}^{k} \sum_{x_i \in C_l} p_i \left(\vec{g}_l - \vec{g}\right)\left(\vec{g}_l - \vec{g}\right)^t = B$$

# Análisis discriminante: solución

Buscar $\boldsymbol{a} \in R^{p}$ tal que los grupos proyectados en $\boldsymbol{a}$ estén lo más separados posible

Tomamos $a$ tal que $\|a\| = 1$




Inercia de $N_g$ proyectada sobre $a$: $I_{\Delta_a^{\perp}}(N_g) = a^t MBMa$

$$V = B + W \Rightarrow MVM = MBM + MWM$$

$$\Rightarrow a^t MVMa = a^t MBMa + a^t MWMa$$

$$\Rightarrow 1 = \underbrace{\frac{a^t MBMa}{a^t MVMa}}_{max = \lambda} + \underbrace{\frac{a^t MWMa}{a^t MVMa}}_{min}$$

# Análisis discriminante: solución

$$\frac{\partial \lambda}{\partial a} = \frac{\left(a^t MVMa\right)\left(2BMa\right) - \left(a^t MBMa\right)\left(2VMa\right)}{\left[a^t MVMa\right]^2}$$

$$\frac{\partial \lambda}{\partial a} = 0 \Leftrightarrow \left(a^t MVMa\right)BMa = \left(a^t MBMa\right)\left(VMa\right)$$

Si $u = Ma$: $u^t Vu \cdot Bu = u^t Bu \cdot Vu$

$$\Leftrightarrow Bu = \frac{u^t Bu}{u^t Vu} Vu = \lambda Vu$$

$$\Leftrightarrow V^{-1} Bu = \lambda u$$

Solución: diagonalizar $V^{-1}B \rightarrow$

- $\lambda$ : valor propio de $V^{-1}B$
- $u$: vector propio de $V^{-1}B$

$c^1 = Xu$ : 1ª variable discriminante.

C

# Análisis Discriminante: un caso particular de un ACP

ACP de $N_g$: nube de los centros de gravedad

- puntos : $g_1, g_2, \dots, g_k$
- pesos : $\mu_1, \mu_2, \dots, \mu_k$
- métrica : $V^{-1}$

$B$ : matriz de varianzas-covarianzas de los $g_l$

$$V^{-1}Bu = \lambda u \Leftrightarrow BV^{-1}\underbrace{(Bu)}_{\upsilon} = \lambda\underbrace{(Bu)}_{\upsilon} \Leftrightarrow BV^{-1}\upsilon = \lambda\upsilon$$

C

# Análisis Discriminante: un caso particular de un ACP

Consecuencia:

- planos principales: - centros de gravedad
  - individuos (suplementarios)

- círculos de correlación: correlación entre variables discriminantes y variables originales.

Calidad del plano 1-2: $$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \ldots + \lambda_{k-1}}$$

$\lambda_j$ : poder discriminante de $C^j$

( $\lambda_j < 1$)

Nota: hay $k$-1 valores propios $\neq 0$

CIMPA-UCR

# Análisis Discriminante: un caso especial del ACP

Diagonalizar $BV^{-1}$ es equivalente a realizar un ACP de $(G, V^{-1}, \mathrm{diag}\,(\mu_k))$, donde

$$
\begin{aligned}
G &= [\overrightarrow{g^1}, \ldots, \overrightarrow{g^k}] \\
\mu_i &= \sum_{j \in C_i} p_j
\end{aligned}
$$

Obtenemos el siguiente esquema de dualidad:

$$
\begin{tikzcd}
\mathbb{R}^p \arrow[d, shift right, "V^{-1}"'] & & (\mathbb{R}^k)^* \arrow[ll, "G^{-1}"'] \\
(\mathbb{R}^p)^* \arrow[u, shift right, "B"'] \arrow[rr, "G"'] & & \mathbb{R}^k \arrow[u, "\mathrm{diag}\,(\mu_i)"']
\end{tikzcd}
$$

# Notation

- The prior probability of class $k$ is $\pi_k$, $\sum_{k=1}^{K} \pi_k = 1$.
  - $\pi_k$ is usually estimated simply by empirical frequencies of the training set

$$\hat{\pi}_k = \frac{\#\text{ samples in class } k}{\text{Total } \#\text{ of samples}}$$

- The class-conditional density of $X$ in class $G = k$ is $f_k(x)$.
- Compute the posterior probability

$$Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

- By MAP (the Bayes rule for 0-1 loss)

$$\begin{aligned} \hat{G}(x) &= \arg\max_k Pr(G = k \mid X = x) \\ &= \arg\max_k f_k(x)\pi_k \end{aligned}$$

# Class Density Estimation

- Linear and quadratic discriminant analysis: Gaussian densities.
- Mixtures of Gaussians.
- General nonparametric density estimates.
- Naive Bayes: assume each of the class densities are products of marginal densities, that is, all the variables are independent.

CI

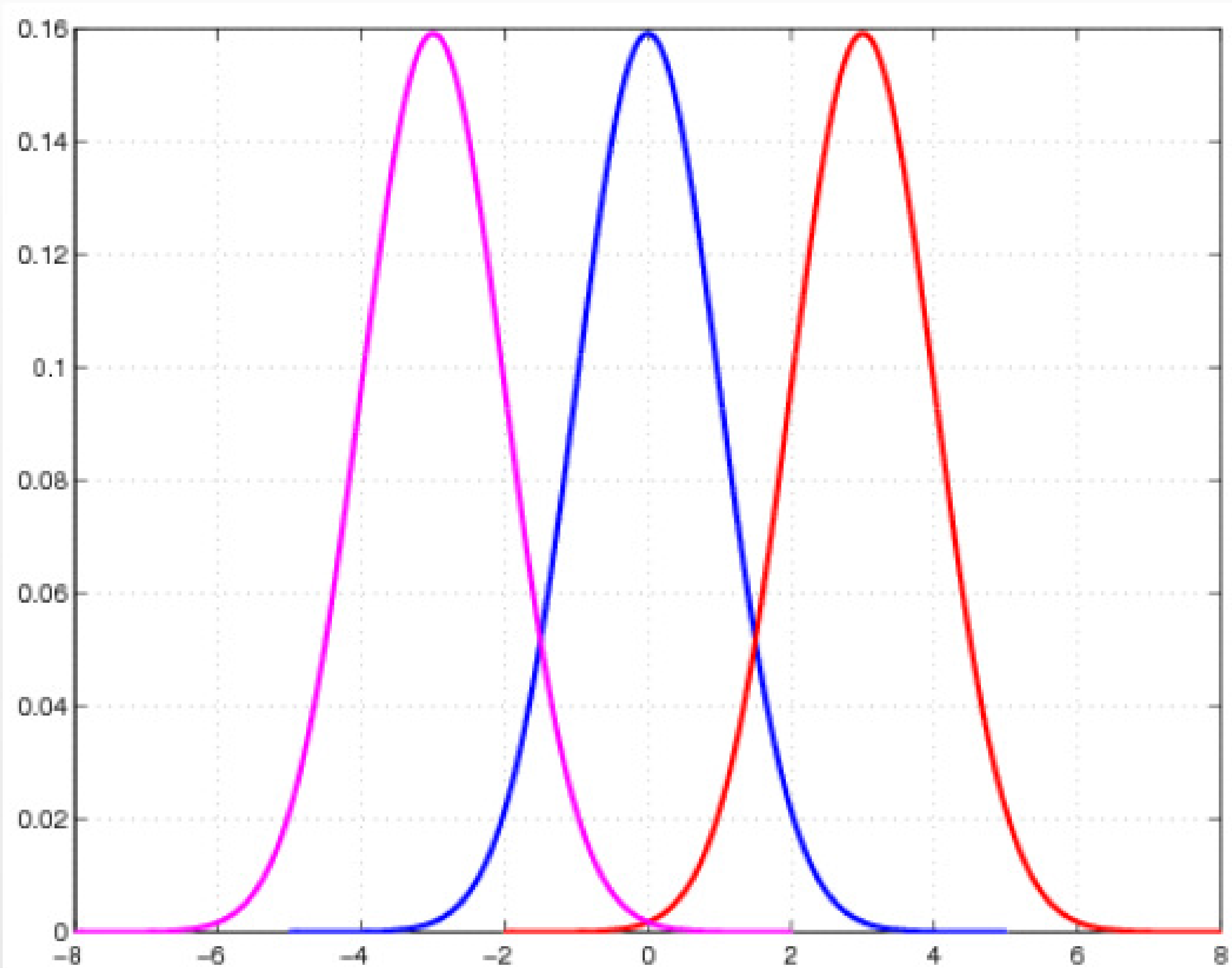
# Linear Discriminant Analysis

- Multivariate Gaussian:

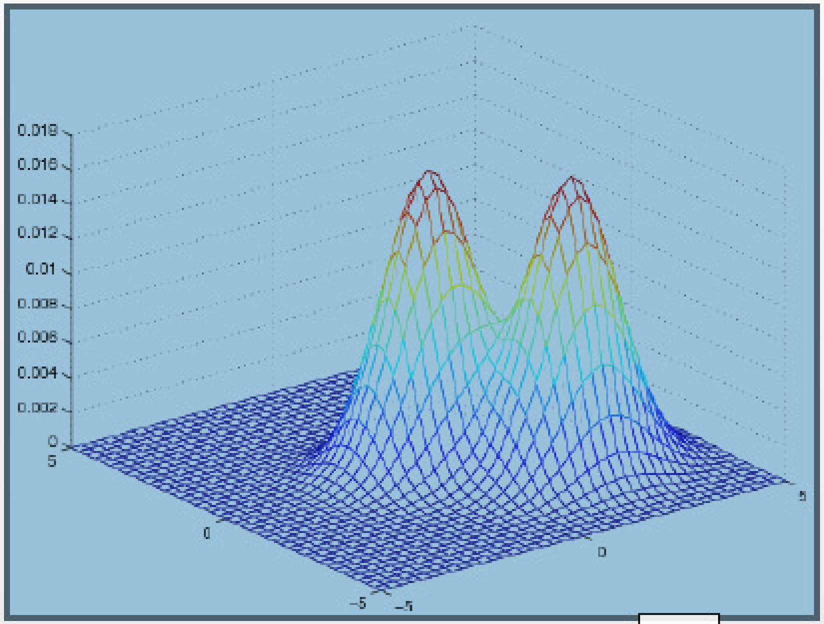
$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

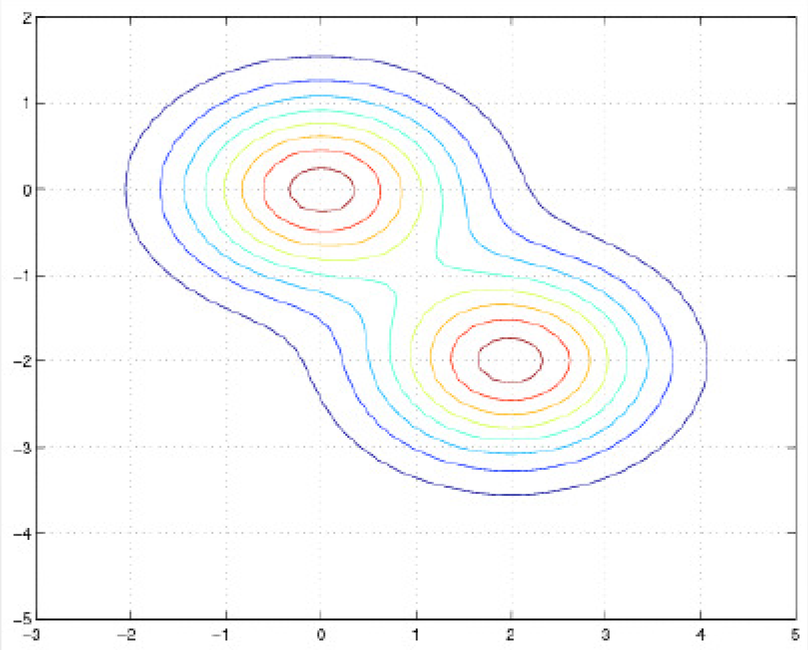
- Linear discriminant analysis (LDA): $\Sigma_k = \Sigma$, $\forall k$.
- The Gaussian distributions are shifted versions of each other.

# Análisis Discriminante



CIMPA-UCR


0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
5
0
-5
-5
0
5


2
1
0
-1
-2
-3
-4
-5
-3
-2
-1
0
1
2
3
4
5

C

- Optimal classification

$$
\begin{aligned}
\hat{G}(x) &= \arg\max_k Pr(G = k \mid X = x) \\
&= \arg\max_k f_k(x)\pi_k = \arg\max_k \log(f_k(x)\pi_k) \\
&= \arg\max_k \Big[-\log((2\pi)^{p/2}|\Sigma|^{1/2}) \\
&\qquad -\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \log(\pi_k)\Big] \\
&= \arg\max_k \Big[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \log(\pi_k)\Big]
\end{aligned}
$$

Note

$$
-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{1}{2}x^T\Sigma^{-1}x
$$

C

## To sum up

$$\hat{G}(x) = \arg\max_k \left[ x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \right]$$

- Define the *linear discriminant function*

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \ .$$

  Then

$$\hat{G}(x) = \arg\max_k \delta_k(x) \ .$$

- The decision boundary between class $k$ and $l$ is:

$$\{x : \delta_k(x) = \delta_l(x)\} \ .$$

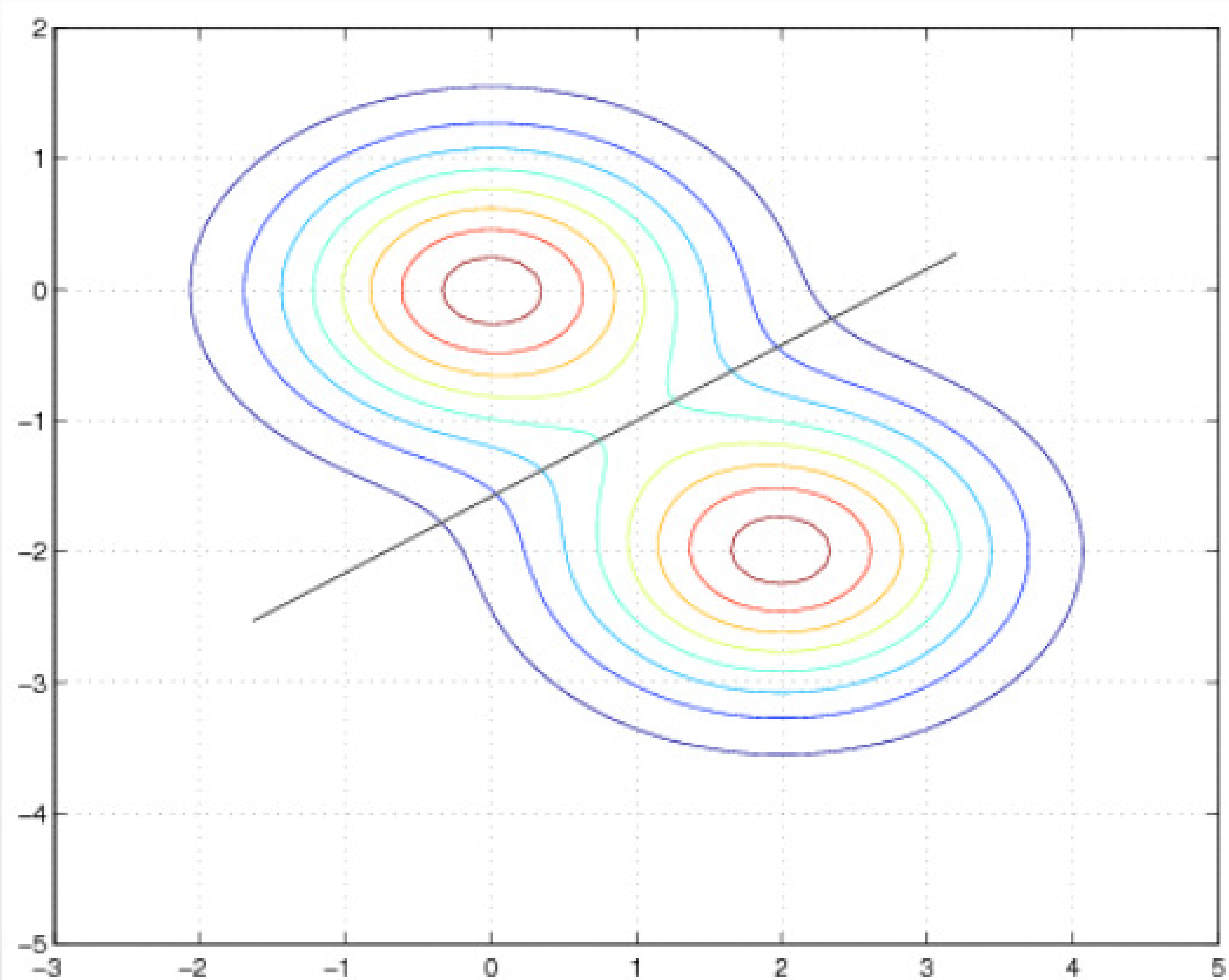
  Or equivalently the following holds

$$\log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) = 0.$$

- Binary classification ($k = 1$, $l = 2$):
  - Define $a_0 = \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$.
  - Define $(a_1, a_2, \ldots, a_p)^T = \Sigma^{-1}(\mu_1 - \mu_2)$.
  - Classify to class 1 if $a_0 + \sum_{j=1}^{p} a_j x_j > 0$; to class 2 otherwise.
  - An example:
    - $\pi_1 = \pi_2 = 0.5$.
    - $\mu_1 = (0, 0)^T$, $\mu_2 = (2, -2)^T$.
    - $\Sigma = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 0.5625 \end{pmatrix}$.
    - Decision boundary:

$$5.56 - 2.00x_1 + 3.56x_2 = 0.0 .$$

# Análisis Discriminante



CIMPA-UCR

# Estimate Gaussian Distributions

- In practice, we need to estimate the Gaussian distribution.
- $\hat{\pi}_k = N_k/N$, where $N_k$ is the number of class-$k$ samples.
- $\hat{\mu}_k = \sum_{g_i=k} x^{(i)}/N_k$ .
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x^{(i)} - \hat{\mu}_k)(x^{(i)} - \hat{\mu}_k)^T/(N-K)$.
- Note that $x^{(i)}$ denotes the $i$th sample vector.

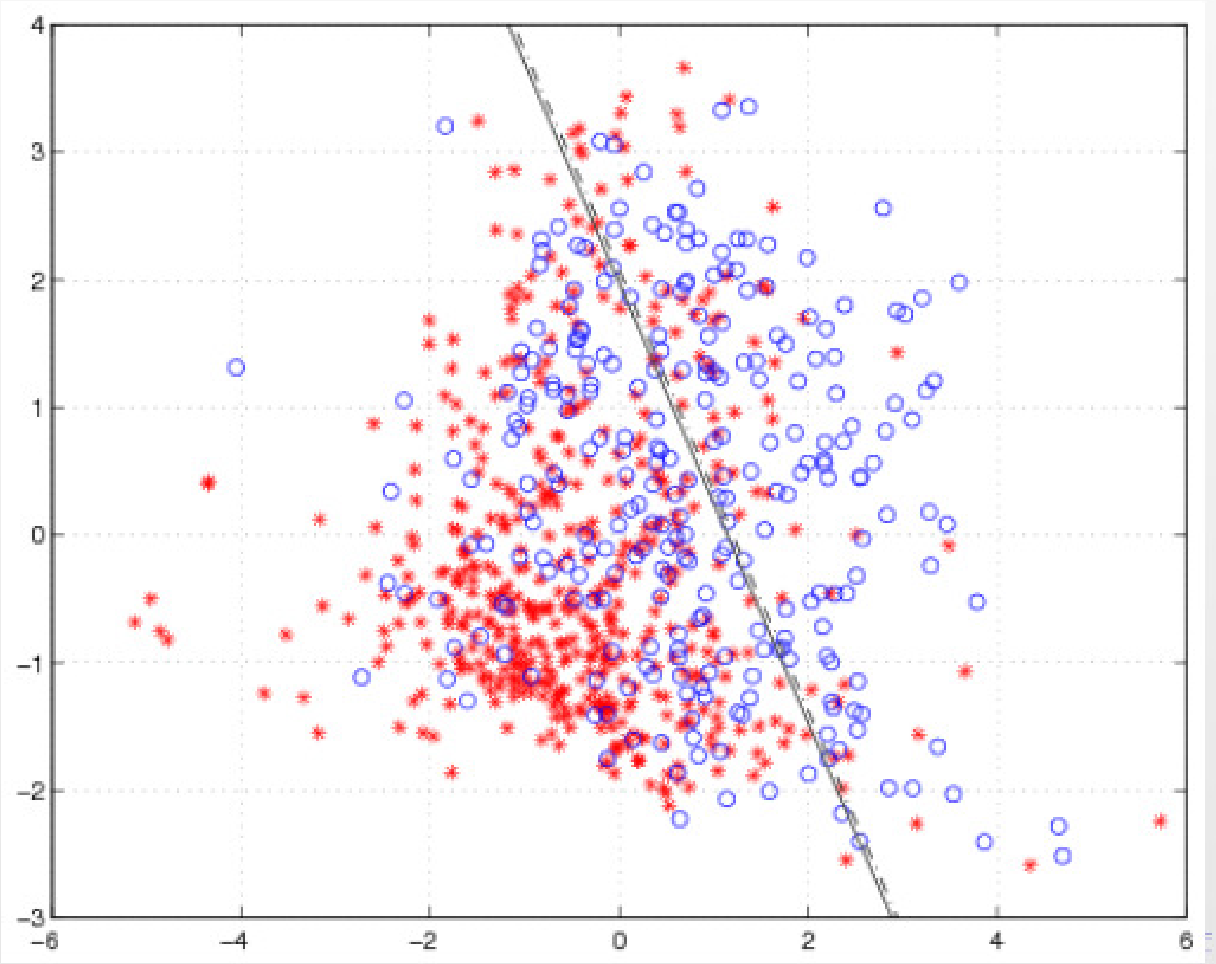
# Diabetes Data Set

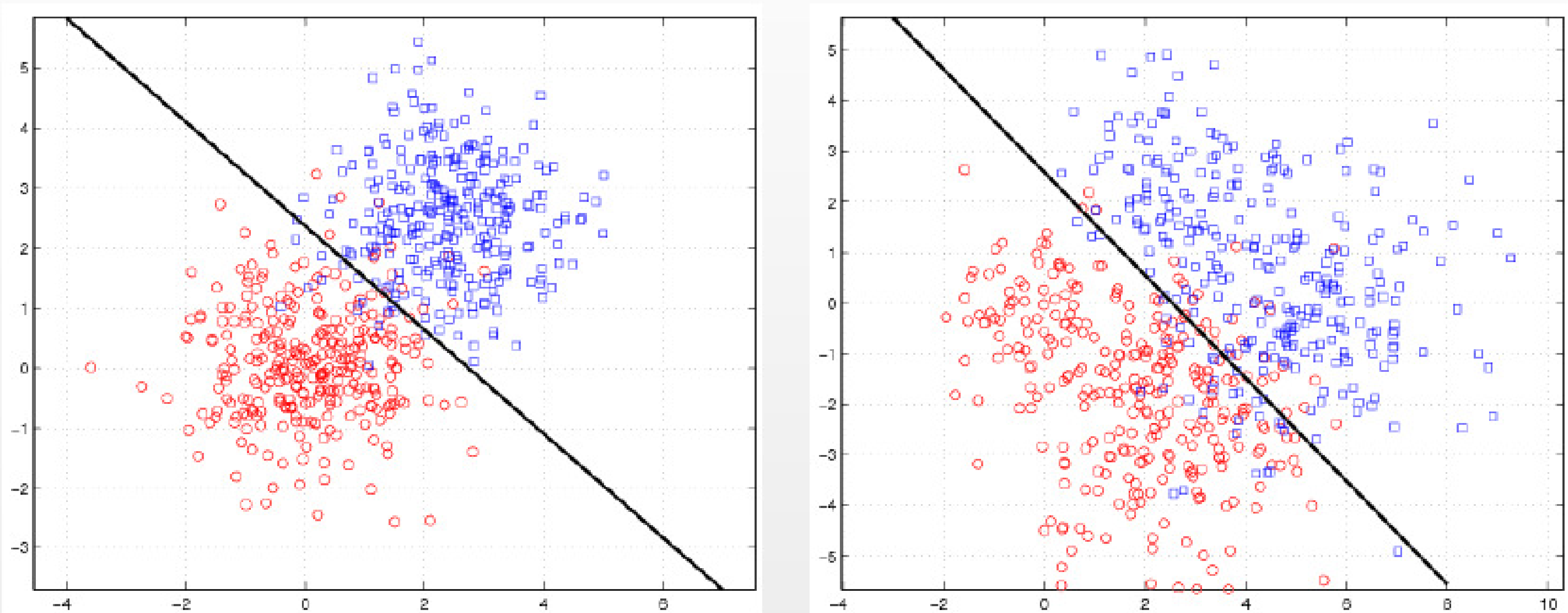
- Two input variables computed from the principal components of the original 8 variables.
- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$.
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$.
- $\hat{\Sigma} = \begin{pmatrix} 1.7925 & -0.1461 \\ -0.1461 & 1.6634 \end{pmatrix}$
- Classification rule:

$$\begin{aligned} \hat{G}(x) &= \begin{cases} 1 & 0.7748 - 0.6771x_1 - 0.3929x_2 \geq 0 \\ 2 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & 1.1443 - x_1 - 0.5802x_2 \geq 0 \\ 2 & \text{otherwise} \end{cases} \end{aligned}$$

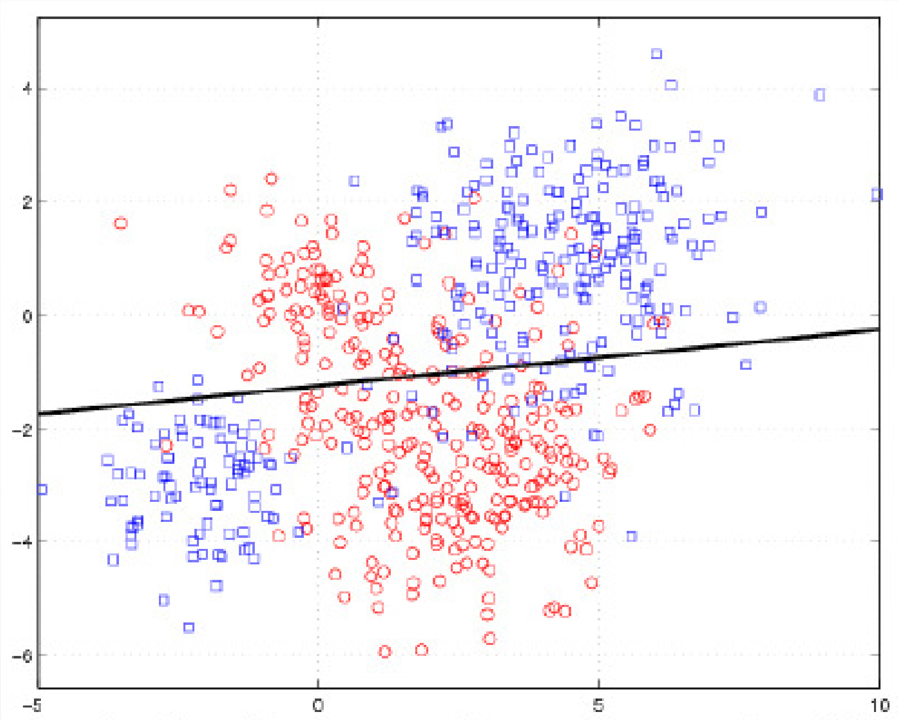
CIMP

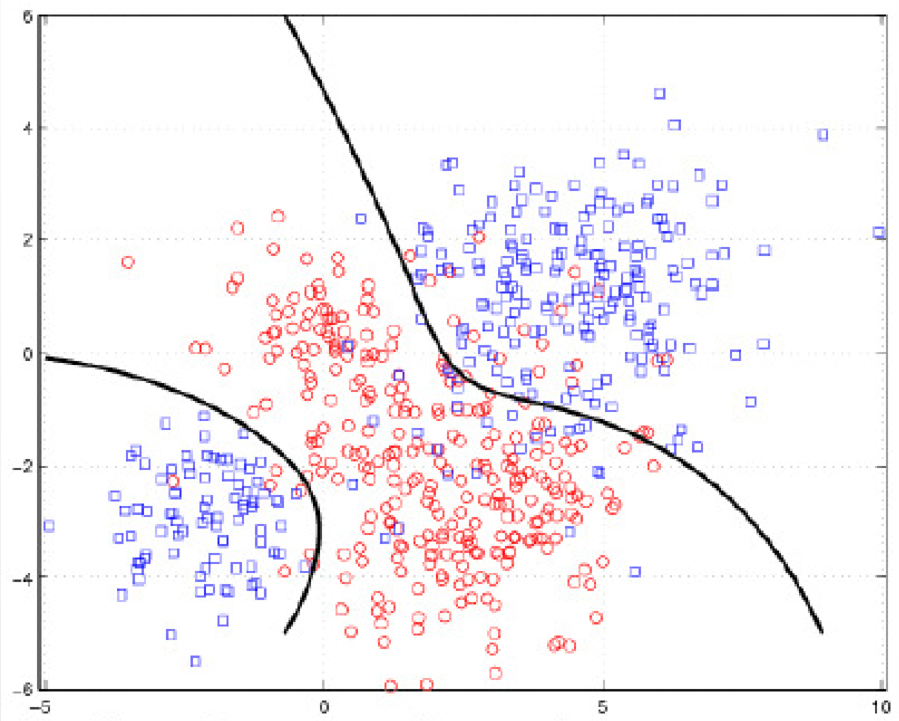
The scatter plot follows. Without diabetes: stars (class 1), with diabetes: circles (class 2). Solid line: classification boundary obtained by LDA. Dash dot line: boundary obtained by linear regression of indicator matrix.






LDA applied to simulated data sets. Left: The true within class densities are Gaussian with identical covariance matrices across classes. Right: The true within class densities are mixtures of two Gaussians.




Left: Decision boundaries by LDA. Right: Decision boundaries obtained by modeling each class by a mixture of two Gaussians.

CI

# Quadratic Discriminant Analysis (QDA)

- Estimate the covariance matrix $\Sigma_k$ separately for each class $k$, $k = 1, 2, ..., K$.
- *Quadratic discriminant function*:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k \ .$$

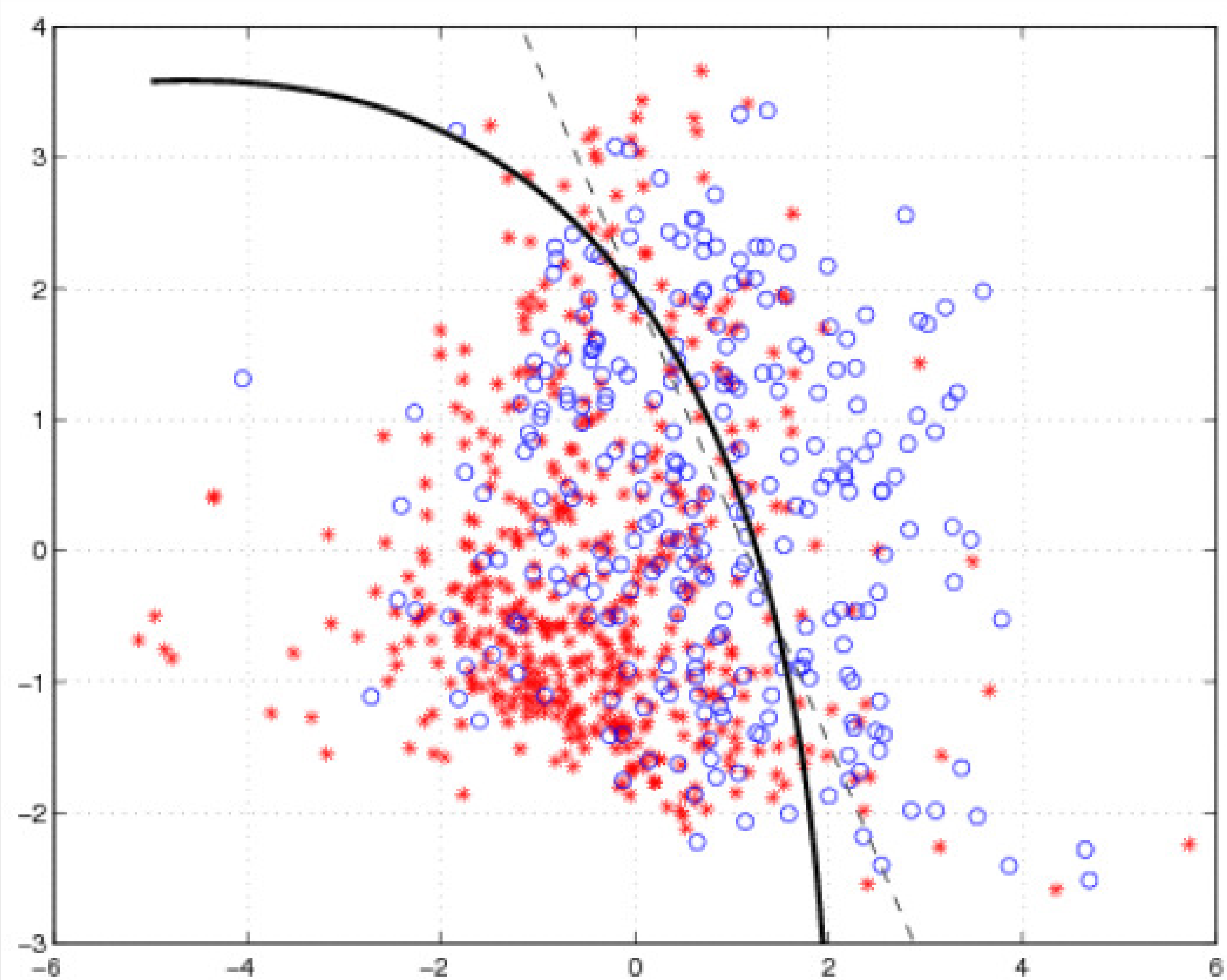
- Classification rule:

$$\hat{G}(x) = \arg\max_k \delta_k(x) \ .$$

- Decision boundaries are quadratic equations in $x$.
- QDA fits the data better than LDA, but has more parameters to estimate.

# Diabetes Data Set

- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$.
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$.
- $\hat{\Sigma}_1 = \begin{pmatrix} 1.6769 & -0.0461 \\ -0.0461 & 1.5964 \end{pmatrix}$
- $\hat{\Sigma}_2 = \begin{pmatrix} 2.0087 & -0.3330 \\ -0.3330 & 1.7887 \end{pmatrix}$


4
3
2
1
0
−1
−2
−3
−6
−4
−2
0
2
4
6

# LDA on Expanded Basis

- Expand input space to include $X_1X_2$, $X_1^2$, and $X_2^2$.
- Input is five dimensional: $X = (X_1, X_2, X_1X_2, X_1^2, X_2^2)$.
-

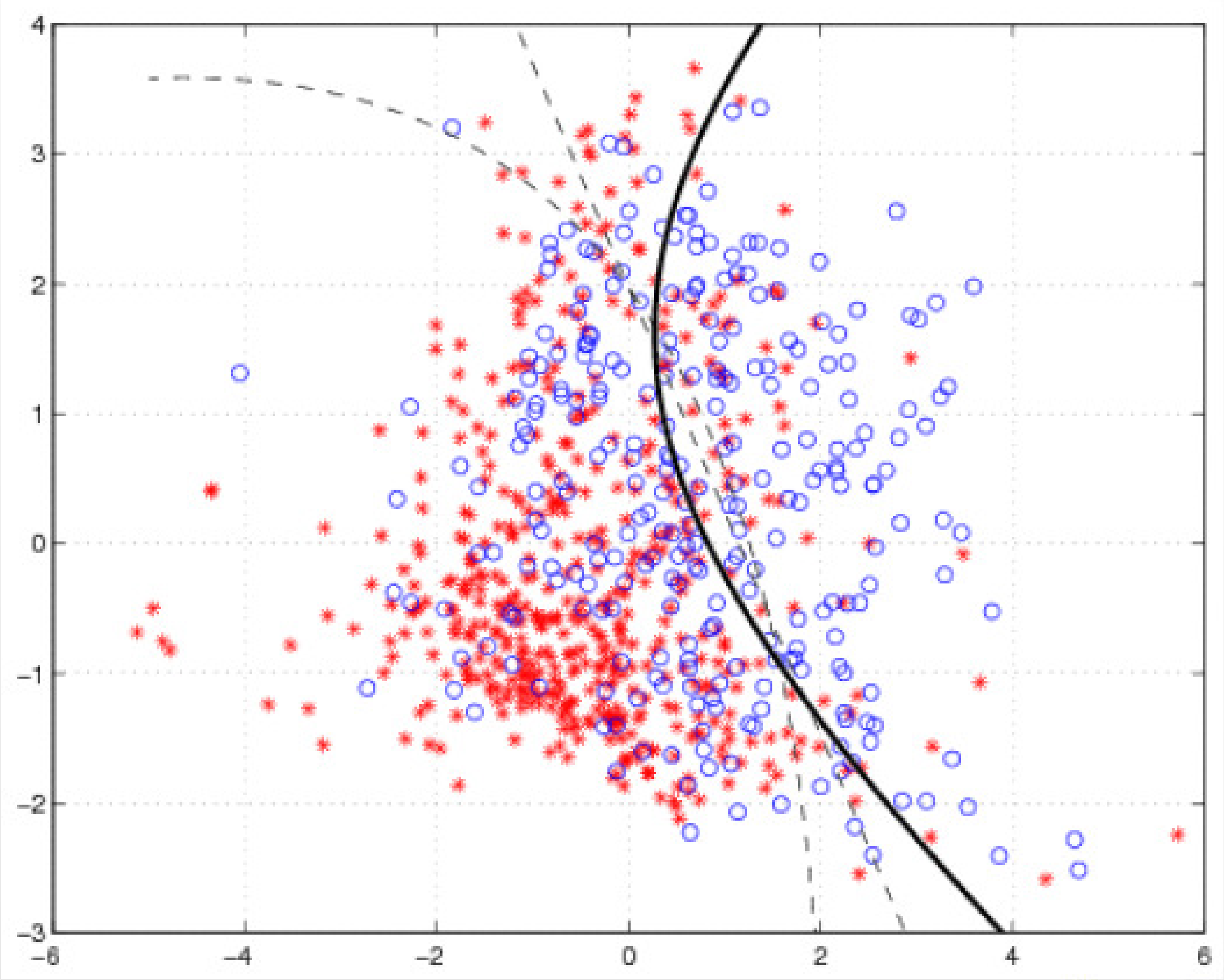
$$\hat{\mu}_1 = \begin{pmatrix} -0.4035 \\ -0.1935 \\ 0.0321 \\ 1.8363 \\ 1.6306 \end{pmatrix} \qquad \hat{\mu}_2 = \begin{pmatrix} 0.7528 \\ 0.3611 \\ -0.0599 \\ 2.5680 \\ 1.9124 \end{pmatrix}$$

-

$$\hat{\Sigma} = \begin{pmatrix} 1.7925 & -0.1461 & -0.6254 & 0.3548 & 0.5215 \\ -0.1461 & 1.6634 & 0.6073 & -0.7421 & 1.2193 \\ -0.6254 & 0.6073 & 3.5751 & -1.1118 & -0.5044 \\ 0.3548 & -0.7421 & -1.1118 & 12.3355 & -0.0957 \\ 0.5215 & 1.2193 & -0.5044 & -0.0957 & 4.4650 \end{pmatrix}$$

CIM

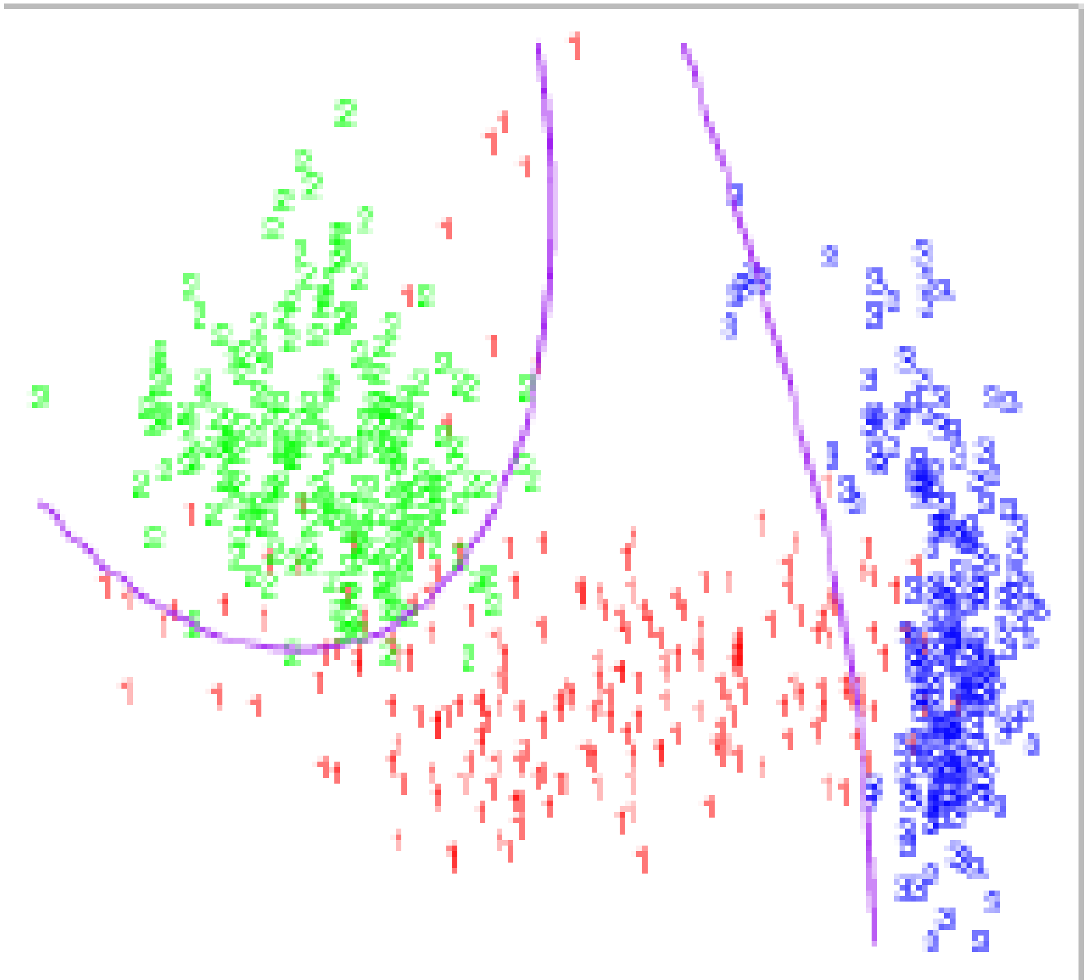
Classification boundaries obtained by LDA using the expanded input space $X_1$, $X_2$, $X_1X_2$, $X_1^2$, $X_2^2$. Boundaries obtained by LDA and QDA using the original input are shown for comparison.



- Classification boundary:

$$0.651 - 0.728x_1 - 0.552x_2 - 0.006x_1x_2 - 0.071x_1^2 + 0.170x_2^2 = 0\ .$$

- If the linear function on the right hand side is non-negative, classify as 1; otherwise 2.



• Quadratic boundaries obtained using QDA in the five dimensional space $x_1, x_2, x_{12}, x_1^2, x_2^2)$

C

# Análisis Discriminante (con R)

```
# Se carga la librería MASS
library(MASS)

# Se hace un análisis discriminante lineal
dis <- lda(Tipo ~ Longitud + Anchura + Altura + Altura.Cara +
Anchura.Cara, data=Tibet, prior=c(0.5,0.5))

dis
Call:
lda(Tipo ~ Longitud + Anchura + Altura + Altura.Cara + Anchura.Cara
    data = Tibet, prior = c(0.5, 0.5))

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
  Longitud  Anchura   Altura Altura.Cara Anchura.Cara
1 174.8235 139.3529 132.0000    69.82353     130.3529
2 185.7333 138.7333 134.7667    76.46667     137.5000
```

```
Coefficients of linear discriminants:
                    LD1
Longitud       0.047726591
Anchura       -0.083247929
Altura        -0.002795841
Altura.Cara    0.094695000
Anchura.Cara   0.094809401


# Se consideran las medidas de dos nuevos craneos
nuevosdatos <-
rbind(c(171,140.5,127.0,69.5,137.0),c(179.0,132.0,140.0,72.0,138.5))

# Asigno a los dos nuevos datos los nombres de las variables
colnames(nuevosdatos) <- colnames(Tibet[,-6])

nuevosdatos <- data.frame(nuevosdatos)

# Se predice el grupo de pertenencia de los nuevos datos
predict(dis,newdata=nuevosdatos)$class

[1] 1 2
Levels: 1 2

$posterior
          1         2
1 0.7545066 0.2454934
2 0.1741016 0.8258984
```