Estadística Aplicada III

Clasificación y Análisis discriminante

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 10: 17/19 de octubre de 2018

Introducción

Propósito de la discriminación y clasificación

- La discriminación intenta *separar* elementos en diferentes grupos predefinidos.
- La clasificación intenta ubicar nuevos objetos a diferentes grupos predefinidos.
- Hay dos objetivos típicos de la discriminación y clasificación:
  1. Descriptivo: encontrar *discriminantes* que mejor separen los grupos.
  2. Ubicación: A través de reglas, poner los nuevos elementos en los grupos vía los *discriminantes*.

Orígenes y primeras aplicaciones

Figura: Ronald Avery Fisher joven

- El modelo original de análisis discriminante fue desarrollado por Ronald Aymler Fisher en 1936 en el artículo: The use of multiple measurements in taxonomic problems, utilizando el conjunto de datos Iris de Edgar Anderson.
- Una de las primeras aplicaciones posteriores consistió en clasificar los restos de un cráneo descubierto en una excavación como humano, utilizando la distribución de medidas físicas para los cráneos humanos y los de antropoides.

Problema general de clasificación I

- Consideren un vector de atributos observables $\mathbf{x}_i \in \Omega \subset \mathbb{R}^p$ para un objeto o individuo i que se sabe que debe pertenecer a uno de g grupos o poblaciones Π_i posibles.
- Ω es el espacio muestral, y se puede representar como una partición de regiones R_i correspondientes a las diferentes poblaciones:

$$\Omega = R_1 \cup R_2 \cup \cdots \cup R_g \ni R_i \cap R_j = \emptyset \quad \forall i \neq j$$

Problema general de clasificación

Dada una observación i con atributos $\mathbf{x}_i$, encontrar una regla de clasificación que defina regiones R_i, usando la información $\mathbf{x}_i$, *junto con lo que sabemos de las poblaciones* Π_j, para ubicar la población de pertenencia de i, con la máxima precisión posible.

- Este pronóstico puede tener un margen de error, usualmente vinculado a que los atributos pueden darse en más de una región R_i.
- Lo anterior supone que las poblaciones Π_j son conocidas de antemano.[1]

Problema general de clasificación II

- El problema de clasificación usualmente se considera como un ejemplo de *aprendizaje supervisado*, en donde un conjunto de *inputs* (o variables independientes o predictores) $\mathbf{x}$, tienen influencia sobre uno o más *outputs* (o variables dependientes, o respuestas) Π_j. Entonces los *inputs* son usados para predecir el valor de los *outputs*[2].
- Las reglas de clasificación o asignación se desarrollan "aprendiendo" de la muestra $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de la que sabemos de qué población proviene. ¿Cómo sabemos esto de algunas observaciones y no de otras? Por diferentes razones:
  - por conocimiento incompleto del desempeño futuro.
  - conocimiento completo requiere la destrucción total del producto.
  - Información costosa o no disponible.

[1] cuando las poblaciones no se conocen de antemano, el análisis se llama *análisis de conglomerados (cluster analysis)*.

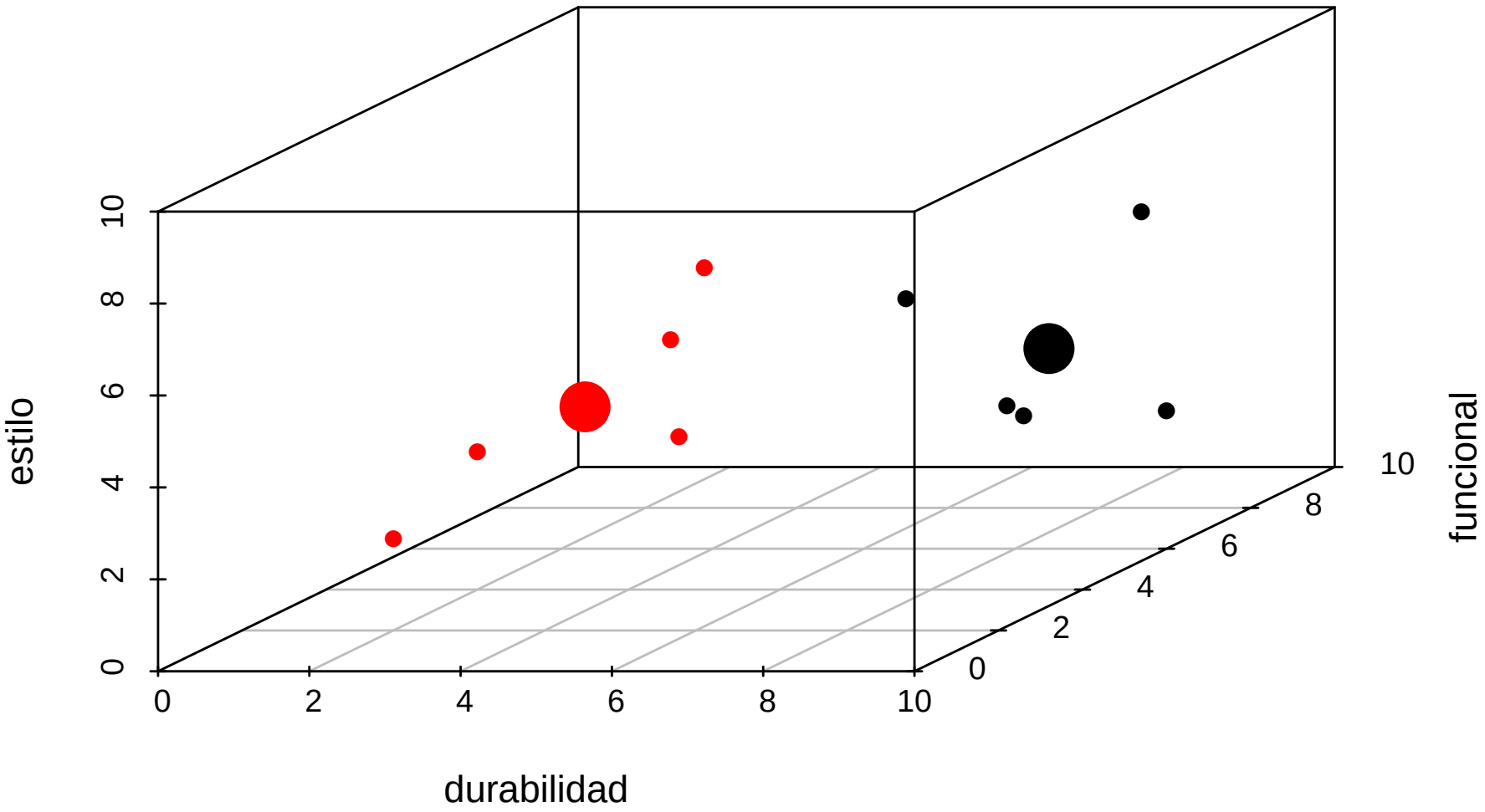
[2] El lenguaje de *inputs* y *outputs* es mucho más común en *Machine Learning*.

Ejemplo I

Supongan que Ben & Frank quiere saber si ciertos armazones serán comercialmente aceptables. Hacen un estudio de mercado para evaluar sus nuevos armazones en tres características: durabilidad, desempeño y estilo, con una escala que va de 0 (muy mal) a 10 (excelente). Los resultados de las tres variables se muestran a continuación con su representación gráfica. Los puntos grandes representan las medias de cada grupo de compra.
Se desea conocer qué características de un nuevo producto son útiles para diferenciar a los compradores de los no-compradores.

```
library(scatterplot3d)
ByF <- data.frame(
        durabilidad = c(8,6,10,9,7,5,3,4,2,2),
        funcional = c(9,7,6,4,8,4,7,5,4,2),
        estilo = c(6,5,3,4,2,7,2,5,3,2),
        compra  = c(1,1,1,1,1,2,2,2,2,2)
)
graf <- scatterplot3d(ByF[,1:3], color = ByF$compra, pch = 16, xlim = c(0,10), ylim = c(0,10), zlim = c(0,10))
graf$points3d(rbind(colMeans(ByF[ByF$compra == 1, -4]),
colMeans(ByF[ByF$compra == 2, -4])),pch = 16,col = c(1,2), cex = 3)
```

Ejemplo II



Ejemplos de aplicaciones I

- **Diagnóstico clínico**: cada Π_i corresponde a una enfermedad. Los atributos $\mathbf{x}$ pueden ser los resultados de varios exámenes médicos para cada paciente. El problema de clasificación consiste en diagnosticar la enfermedad sobre la base de resultados médicos obtenidos.
- **Bancarrota de una institución financiera**: Se pueden considerar dos poblaciones: Π_1 son IF's que caerán en bancarrota en los próximos 12 meses, y Π_2 las que no están en esa situación. Entonces $\mathbf{x}$ pueden ser los predictores de default, como número de auditorías, capital social respecto a capital total, etc. (Este modelo fue desarrollado por Edward Altman).
- **Identificación de especies**: Π_i representa los taxones previamente definidos. $\mathbf{x}$ pueden ser las medidas morfológicas de la planta que un investigador está recolectando. El problema del investigador consiste en clasificar los especímenes que va encontrando en su respectivo taxón.
- **Credit Scoring**: El problema consiste en clasificar a un cliente en dos posibles poblaciones: Π_1 son los que pagan a tiempo y Π_2 son los que no pagan. Los atributos pueden ser características demográficas (edad, género, grupo social, etc), económicas (número de dependientes, ingreso, tipo de trabajo, etc), entre otras.

Ejemplos de aplicaciones II

- **Pattern Recognition**: En ingeniería, el problema de discriminación se estudia bajo el nombre anterior, para diseñar máquinas y sistemas capaces de clasificar de manera automática. Ejemplos incluyen los ATM's, las máquinas para pagar el estacionamiento, lectoclasificadoras, etc.

Modelo

Reglas de clasificación óptimas

Consideremos inicialmente el caso de $g = 2$ poblaciones, para simplificar el análisis y ganar intuición.

- Una regla de clasificación debe contemplar los siguientes elementos:
  - La información inicial o *a priori* disponible a través de las distribuciones π_i de cada una de las poblaciones Π_i.
  - una función de costo asociada a una clasificación errónea. Usualmente tenemos un caso similar al de los errores tipo I y II:

		Clasificación	
		Π_1	Π_2
Pertenencia Real	Π_1	✓	☹
	Π_2	☹	✓

  - Una opción para sustituir la función de costo, son las probabilidades de clasificación errónea.
- Entonces una regla de clasificación óptima debería tomar en cuenta la información inicial y reducir el costo total o la probabilidad total de clasificación errónea.

Probabilidades de clasificación I

Def (Probabilidades asociadas a un problema de clasificación)

- En la población Π_i, los atributos tienen una densidad $f_i(\mathbf{x})$.
- Sea $P(i|j)$ la probabilidad condicional de clasificar un elemento en la población i, dado que pertenece a la población j:

$$P(i|j) = P(\mathbf{x} \in R_i|\Pi_j) = \int_{R_i} f_j(\mathbf{x})d\mathbf{x}$$

  Noten que $\sum_{i=1}^{g} P(i|j) = 1$ para cada j.
- La probabilidad inicial o a priori de pertenecer a la población i es π_i, $\sum_{i=1}^{g} \pi_i = 1$.

- Podemos pensar que la distribución de las observaciones se puede considerar como una *mezcla* de las densidades en cada una de las poblaciones:

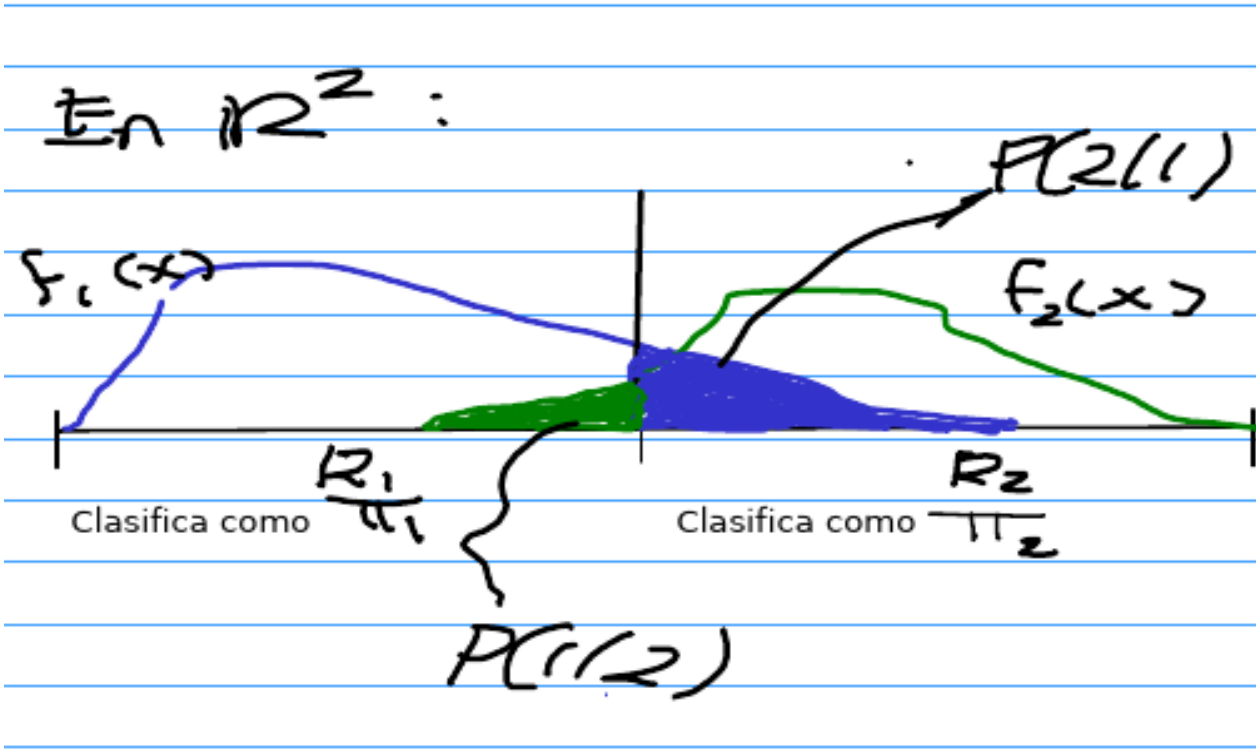
$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x}).$$

Probabilidades de clasificación II

- Las probabilidades de clasificación correcta o incorrecta, se pueden expresar en términos de las probabilidades de clasificación errónea y la distribución inicial:

$$
\begin{aligned}
P(\mathbf{x} \text{ clasificada correctamente en } \Pi_j) &= P(\mathbf{x} \in R_j \cap \mathbf{x} \in \Pi_j) \\
&= P(\mathbf{x} \in R_j | \Pi_j)\pi_j \\
&= P(j|j)\pi_j \\
P(\mathbf{x} \text{ clasificada incorrectamente en } \Pi_j) &= P(\mathbf{x} \in R_j \cap \mathbf{x} \in \Pi_i) \\
&= P(\mathbf{x} \in R_j | \Pi_i)\pi_i \\
&= P(j|i)\pi_i
\end{aligned}
$$

Probabilidades de clasificación III



Costos de clasificación errónea I

- Los costos de clasificación pueden ser muy relevantes y diferentes según el error cometido:
  - No es el mismo costo clasificar erróneamente a un paciente con cáncer en el grupo de los pacientes sanos, que clasificar al que no tiene cáncer en el grupo de los enfermos.
  - No es el mismo costo clasificar a un cliente como buen pagador cuando no paga (costo real) que clasificarlo como mal pagador cuando sí paga (costo de pérdida de oportunidad).
- Los costos de clasificación se especifican en la matriz de costos que se mencionó antes:

		Clasificación	
		Π_1	Π_2
Pertenencia Real	Π_1	0	$c(2\|1)$
	Π_2	$c(1\|2)$	0

Costo esperado de clasificación errónea

Para cualquier regla de clasificación, el costo total esperado de clasificación errónea (ECM) está dado por:

$$ECM = c(2|1)P(2|1)\pi_1 + c(1|2)P(1|2)\pi_2$$

Costos de clasificación errónea II

- Como alternativa al costo, se puede usar la probabilidad total de clasificación errónea como un criterio de valuación, dada por:

$$TPM = P(2|1)\pi_1 + P(1|2)\pi_2$$

Criterios para obtener las reglas óptimas de clasificación I

- Se consideran tres criterios diferentes para obtener las regiones óptimas de clasificación:
  - Costo total esperado de clasificación errónea mínimo (ECM),
  - Probabilidad total de clasificación errónea mínima (TPM), y
  - Maximiazación de la distribución posterior.
- Una regla de clasificación que sea razonable, debería tener un ECM (TPM) tan pequeño como sea posible, así que si $\hat{\Pi}_a$ representa una regla de clasificación, lo que permite definir un órden en el conjunto de las reglas de clasificación:

Criterio ECM (TPM)

Se prefiere la regla $\hat{\Pi}_a$ a $\hat{\Pi}_b$ si su costo total (probabilidad total de clasificación errónea) esperado(a) es menor,

$$ECM(\hat{\Pi}_a) < ECM(\hat{\Pi}_b)$$

$$(TPM(\hat{\Pi}_a) < TPM(\hat{\Pi}_b))$$

Criterios para obtener las reglas óptimas de clasificación II

- En el caso de la probabilidad posterior, se utilizar la probabilidad posterior de clasificación una vez que se observa un nuevo caso $\mathbf{x}_0$

Criterio de probabilidad posterior máxima

Se prefiere la regla $\hat{\Pi}_a$: asigna $\mathbf{x}_0$ a Π_1 a la regla $\hat{\Pi}_b$: asigna $\mathbf{x}_0$ a Π_j, $j \neq 1$ siempre que

$$P(\Pi_1|\mathbf{x}_0) > P(\Pi_j, j \neq 1|\mathbf{x}_0).$$

Criterio ECM y TPM I

- El siguiente resultado define cuál es la regla óptima de clasificación que minimiza el costo.

Teorema (Regiones que minimizan el ECM)

Las regiones R_1 y R_2 que minimizan el ECM están definidas por los siguientes conjuntos:

$$R_1 = \left\{ \mathbf{x} \in \Omega \mid \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right\}$$

y

$$R_2 = \Omega \setminus R_1 = \left\{ \mathbf{x} \in \Omega \mid \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right\}$$

En los casos en donde los costos son iguales o las probabilidades iniciales son iguales, las expresiones se reducen de manera directa.

Demostración.

Criterio ECM y TPM II

Utilizando las definiciones dadas previamente,

$$
\begin{aligned}
ECM &= c(2|1)P(2|1)\pi_1 + c(1|2)P(1|2)\pi_2 && (1)\\
&= c(2|1)\pi_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + c(1|2)\pi_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} && (2)\\
&= c(2|1)\pi_1 \left(1 - \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x}\right) + c(1|2)\pi_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} && (3)\\
&= c(2|1)\pi_1 + \int_{R_1} [c(1|2)\pi_2 f_2(\mathbf{x}) - c(2|1)\pi_1 f_1(\mathbf{x})] d\mathbf{x} && (4)
\end{aligned}
$$

La sustitución en (3) es porque $R_1 \cup R_2 = \Omega$. Como los costos y las probabilidades son positivos, entonces el ECM se minimiza cuando $c(1|2)\pi_2 f_2(\mathbf{x}) - c(2|1)\pi_1 f_1(\mathbf{x}) \leq 0$. Entonces, definimos R_1 como el conjunto de los valores de $\mathbf{x}$ tales que

$$c(1|2)\pi_2 f_2(\mathbf{x}) \leq c(2|1)\pi_1 f_1(\mathbf{x})$$

y reacomodando términos,

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)\pi_2}{c(2|1)\pi_1}$$

Criterio ECM y TPM III

como se quería demostrar.

□

Criterio de probabilidad total de clasificación errónea (TPM) mínima

Se prefiere $\hat{\Pi}_a$ a $\hat{\Pi}_b$ si su TPM es menor,

$$TPM(\hat{\Pi}_a) < TPM(\hat{\Pi}_b)$$

- En este caso, el problema es equivalente a minimizar el ECM cuando $c(1|2) = c(2|1)$. (Tarea)

Criterio basado en maximizar la distribución posterior I

- En este criterio, se busca clasificar a una nueva observación $\mathbf{x}_0$ a la población que obtenga la probabilidad posterior $P(\Pi_i|\mathbf{x}_0)$ mayor. De acuerdo al teorema de Bayes,

$$\begin{aligned} P(\Pi_1|\mathbf{x}_0) &= \frac{P(\mathbf{x}_0|\Pi_1)\pi_1}{P(\mathbf{x}_0|\Pi_1)\pi_1 + P(\mathbf{x}_0|\Pi_2)\pi_2} \\ &= \frac{f_1(\mathbf{x}_0)\pi_1}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2} \\ P(\Pi_2|\mathbf{x}_0) &= 1 - P(\Pi_1|\mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)\pi_2}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2} \end{aligned}$$

- Entonces la regla es: Clasificar $\mathbf{x}_0 \in \Pi_1$ si $P(\Pi_1|\mathbf{x}_0) > P(\Pi_2|\mathbf{x}_0)$.
- La solución matemática de este problema es equivalente a la solución óptima del ECM con costos $c(1|2) = c(2|1)$ (Tarea).

Casos con distribución normal

Caso normal multivariado, poblaciones con varianza común I

- Ahora consideremos el caso donde se tiene un vector $\mathbf{x}$ con p atributos que tiene una distribución normal multivariada $f_1 \sim \mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ cuando pertenece a Π_1 o $f \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ cuando pertenece a Π_2:

$$f_i(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)'\right\}$$

- Para simplificar la notación, consideremos $D_i^2(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ la distancia de Mahalanobis de la observación $\mathbf{x}$ a la media $\boldsymbol{\mu}_i$.
- De acuerdo al criterio de ECM, la regla queda del siguiente modo:
  Se asigna a Π_1 si:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)\pi_2}{c(2|1)\pi_1}$$

  Sustituyendo las definiciones de f_i, tomando logaritmos:

$$\log\left(\frac{e^{D_1^2/2}}{e^{D_2^2/2}}\right) = \frac{1}{2}(D_2^2 - D_1^2) \geq \log\left(\frac{\pi_2 c(1|2)}{\pi_1 c(2|1)}\right)$$

Caso normal multivariado, poblaciones con varianza común II

Regla óptima en normal multivariada, con $\Sigma_1 = \Sigma_2 = \Sigma$

Se asigna a $\mathbf{x} \in \Pi_1$ si:

$$D_2^2(\mathbf{x}) - D_1^2(\mathbf{x}) \geq 2 \log \left(\frac{\pi_2 c(1|2)}{\pi_1 c(2|1)} \right)$$

donde $D_i^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$ es la distancia de Mahalanobis del punto $\mathbf{x}$ a la media $\boldsymbol{\mu}_i$.

- Noten que cuando los costos y las probabilidades iniciales, la regla asigna la observación a la población que tenga la distancia (medida en términos de la varianza) a la media más cercana: se asigna $\mathbf{x}$ a Π_1 si $D_2(\mathbf{x}) \geq D_1(\mathbf{x})$.
- En la práctica, los parámetros poblacionales se sustituyen por los respectivos muestrales, $\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$, y tomando en consideración la varianza combinada:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

Ejemplos: datos de Salmón I

- Para distinguir a los salmones canadienses de los americanos (Alaska) se mide el diámetro de los anillos. Usualmente, los anillos de los salmones de río (nacidos en Canadá) son menores a los de los americanos. los datos corresponden a las siguientes variables
  - `lugar`: lugar de nacimiento (1=Alaska, 2 = Canadá)
  - `genero`: 1 = macho, 2 = hembra.
  - `rio`: 100* diámetro de anillos del primer año en río (pulgadas)
  - `mar`: 100* diámetro de anillos del primer año en mar (pulgadas)
- Bajo el supuesto de normalidad con igualdad de varianzas, podemos estimar los parámetros poblacionales y calcular la matriz de varianza combinada:

Ejemplos: datos de Salmón II

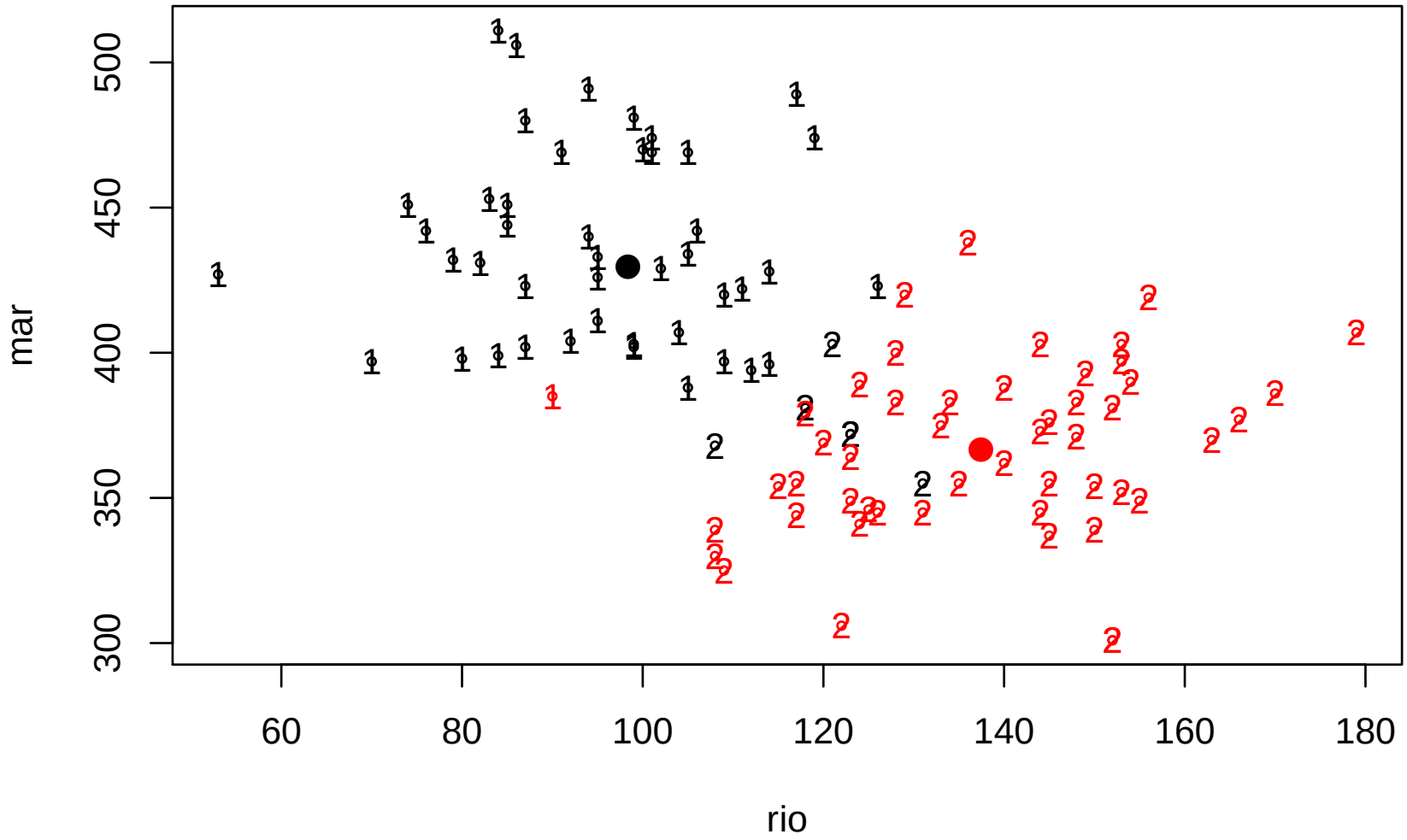
```
salmon <- read.table("../data/Johnson & Wichern/T11-2.DAT")
names(salmon) <- c("lugar","genero","rio","mar")

n1 <- length(salmon[salmon$lugar==1,]$lugar)
mu1 <- colMeans(salmon[salmon$lugar==1,c(3,4)]); S1 <- var(salmon[salmon$lugar==1,c(3,4)])
mu1; S1

   rio    mar
 98.38 429.66
          rio       mar
rio  260.6078 -188.0927
mar -188.0927 1399.0861

n2 <- length(salmon[salmon$lugar==2,]$lugar)
mu2 <- colMeans(salmon[salmon$lugar==2,c(3,4)]); S2 <- var(salmon[salmon$lugar==2,c(3,4)])
mu2; S2

   rio    mar
137.46 366.62
         rio      mar
rio 326.0902 133.5049
mar 133.5049 893.2608

Sp <- ((n1-1)*S1 + (n2-1)*S2)/(n1+n2-2); Sp

          rio        mar
rio 293.34898  -27.29388
mar -27.29388 1146.17347
```

Graficando los puntos, y las medias, obtenemos la población en la que se asignó originalmente cada punto y en la que la regla finalmente asigna:

Ejemplos: datos de Salmón III

```
plot(salmon[,3:4],col=salmon$lugar, cex = 0.5)
points(mu1[1],mu1[2],col=1,pch=16, cex = 1.5)
points(mu2[1],mu2[2],col=2,pch=16, cex = 1.5)

D12 <- function(x) as.numeric(as.numeric((x-mu1)) %*% solve(Sp) %*% as.numeric(x-mu1))
D22 <- function(x) as.numeric(as.numeric((x-mu2)) %*% solve(Sp) %*% as.numeric(x-mu2))

for(i in 1:(n1+n2)) points(salmon[i,3], salmon[i,4],
                                                    pch = ifelse(D12(salmon[i,3:4]) < D22(salmon[i,3:4]),"1"
                                                    col = ifelse(salmon$lugar[i] == 1, 1,2))
```

Ejemplos: datos de Salmón IV



Interpretación geométrica caso normal con varianza fija. I

- Como las distancias D_1^2 y D_2^2 tienen el término común $\mathbf{x}'\Sigma^{-1}\mathbf{x}$, que no depende de la población de la que viene la observación (por tener varianzas iguales), se pueden eliminar y calcular la función indicadora $l_i(\mathbf{x})$ para cada i:

$$l_i(\mathbf{x}) = -\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \log\frac{\pi_i}{c(i|j)}$$

- La función $l_i(\mathbf{x})$ es lineal en $\mathbf{x}$. La regla clasifica a $\mathbf{x}$ en donde esta función lineal sea mínima. Esta regla divide las regiones R_1 y R_2 con frontera definida en donde $l_1 = l_2$:

$$-\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \log\frac{\pi_2 c(1|2)}{\pi_1 c(2|1)} \tag{5}$$

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}\mathbf{x} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}\left(\frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right) - \log\frac{\pi_2 c(1|2)}{\pi_1 c(2|1)} \tag{6}$$

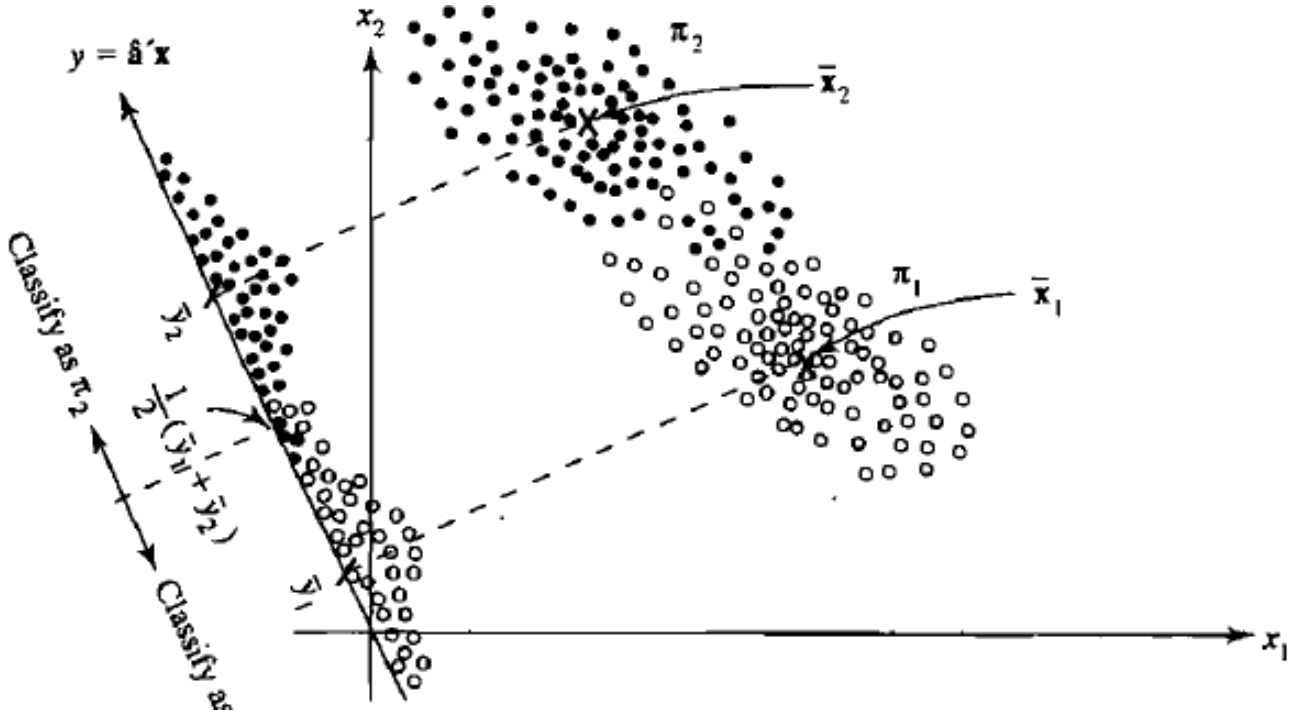
$$\mathbf{w}'\mathbf{x} = \frac{\mathbf{w}'\boldsymbol{\mu}_1 + \mathbf{w}'\boldsymbol{\mu}_2}{2} - \log\frac{\pi_2 c(1|2)}{\pi_1 c(2|1)} \tag{7}$$

$$z = \frac{\bar{z}_1 + \bar{z}_2}{2} = m \tag{8}$$

- De (5) a (6) se reacomodan los términos, de (6) a (7) definimos $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ y de (7) a (8) se definen $z = \mathbf{w}'\mathbf{x}$, $\bar{z}_i = \mathbf{w}'\boldsymbol{\mu}_i - \log\frac{\pi_i}{c(i|j)}$

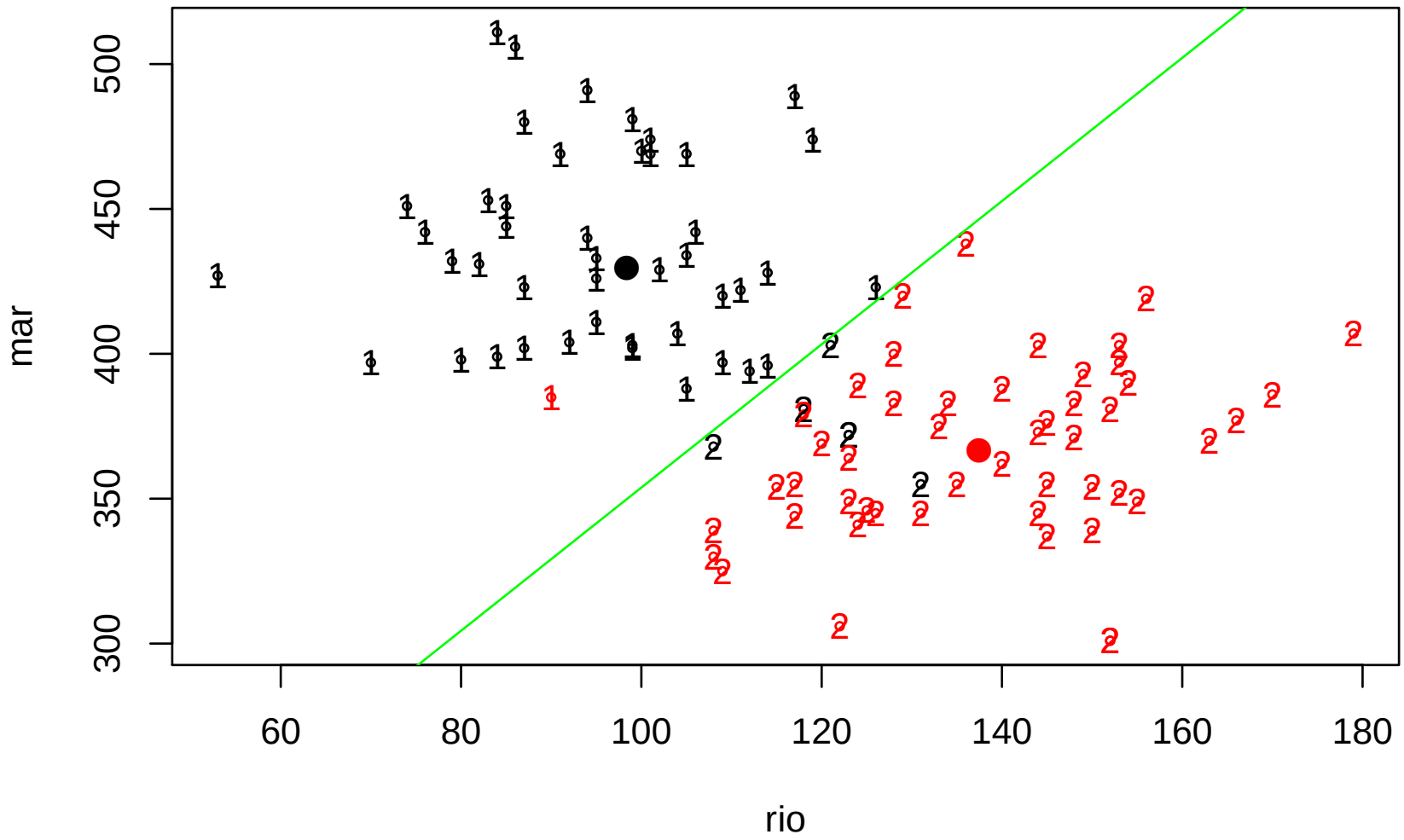
Interpretación geométrica caso normal con varianza fija. II

Normalizando $\mathbf{w}$, $\mathbf{u} = \mathbf{w}/||\mathbf{w}||$, vemos que $\mathbf{u}$ proyecta a $\mathbf{x}$ en la dirección en donde las distancias (de Mahalanobis) a las medias de las dos poblaciones se maximiza y la regla clasifica en Π_1 o Π_2 de acuerdo al lado donde quede la observación respecto al promedio de las medias m proyectadas en esa dirección.



Ejemplo (Salmón) I

- Para los datos de Salmón, calculando los coeficientes de la función lineal frontera:



Ejemplo (Salmón) II

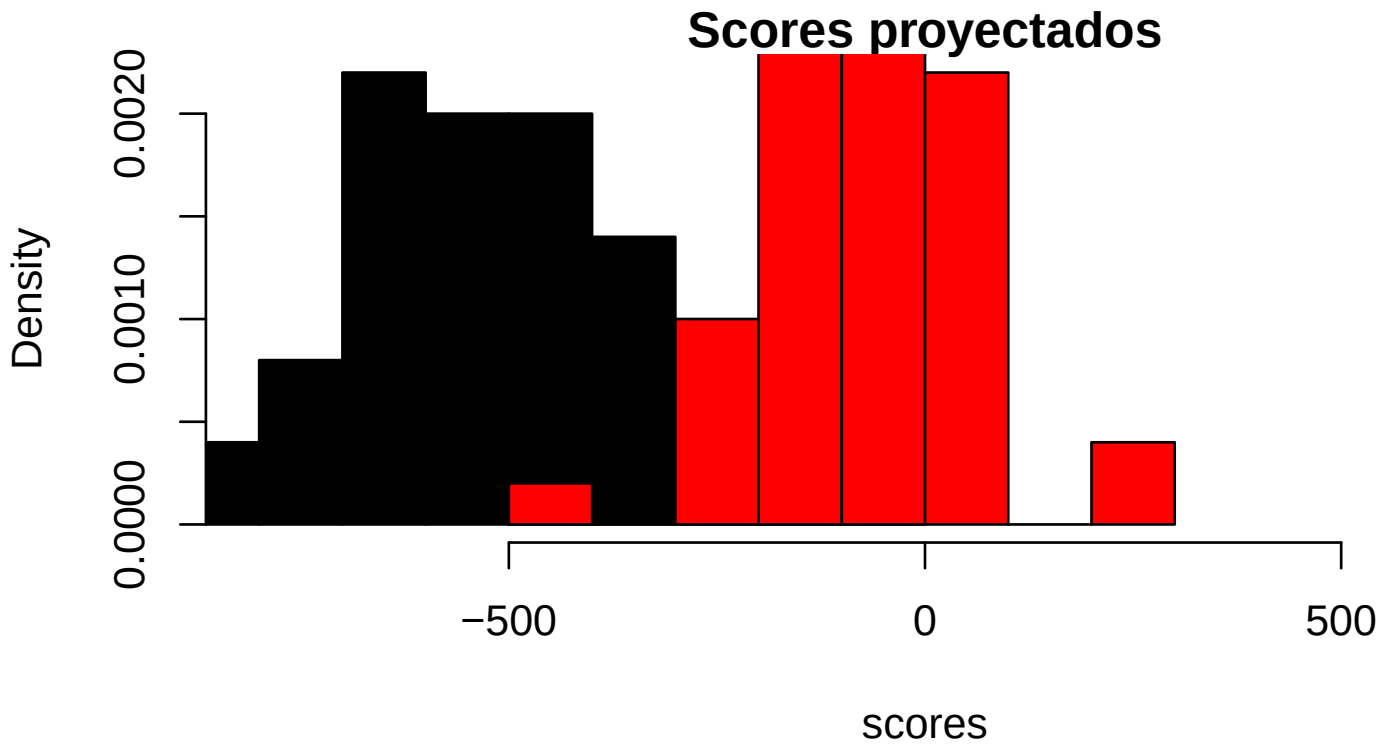
```
- mu1 %*% solve(Sp)

             rio       mar
[1,] -0.3710689 -0.383701

  mu1 %*% solve(Sp) %*% mu1/2

         [,1]
[1,] 100.6834

- mu2 %*% solve(Sp)

             rio        mar
[1,] -0.4994562 -0.3317579

      mu2 %*% solve(Sp) %*% mu2/2

         [,1]
[1,] 95.14216

      w <- solve(Sp)%*%(mu2-mu1)
      w

            [,1]
rio  0.12838726
mar -0.05194311

      u <- w/sum(w^2)
      a <- as.numeric(t(u) %*% (mu1+mu2)/2)
      a

[1] -288.8846

      # y <- a/u[2] - u[1]/u[2]*seq(60,180,1)
      # lines(seq(60,180,1),y,col="green")
```

Ejemplo (Salmón) III

Entonces, $l_1 = -0.371x_1 - 0.384x_2 + 100.683$ y $l_2 = -0.499x_1 - 0.332x_2 + 95.142$ y la línea frontera está dada por $y = 106.6783289 + 2.4716899x$. Esta linea corresponde a la media donde la separación es mayor. la dirección de proyección es perpendicular a esta linea verde.

- Considerando los puntos proyectados en esa dirección, podemos obtener las distribuciones de cada población:

```
scores <- as.matrix(salmon[,3:4]) %*% u
hist(scores[salmon[,1]==1],prob=T,xlim=c(-800,800),col = 1,main="Scores proyectados",xlab="scores")
hist(scores[salmon[,1]==2],prob=T,add=T,col = 2)
```



Caso normal multivariado, diferentes varianzas I

- Ahora consideremos el caso donde se tiene un vector $\mathbf{x}$ con p atributos que tiene una distribución normal multivariada $f_1 \sim \mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ cuando pertenece a Π_1 o $f \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ cuando pertenece a Π_2:

$$f_i(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_i|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)'\right\}$$

- De acuerdo al criterio de ECM, la regla queda del siguiente modo:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}$$

Tomando logaritmos y agrupando los términos con $\mathbf{x}$ y los términos constantes:

$$\log\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = y - m$$

donde:

$$y = -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x}$$

y

$$m = \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$$

Caso normal multivariado, diferentes varianzas II

Entonces, finalmente la regla queda como:

Regla óptima en normal multivariada, con $\Sigma_1 \neq \Sigma_2$

Se asigna a $\mathbf{x} \in \Pi_1$ si:

$$y \geq m + 2 \log \left(\frac{\pi_2 c(1|2)}{\pi_1 c(2|1)} \right)$$

donde y y m se definen como: donde:

$$y = -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x}$$

y

$$m = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$$

- Igual que en el caso de varianzas iguales, los valores poblacionales se sustituyen por los valores muestrales.
- Noten que ahora y es una función cuadrática de $\mathbf{x}$, y entonces se obtiene una regla de clasificación cuadrática.

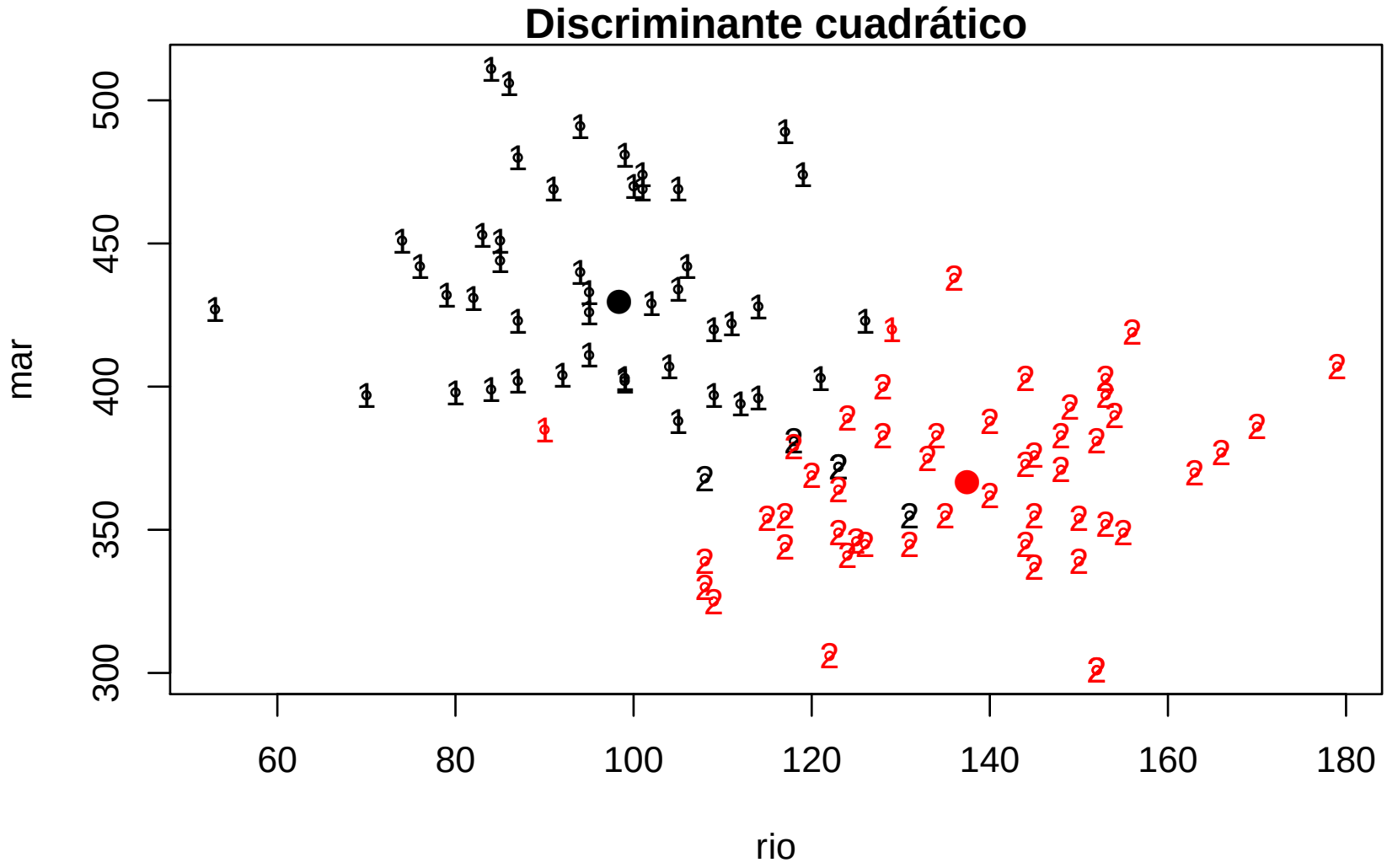
Ejemplo (Salmón) I

- Utilizando la regla de clasificación cuadrática, se obtiene lo siguiente:

```
y <- function(x){
        S1inv <- solve(S1)
        S2inv <- solve(S2)
        as.numeric(-1/2*(as.numeric(x) %*% (S1inv-S2inv) %*% t(x)) +
                         (mu1 %*% S1inv - mu2 %*% S2inv) %*% t(x) )
}
m <- as.numeric( 1/2*log(det(S1)/det(S2)) +
                                  1/2*(as.numeric(mu1 %*% solve(S1) %*% mu1) -
                              as.numeric(mu2 %*% solve(S2) %*% mu2)))
m

[1] 31.47367

scores2 <- NULL
for(i in 1:dim(salmon)[1]) scores2[i] <- y(salmon[i,3:4])
scores2

  [1] 31.45684 28.41723 40.11924 48.69819 34.91354 38.48761 39.17592 39.45510 40.60294 34.99625
 [11] 34.72438 29.98487 29.98487 34.92638 32.60689 34.64363 34.76196 32.27267 36.75946 37.57611
 [21] 32.46446 41.24856 36.45559 36.80788 40.93785 41.07534 33.21250 37.30561 40.06059 31.04966
 [31] 33.09064 31.88088 41.88440 42.45722 43.24771 36.22155 42.23316 37.62194 36.39959 37.70136
 [41] 49.85599 42.91559 43.67171 41.47910 36.79374 38.30632 35.97825 42.68253 44.95221 45.04630
 [51] 31.55705 26.62076 21.38283 26.19195 23.85223 30.66861 26.35560 28.13969 27.52626 26.74623
 [61] 26.94848 29.49139 27.96379 23.15949 26.04020 26.35316 26.04020 31.25920 27.93581 26.86757
 [71] 34.53457 26.55765 29.63729 27.16322 30.02721 28.98004 29.88042 28.63140 30.94255 30.20218
 [81] 26.50011 26.11388 26.02937 25.44189 29.21838 29.41416 30.63511 28.11519 24.33223 26.76138
 [91] 28.82546 29.82641 29.05474 27.23873 28.26054 26.02057 28.69174 28.75910 25.71112 29.90457
```

Ejemplo (Salmón) II

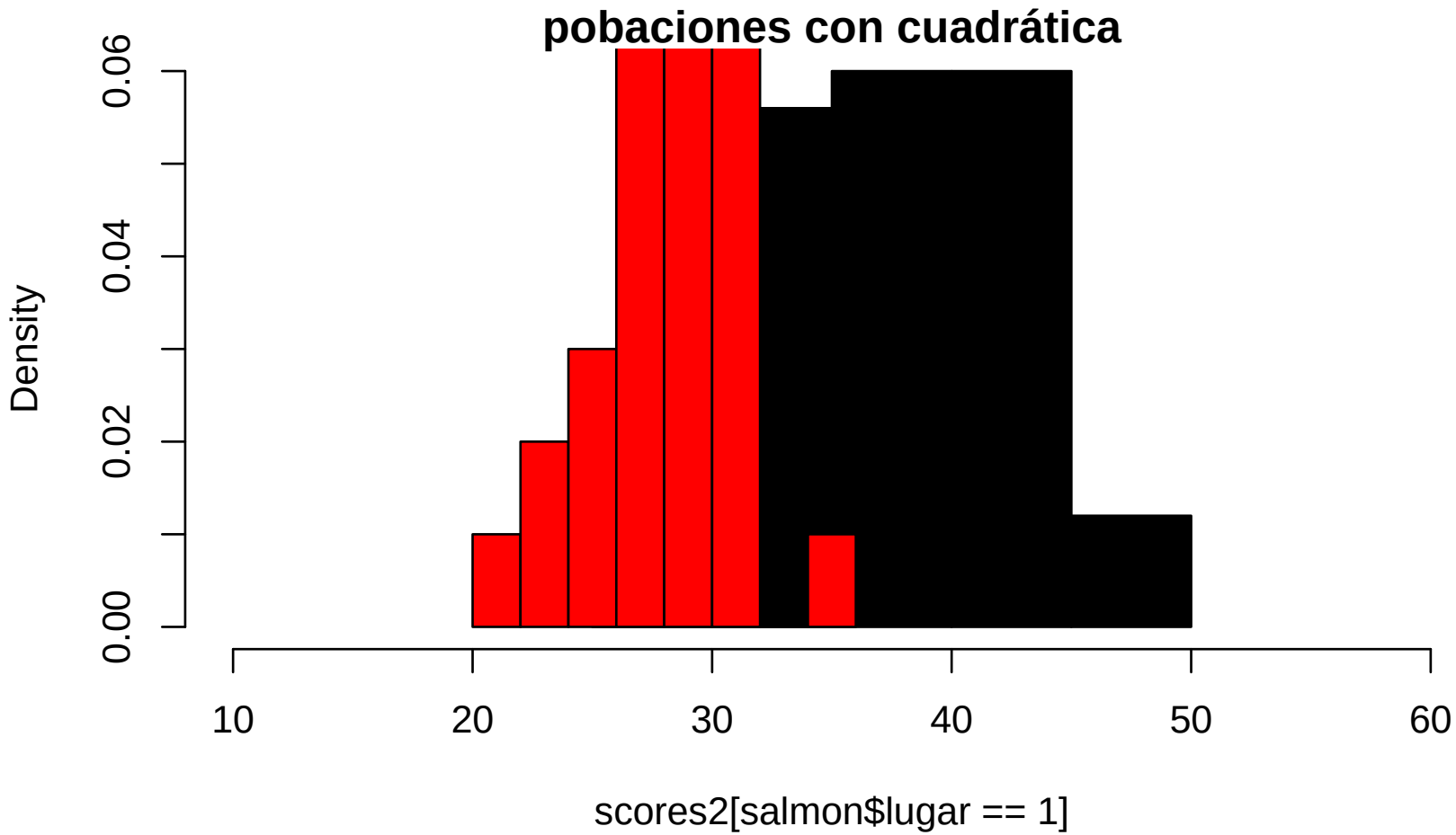
- Graficando los resultados se obtiene:



Ejemplo (Salmón) III

- En este caso, los histogramas quedan de la siguiente manera:

```
hist(scores2[salmon$lugar==1], prob = T, col = 1, xlim = c(10,60),
     main = "pobaciones con cuadrática")
hist(scores2[salmon$lugar==2], prob = T, col = 2, add = T)
```



Notas sobre reglas cuadráticas para distribuciones no normales. I

- Las regiones resultantes con funciones cuadráticas son típicamente disjuntas y difíciles de interpretar en varias dimensiones.

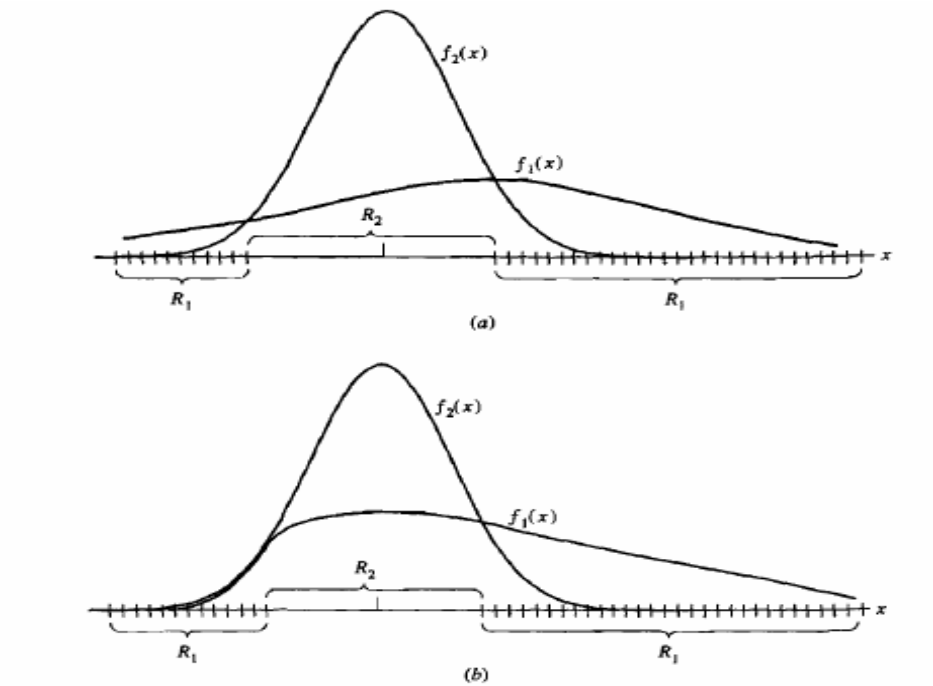


Figure 11.6 Quadratic rules for (a) two normal distribution with unequal variances and (b) two distributions, one of which is nonnormal—rule not appropriate.

Notas sobre reglas cuadráticas para distribuciones no normales. II

- El número de parámetros a estimar en el caso lineal es de $gp + \frac{p(p+1)}{2}$ y en el caso cuadrático es de $g(p + \frac{p(p+1)}{2})$. Un número de parámetros puede hacer muy inestable la discriminación cuadrática, salvo que la muestra sea muy grande.
- La regla de discriminación cuadrática es muy sensible a desviaciones de la distribución normal. Usualmente la discriminación lineal es mucho más robusta en esos casos.

Evaluación de las reglas de clasificación

Evaluando la regla de clasificación I

- Para determinar si la regla de clasificación es efectiva, se pueden examinar las probabilidades de clasificación errónea, a través de la probabilidad de clasificación errónea mínima (TPM) introducida antes.
- La Tasa óptima de error (OER) es el valor mínimo posible de TPM:

$$OER = \min_{R_1, R_2} TPM(R_1, R_2) \quad \text{sujeto a: } \Omega = R_1 \cup R_2$$

que se obtiene como se vió antes, cuando $\mathbf{x}$ se asigna a R_1 si $\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{\pi_2}{\pi_1}$

- El OER requiere que se conozcan los parámetros de las densidades f_1 y f_2 que usualmente son desconocidos, por lo que se considera la tasa de error real (AER):

$$AER(\hat{R}_1, \hat{R}_2) = TPM(\hat{R}_1, \hat{R}_2)$$

- Sin embargo, la AER depende de las funciones de densidad de las poblaciones que usualmente son desconocidas.

Evaluando la regla de clasificación II

- Una medidad de eficiencia que no depende de la forma de las poblaciones es la Tasa de error aparente (APER), que se puede calcular para cualquier regla de clasificación:

$$APER = \frac{n_{M1} + n_{M2}}{n_1 + n_2}$$

es la proporción del total de datos mal clasificados en relación al tamaño de la muestra; los valores de n_{M1} y n_{M2} se obtienen de la matriz de confusión:

		Clasificación Π_1	Clasificación Π_2	
Pertenencia Real	Π_1	n_{C1}	n_{M1}	n_1
	Π_2	n_{M2}	n_{C2}	n_2

Ejemplo. [Ejemplo para datos de Salmón]

Para el caso lineal de los datos de salmón, se obtiene:

Evaluando la regla de clasificación III

```
## APER
salmon$lda <- NULL
for(i in 1:(n1+n2)) salmon$lda[i] <- ifelse(D12(salmon[i,3:4]) < D22(salmon[i,3:4]),1,2)
CM <- table(salmon$lugar,salmon$lda);CM #matriz de confusión
```

```
   1  2
1 44  6
2  1 49
```

```
APER <- (CM[1,2]+CM[2,1])/sum(CM); APER #tasa de error aparente
```

```
[1] 0.07
```

Entonces:

		Clasificación Π_1	Clasificación Π_2	
Pertenencia Real	Π_1	44	6	50
	Π_2	1	49	50

y el $APER = 0.07$.
Para el caso cuadrático:

Evaluando la regla de clasificación IV

```
## APER cuadrático
salmon$qda <- NULL
for(i in 1:(n1+n2)) salmon$qda[i] <- ifelse(scores2[i] >= m,1,2)
CM <- table(salmon$lugar,salmon$qda);CM #matriz de confusión
```

```
   1  2
1 45  5
2  2 48
```

```
APER <- (CM[1,2]+CM[2,1])/sum(CM); APER #tasa de error aparente
```

```
[1] 0.07
```

Prácticamente en este ejemplo no hay diferencia en el ajuste.

□

Validación cruzada tipo jackknife I

- Otro estimador del AER fue propuesto por Lachenbruch (1975) está basado en validación cruzada y jackknife. Para la población Π_i, tomando $j = 1, \ldots, n_i$:
  1. Crear la regla de clasificación para Π_i sin la observación j
  2. Usar la regla de clasificación anterior para clasificar la observación i.
- Del procedimiento anterior se obtienen n^*_{M1}, el total de veces que las observaciones retenidas fueron clasificadas incorrectamente en la población 1 y n^*_{M2} el total de veces que las observaciones retenidas fueron clasificadas en la población 2. Entonces el estimador del AER está dado por:

$$\hat{\theta} = \mathrm{E}(\hat{AER}) = \frac{n^*_{M1} + n^*_{M2}}{n_1 + n_2}$$

- Más adelante haremos este ejercicio con las funciones disponibles, porque de otra manera hay que programar el procedimiento.

Validación Cruzada

Validación cruzada: ideas básicas I

- La validación cruzada (cross-validation) es una forma de medir el *desempeño predictivo* de un modelo estadístico.
  - Las estadísticas de ajuste de un modelo no son guía adecuada del poder predictivo del modelo. Por ejemplo, en regresión una R^2 alta no necesariamente indican que el modelo es bueno para predecir (se pueden incluir más términos para mejorar R^2 pero su poder predictivo empeora con el número de términos)
- Con la validación cruzada, podemos evaluar: (1) la estabilidad de los parámetros estimados, (2) la exactitud de un problema de clasificación, (3) la adecuación de un modelo ajustado, etc.
- El jackknife es un caso particular de la validación cruzada.
- El enfoque de la validación cruzada es dividir los datos disponibles en dos conjuntos: un conjunto de *entrenamiento*, que se usa para estimar el modelo y un conjunto de *prueba*, en el que se evalúa el modelo y se obtiene un estimador del error de ajuste del modelo.
- Hay diversas maneras de hacer este procedimiento:
  - **uno-afuera**. se usan $n-1$ datos para estimar el modelo. El modelo se prueba en el dato que se dejó afuera. Esto se puede realizar n veces se utilizan los errores $e_i^* = y_i - \hat{y}_i$ para calcular el error cuadrático medio de validación cruzada: $MSE_{cv} = \frac{\sum_{i=1}^{n} e_i^2}{n}$
  - **k-afuera**.
  - **Muestreo aleatorio**.

Ejemplo de validación cruzada I

- En el siguiente ejemplo, los datos x y y están relacionados, tienen correlación, pero la relación posiblemente no es lineal.

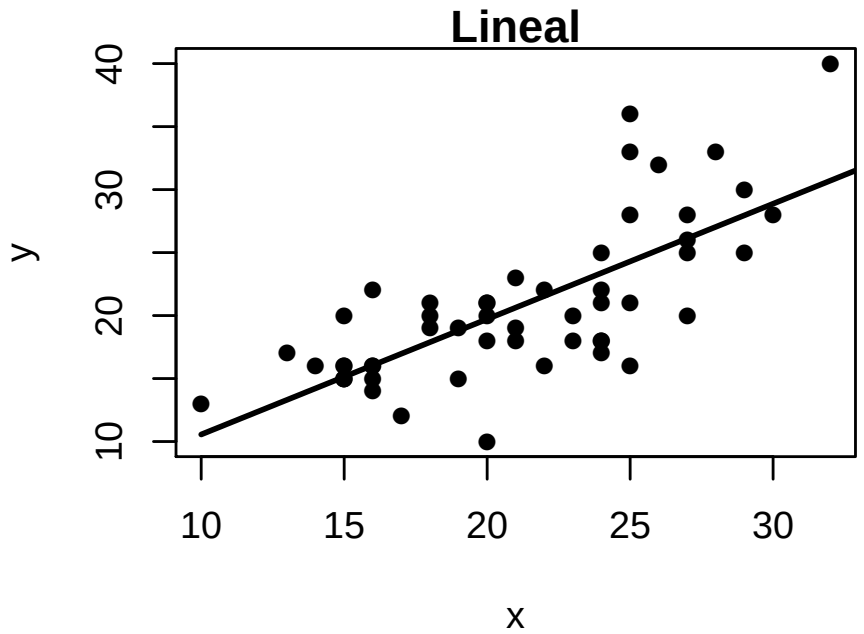
```
x

 [1] 24 16 24 18 18 10 14 16 18 20 21 20 21 15 16 15 17 19 16 15 15 13 24 22 21 24 15 20 20 25 27 22
[33] 20 24 24 23 29 27 23 19 25 15 16 27 27 30 29 26 25 25 32 28 25

y

 [1] 25 22 17 21 20 13 16 14 19 10 23 20 19 15 16 16 12 15 15 15 15 17 18 16 18 22 20 21 21 21 25 22
[33] 18 21 18 20 25 20 18 19 16 16 16 26 28 28 30 32 28 36 40 33 33
```

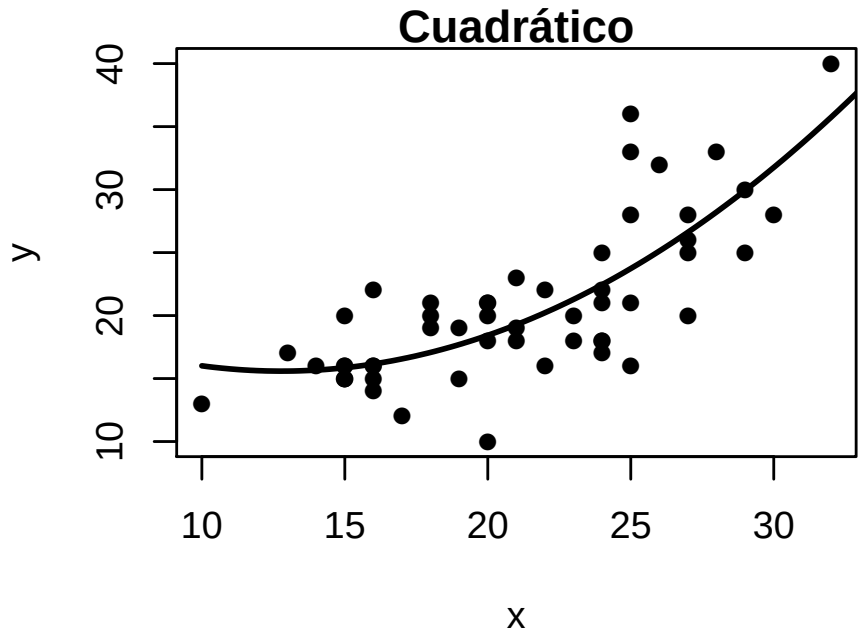
- En este ejemplo, nos concentraremos en el error de predicción, que puede ser estimado por validación cruzada, sin hacer supuestos fuertes acerca del error de la variable.
- Los modelos que se propondrán para la relación son los siguientes:
  1. Lineal: $y = \beta_0 + \beta_1 x + \epsilon$
  2. Cuadrático: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
  3. Exponencial: $log(y) = log(\beta_0) + \beta_1 x + \epsilon$
  4. Log-Log: $log(y) = \beta_0 + \beta_1 log(x) + \epsilon$

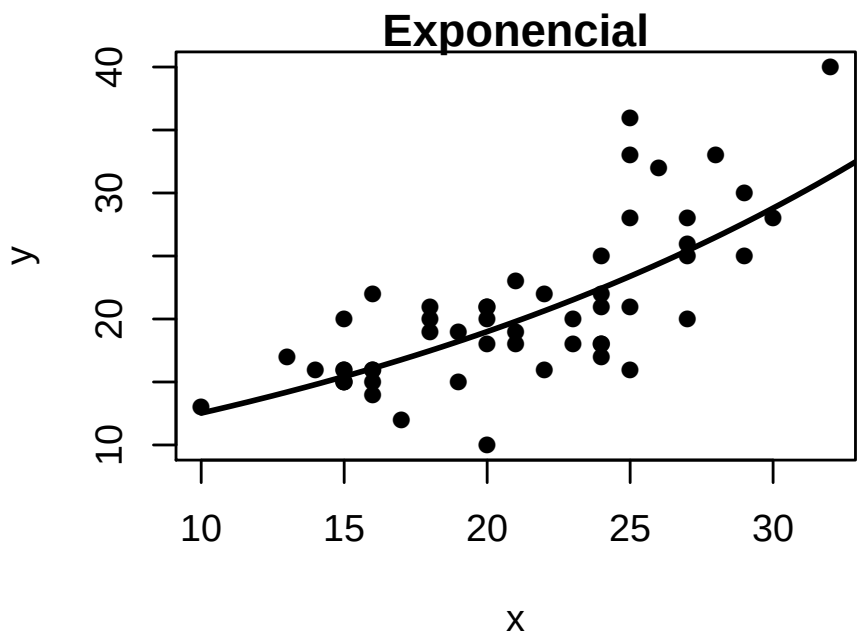
Modelos I

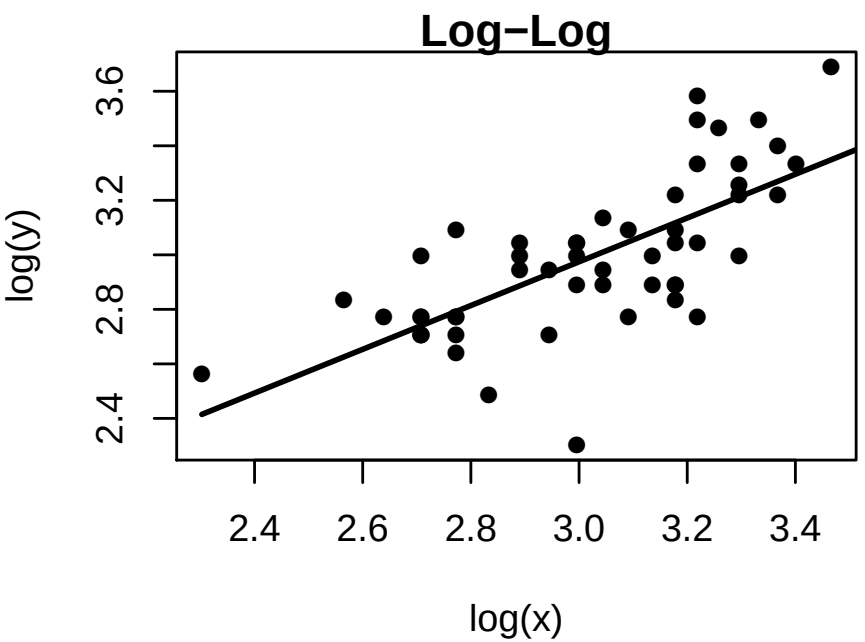
```
par(mfrow=c(2,2))
par(oma=c(1,1,1,1),mar=c(4,4,1,1))
a <- seq(10,40,0.1) #sucesión para graficar los ajustes

L1 <- lm(y ~ x)
plot(x,y,main="Lineal",pch=16)
yhat1 <- L1$coef[1] + L1$coef[2]*a
lines(a,yhat1,lwd=2)

L2 <- lm(y ~ x + I(x^2))
plot(x,y,main="Cuadrático",pch=16)
yhat2 <- L2$coef[1] + L2$coef[2]*a +L2$coef[3]*a^2
lines(a,yhat2,lwd=2)

L3 <- lm(log(y) ~ x)
plot(x,y,main="Exponencial",pch=16)
logyhat3 <- L3$coef[1] + L3$coef[2]*a
yhat3 <- exp(logyhat3)
lines(a,yhat3,lwd=2)

L4 <- lm(log(y) ~ log(x))
plot(log(x),log(y),main="Log-Log",pch=16)
logyhat4 <- L4$coef[1] + L4$coef[2]*log(a)
lines(log(a),logyhat4,lwd=2)
```

Modelos II


Lineal
y
x
10
15
20
25
30
10
20
30
40


Cuadrático
y
x
10
15
20
25
30
10
20
30
40


Exponencial
y
x
10
15
20
25
30
10
20
30
40


Log−Log
log(y)
log(x)
2.4
2.6
2.8
3.0
3.2
3.4
2.4
2.8
3.2
3.6

Ejemplo Validación Cruzada I

- Una vez que el modelo es ajustado, se evalúa el ajuste.

Procedimiento para estimar el error de predicción usando validación cruzada (uno afuera)

1. Para $k = 1, \ldots, n$ dejar la observación (x_k, y_k) para ser el punto de prueba y usar las observaciones restantes para ajustar el modelo.
   a. Ajusta el modelo usando sólo $n - 1$ observaciones en el conjunto de entrenamiento.
   b. Calcular la respuesta predictiva $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ para el punto de prueba
   c. Calcula el error de predicción $e_k = y_k - \hat{y}_k$.
2. Estima la media de los errores de predicción al cuadrado $\hat{\sigma}^2_\epsilon = \frac{1}{n} \sum_{k=1}^{n} e_k^2$.

Ejemplo Validación Cruzada I

```
n <- length(x)
e1 <- e2 <- e3 <- e4 <- numeric(n)

for(k in 1:n){
yy <- y[-k]
xx <- x[-k]
J1 <- lm(yy ~ xx)
e1[k] <- y[k] - (J1$coef[1] + J1$coef[2]*x[k])
J2 <- lm(yy ~ xx + I(xx^2))
e2[k] <- y[k] - (J2$coef[1] + J2$coef[2]*x[k] + J2$coef[3]*x[k]^2)
J3 <- lm(log(yy) ~ xx)
yhat3 <- exp(J3$coef[1] + J3$coef[2]*x[k])
e3[k] <- y[k] - yhat3
J4 <- lm(log(yy) ~ log(xx))
yhat4 <- exp(J4$coef[1] + J4$coef[2]*log(x[k]))
e4[k] <- y[k] - yhat4
}
```

Los siguientes son los estimados de los errores de predicción

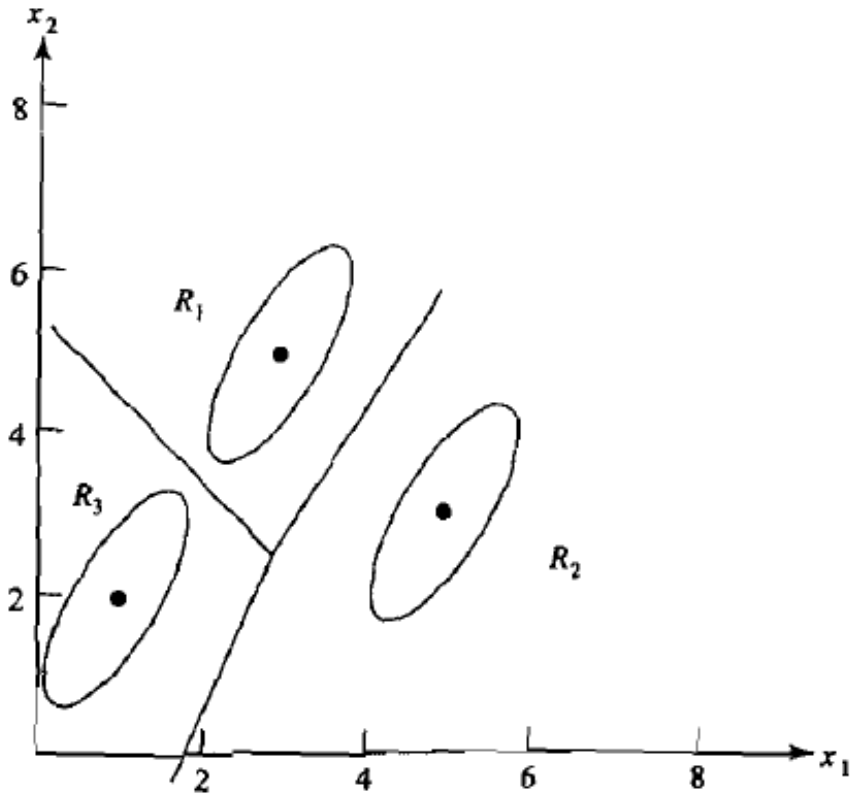
```
c(mean(e1^2),mean(e2^2), mean(e3^2), mean(e4^2))

[1] 19.55644 17.85248 18.44188 20.45424
```

Entonces el mejor modelo es el modelo cuadrático que tiene el menor error cuadrático medio de predicción.

Generalización a $g > 2$ poblaciones

Generalización a $g > 2$ poblaciones

- La generalización a más de 2 poblaciones es directa. Las definiciones que hemos hecho antes se extienden de manera directa a más de dos poblaciones.
- Ahora tenemos probabilidades condicionales $P(l|k)$ como las probabilidades condicionales de clasificar en la población Π_l dado que la observación viene de la población k.
- Los costos serán $c(l|k)$ y las probabilidades iniciales serán π_k para $k = 1, 2, \ldots, g$.



Criterios de evaluación de regla de clasificación cuando $g > 2$ I

- El costo esperado condicional de clasificación errónea de un elemento de la población Π_k es

$$ECM(k) = \sum_{l \neq k} P(l|k)c(l|k)$$

y el costo total esperado es:

$$ECM = \sum_{i=1}^{g} \pi_i ECM(i) = \sum_{i=1}^{g} \sum_{\substack{k=1 \\ k \neq i}}^{g} \pi_i P(k|i)c(k|i)$$

- Las regiones $\{R_1, R_2, \ldots, R_g\}$ que minimizan el ECM están definidas de tal manera que: asigna $\mathbf{x}$ a Π_k si se maximiza $f_k(\mathbf{x})\pi_k$, o equivalentemente, si se minimiza

$$\sum_{l \neq k} \pi_k f_l(\mathbf{x})c(k|l)$$

- Cuando los costos de clasificación errónea $c(l|k)$ son todos iguales, entonces la regla anterior también es equivalente a asignar $\mathbf{x}$ a Π_k si $f_k(\mathbf{x})\pi_k > f_l(\mathbf{x})\pi_l$ para todo $l \neq k$.
- El último caso es equivalente también a maximizar la probabilidad posterior $P(\Pi_k|\mathbf{x})$.

Casos para distribuciones normales I

- Consideraremos los enunciados para los casos normales para múltiples poblaciones, que son extensiones del caso de 2 poblaciones.

Regla de clasificación en poblaciones normales con diferentes varianzas, basada en TPM

Clasifica $\mathbf{x} \in \Pi_k$ si

$$qda_k(\mathbf{x}) = \max_{i \in \{1,\ldots,g\}} \{qda_i(\mathbf{x})\}$$

donde

$$qda_i(\mathbf{x}) = -\frac{1}{2}\left(\log|\mathbf{\Sigma}_i| + D_i^2(\mathbf{x})\right) + \log \pi_i$$

y $D_i(\mathbf{x})$ es la distancia de Mahalanobis con covarianza $\mathbf{\Sigma}_i$ a la media de la población i.

Casos para distribuciones normales II

- Para varianzas iguales:

Regla de clasificación en poblaciones normales con misma varianza, basada en TPM

Clasifica $\mathbf{x} \in \Pi_k$ si

$$lda_k(\mathbf{x}) = \max_{i \in \{1,\ldots,g\}} \{lda_i(\mathbf{x})\}$$

donde

$$lda_i(\mathbf{x}) = \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \log \pi_i$$

y $D_i(\mathbf{x})$ es la distancia de Mahalanobis con covarianza $\boldsymbol{\Sigma}$ a la media de la población i.

Enfoque de Fisher para LDA cuando $g > 2$ I

- Para derivar el modelo de más de dos grupos, consideraremos el desarrollo seguido por Ronald Fisher para resolver el problema. Esta versión da los mismos resultados que seguimos previamente para $g = 2$.
- El modelo de Fisher no requiere el supuesto de normalidad, pero supone, inicialmente, que las varianzas son iguales en todas las poblaciones.
- El enfoque de Fisher consiste en encontrar el menor número de combinaciones lineales $y_1 = \mathbf{a}_1'\mathbf{x}, \ldots, y_m = \mathbf{a}_m'\mathbf{x}$, $m < g$, que mejor separe las poblaciones. Supongamos que $f_k(\mathbf{x})$ tiene media $\boldsymbol{\mu}_k$ y varianza $\boldsymbol{\Sigma}$ para la población Π_k (se considera la varianza constante entre poblaciones).
- Sea

$$\mathbf{B}\boldsymbol{\mu} = \sum_{i=1}^{g} (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})'$$

donde $\bar{\boldsymbol{\mu}} = \frac{1}{g}\sum_i^g \boldsymbol{\mu}_i$, la suma de productos cruzados *entre* las medias de los grupos.

Enfoque de Fisher para LDA cuando $g > 2$ II

- Una combinación lineal $y = \mathbf{a}'\mathbf{x}$ cambiará su media según la población de la que provenga $\mathbf{x}$, si $\mathbf{x} \in \Pi_k$:

$$\begin{aligned} \mathrm{E}(y_k|\Pi_k) &= \mathbf{a}'\mathrm{E}(\mathbf{x}|\Pi_k) = \mathbf{a}'\boldsymbol{\mu}_k \\ \mathrm{Var}(y_k|\Pi_k) &= \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \end{aligned}$$

Además, para la media global de las combinaciones lineales

$$\bar{\boldsymbol{\mu}}_y = \frac{1}{g}\sum_{i=1}^{g}\mathrm{E}(y_i|\Pi_i) = \mathbf{a}'\bar{\boldsymbol{\mu}}$$

Variabilidad entre vs. intra grupo I

- la razón:

$$Q = \frac{\sum_{i=1}^{g} (\mu_{y_k} - \bar{\mu}_y)^2}{\sigma_y^2} = \frac{\mathbf{a}'\mathbf{B}\boldsymbol{\mu}\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}$$

  mide la variabilidad *entre* los grupos de valores de y relativos a la variabilidad total, dentro de los grupos.
- Fisher selecciona el vector a que maximiza este cociente.
- Usualmente, no conocemos los valores poblacionales, pero se cuenta con un conjunto de entrenamiento de observaciones clasificadas correctamente.
- Supongan que el conjunto de entrenamiento es una muestra de n_i observaciones de la población Π_i. Entonces podemos estimar el cociente Q con las siguientes cantidades:

Variabilidad entre vs. intra grupo II

$$
\begin{aligned}
\bar{\mathbf{x}}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij} \quad i = 1, 2, \ldots, g \\
\bar{\mathbf{x}} &= \frac{1}{g}\sum_{i=1}^{g}\bar{\mathbf{x}}_i \\
\mathbf{B} &= \sum_{i=1}^{g}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \\
\mathbf{W} &= \sum_{i=1}^{g}(n_i - 1)\mathbf{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})'
\end{aligned}
$$

- Entonces $\hat{Q} = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$

Discriminantes muestrales I

Discriminantes muestrales

El $k-$**ésimo discriminante muestral** es la combinación lineal

$$\hat{y}_k = \mathbf{a}_k' \mathbf{x}$$

para $k = 1, 2, \ldots, s = \text{mín}\{g-1, p\}$, donde $\mathbf{a}_k$ es proporcional al $k-$ésimo vector propio de $\mathbf{W}^{-1}\mathbf{B}$.
Los vectores $\mathbf{a}_k$ se escalan para hacer que $\hat{y}_k$ tenga varianza unitaria: $\mathbf{a}_k' \hat{\mathbf{\Sigma}} \mathbf{a}_k = 1$ donde

$$\hat{\mathbf{\Sigma}} = \mathbf{S}_p = \frac{1}{n-g}\mathbf{W}$$

con $n = \sum_{k=1}^{g} n_k$.

Clasificación de nuevos elementos con discriminantes muestrales para distribuciones normales multivariadas, caso general I

- Cuando $\mathbf{x}$ proviene de distribuciones normales, $f_k(\mathbf{x}) \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- En este caso, se suponen costos iguales de clasificación errónea. Entonces se asigna $\mathbf{x}$ a la población Π_k que minimiza $\sum_{l \neq k} \pi_l f_l(\mathbf{x})$, que equivale a maximizar $\pi_k f_k(\mathbf{x})$.
- Similar al caso $g = 2$, tenemos las siguientes reglas:

Regla de clasificación para $g > 2$ poblaciones normales

Se asigna $\mathbf{x}$ a Π_k que maximiza:

- score discriminante cuadrático

$$d_k^Q(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_k| - \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_k)'\mathbf{S}_k^{-1}(\mathbf{x} - \bar{\mathbf{x}}_k) + \log \pi_k$$

  cuando las varianzas son diferentes, y

- score discriminante lineal

$$d_k^L(\mathbf{x}) = \bar{\mathbf{x}}_k)'\mathbf{S}_p^{-1}\mathbf{x} - \frac{1}{2}\bar{\mathbf{x}}_k'\mathbf{S}_p^{-1}\bar{\mathbf{x}}_k + \log \pi_k$$

  cuando las varianzas son iguales.

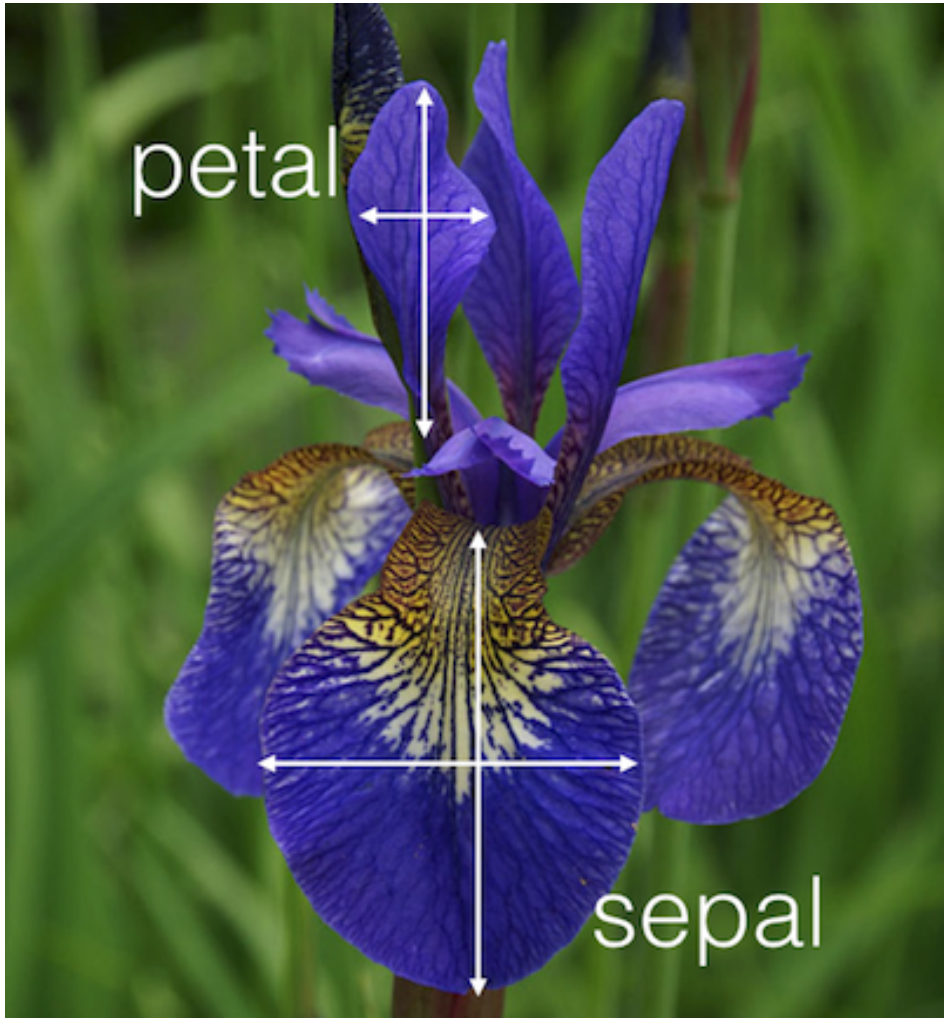
Herramientas en R para clasificación y Discriminación I

- hay métodos de componentes principales en varios paquetes de R:
  - La función `lda` y `qda` del paquete `MASS` han sido las principales herramientas durante muchos años.
  - `klaR`: Funciones misceláneas para clasificación y visualización, algunos métodos no parampetricos y Bayesianos.

Ejemplo: datos Iris

Datos Iris

- Los datos consisten en mediciones de $p = 4$ variables (longitud del sépalo, ancho del sépalo, longitud del pétalo, ancho del pétalo) tomadas de $n_k = 50$ flores muestreadas alaetoriamente de $g = 3$ especies (setosa, versicolor, virginica).
- El objetivo es crear una función discriminante que mejor clasifique una nueva flor en una de las tres especies.



Datos Iris I

```
head(iris); p <- 4; g <- 3
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Calculamos las medias muestrales de cada población: $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{\mathbf{x}}_3$

```
especies <- unique(iris$Species)
xbari <- list(NULL)
for(i in especies)xbari[[match(i,especies)]] <- colMeans(iris[iris$Species==i,1:p])
xbari
```

```
[[1]]
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.006        3.428        1.462        0.246

[[2]]
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.936        2.770        4.260        1.326

[[3]]
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       6.588        2.974        5.552        2.026
```

Calculamos la matriz combinada $\mathbf{S}_p$

Datos Iris II

```
Sp <- matrix(0,p,p)
nx <- rep(50,g)  #tamaños de muestra por población
for(j in 1:g){
        x <- iris[iris$Species==especies[j],1:p]
        Sp <- Sp + cov(x)*(nx[j]-1)
}
Sp <- Sp/(sum(nx)-g); round(Sp, 4)
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length       0.2650      0.0927       0.1675      0.0384
Sepal.Width        0.0927      0.1154       0.0552      0.0327
Petal.Length       0.1675      0.0552       0.1852      0.0427
Petal.Width        0.0384      0.0327       0.0427      0.0419
```

Ajustamos un modelo discriminante lineal con la función `lda` del paquete `MASS`:

Datos Iris III

```
library(MASS)


Attaching package: 'MASS'
The following object is masked from 'package:DAAG':
    hills

lda1 <- lda(Species ~ . , data = iris, prior=rep(1/3,3))
lda1

Call:
lda(Species ~ ., data = iris, prior = rep(1/3, 3))

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

Datos Iris IV

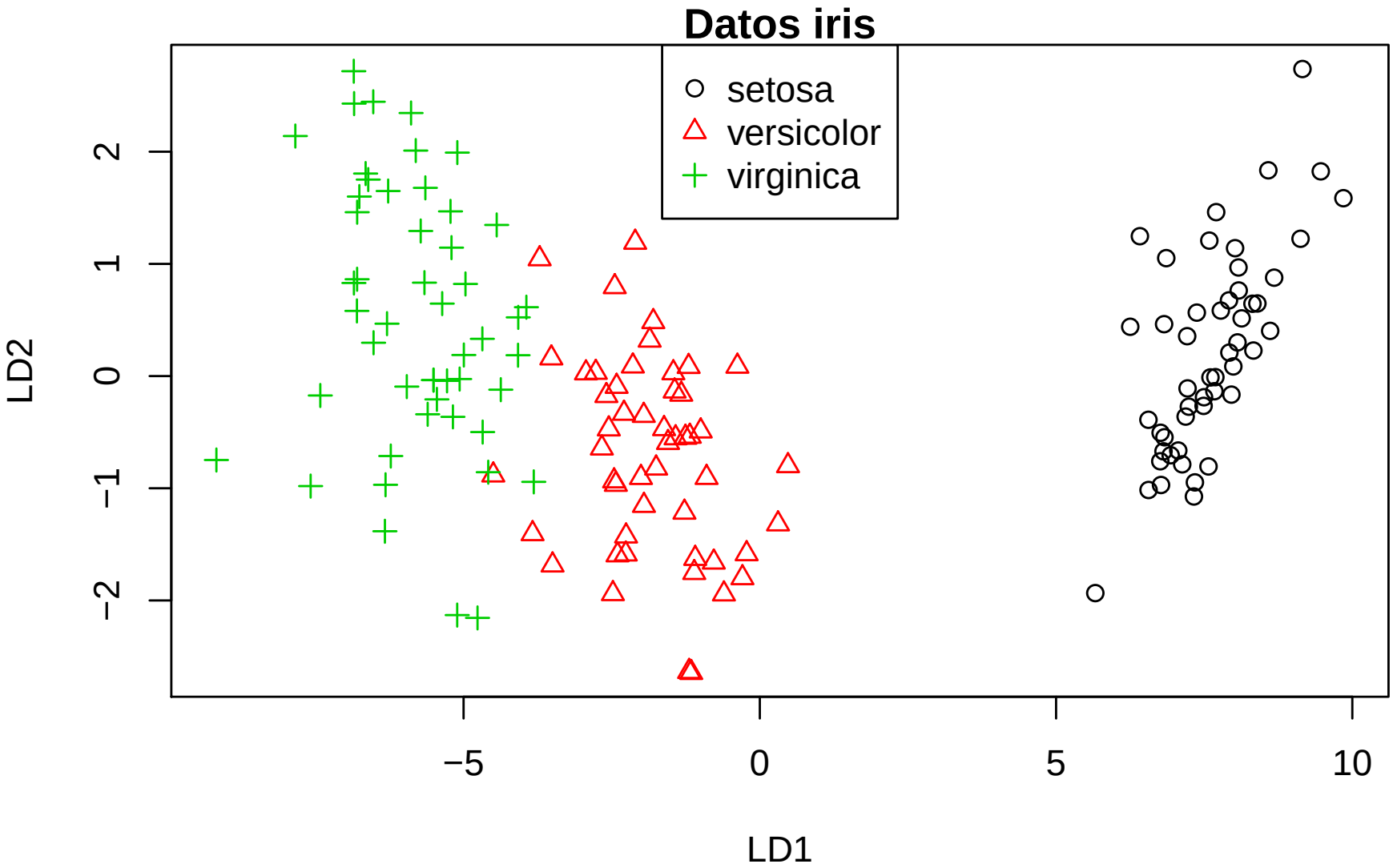
- Noten que los coeficientes de los discriminantes lineales corresponden a la matriz que definimos $\mathbf{A}$.
- También vemos que el primer discriminante es el que más contribuye a la separación de las medias y el segundo discriminante prácticamente no contribuye a la separación.
- Calculamos los scores $\hat{\mathbf{y}} = \mathbf{A}'(\mathbf{x} - \bar{\mathbf{x}}_i)$ para todas las observaciones:

```
mu_k <- lda1$means
mu <- colMeans(mu_k)    #media global
scores <- scale(iris[,1:p], center = mu, scale=F) %*% lda1$scaling
scores2 <- predict(lda1)$x      #otra manera de calcular los scores
#comparamos los scores:
sum(scores-scores2)^2

[1] 9.653312e-27
```

A continuación se grafican los scores en las direcciones de los dos discriminantes.

```
plot(scores, xlab="LD1", ylab="LD2", col=as.numeric(iris$Species),
                 pch = as.numeric(iris$Species), main= "Datos iris")
legend("top", legend = especies, pch = 1:3, col = 1:3)
```

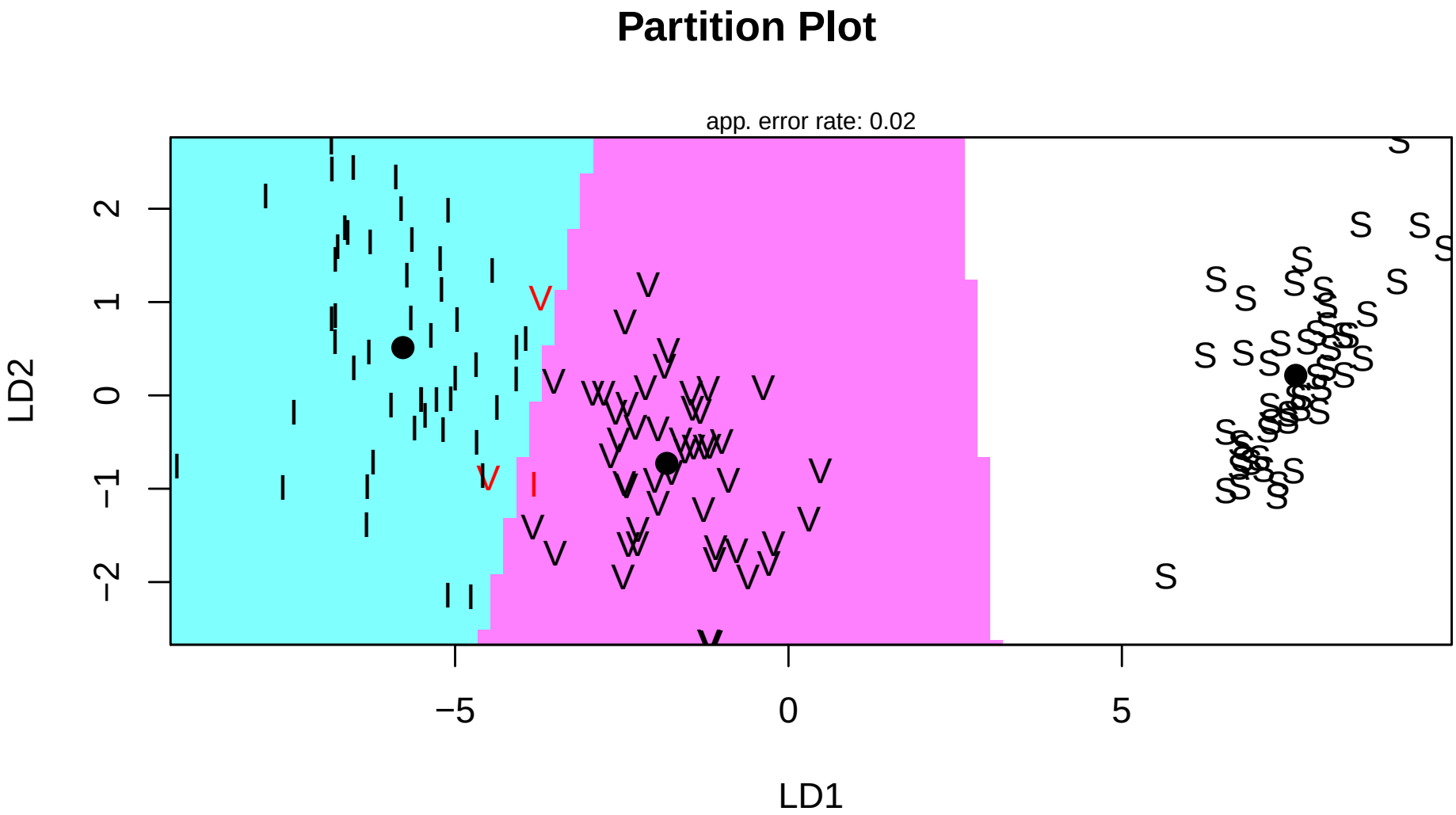
Datos Iris V



En términos de gráficas, podemos encontrar explícitamente las regiones R_i una vez que tenemos las fronteras de las regiones, utilizando el paquete `klaR`.

Datos Iris VI

```
library(klaR)
especies <- factor(rep(c("S","V","I"),rep(50,3)))
partimat(x = scores[,2:1], grouping = especies, method = "lda")
```

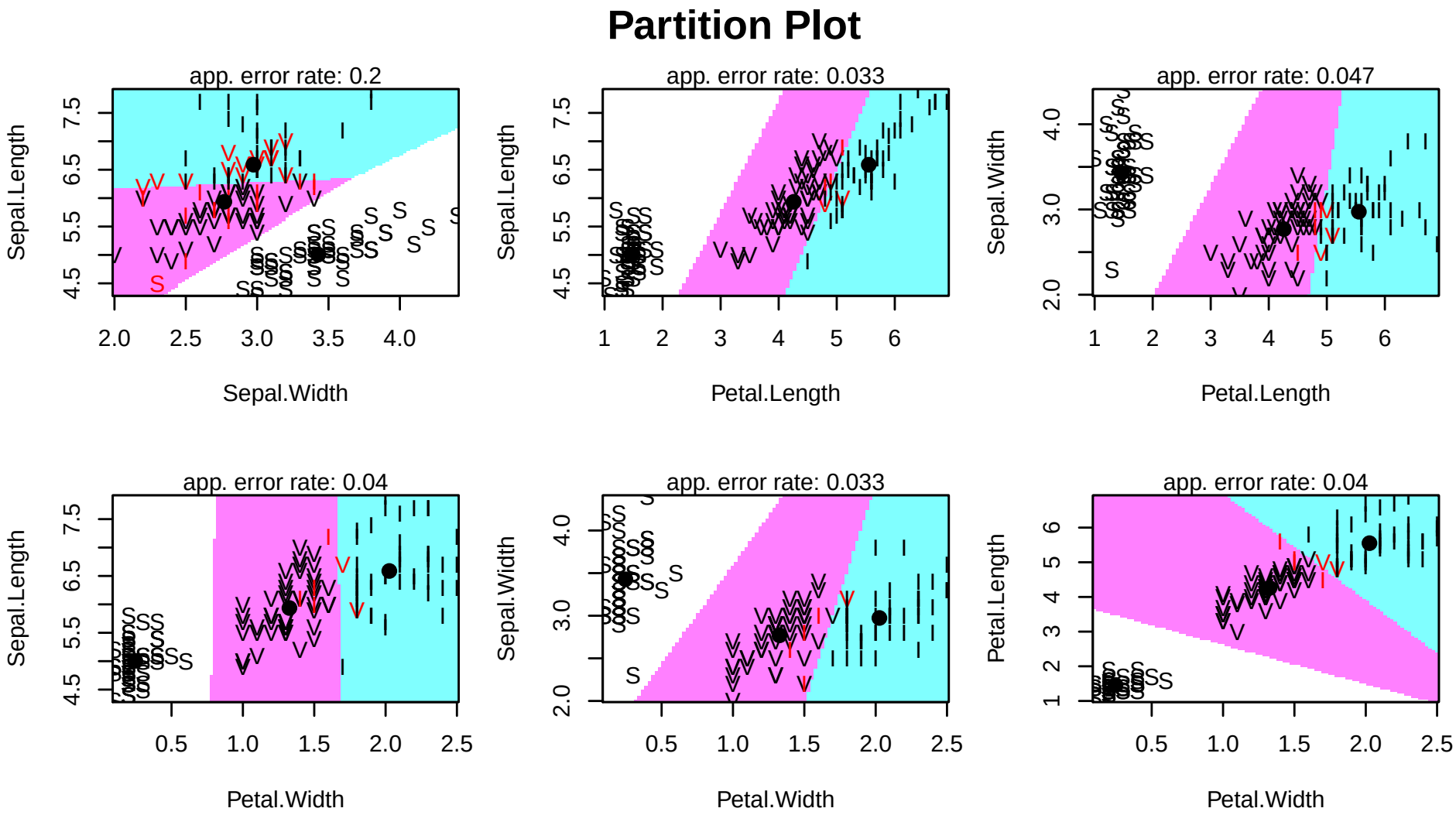


Datos Iris VII

Podemos graficar todos los pares de variables y marcar las regiones:

```
par(mfrow = c(2,3))
partimat(x = iris[,-5], grouping = especies, method = "lda")
```

Datos Iris VIII



Podemos también ver las regiones cuadráticas:

```
par(mfrow = c(2,3))
partimat(x = iris[,-5], grouping = especies, method = "qda")
```

Datos Iris IX