

Minería de Datos

Clasificación Supervisada

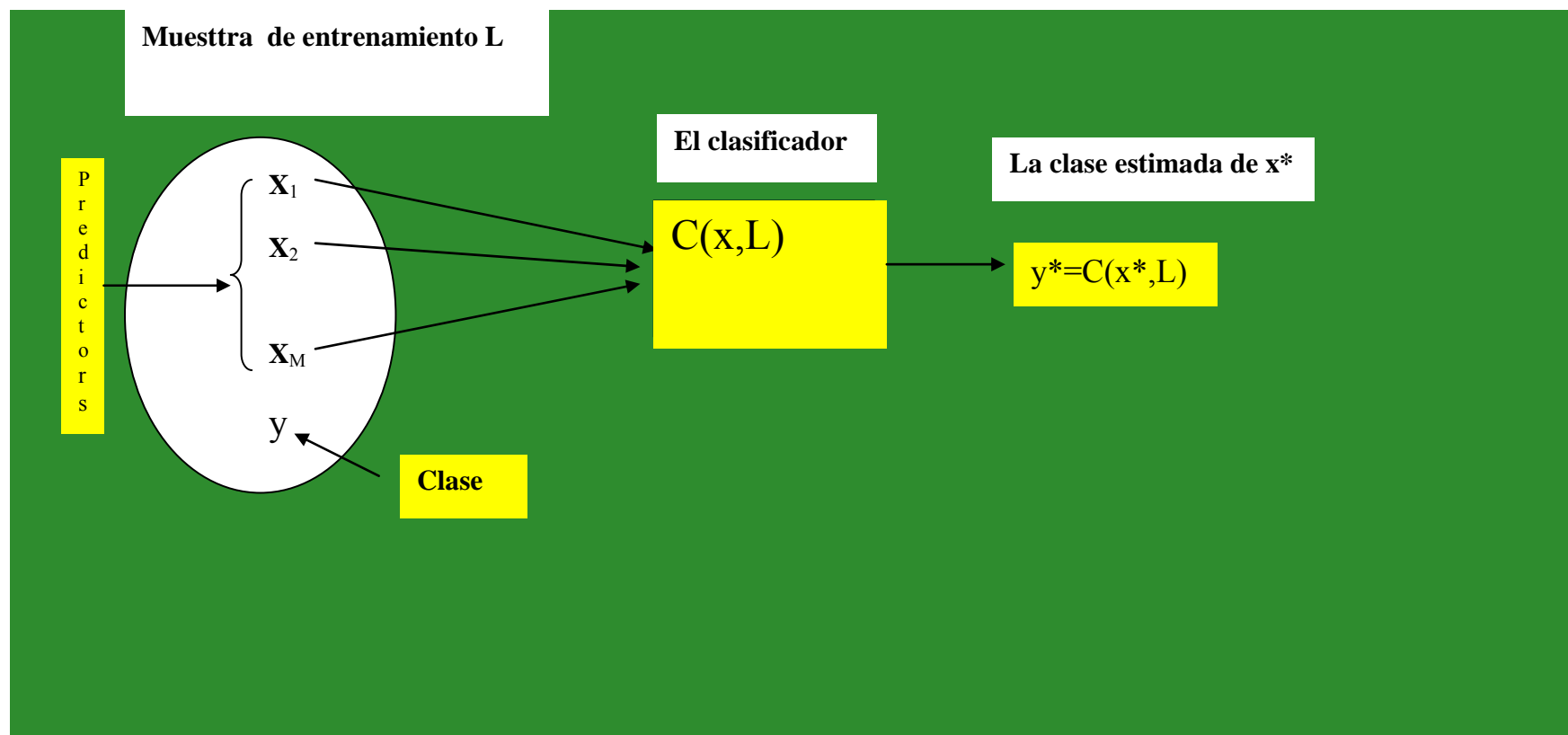
Dr. Edgar Acuña

Departamento de Ciencias Matemáticas
Universidad de Puerto Rico-Mayaguez

E-mail: edgar@math.uprm.edu, eacunaf@gmail.com

Website: math.uprm.edu/~edgar

El problema de clasificacion supervisada



Clasificación Supervisada—un proceso de dos etapas.

- Construcción del modelo:
 - Se asume que cada objeto pertenece a una clase predefinida, según es definido en la columna de clases.
 - El modelo es representado por reglas de clasificación, árboles de decisión, o fórmulas matemáticas.
- Uso del modelo, para clasificar futuros objetos.
 - Estimar la precisión del modelo
 - La columna de clases de los objetos de la muestra de prueba es comparado con las clases predichas por el modelo. La muestra de prueba debe ser independiente de la muestra de entrenamiento de lo contrario puede ocurrir “sobreajuste”.
 - La tasa de precisión es el porcentaje de objetos de la muestra de prueba que son correctamente clasificadas por el modelo.

Clasificación Supervisada vs. Regresión

- **Clasificación Supervisada:**
 - Predice valores categoricos que representan clases.
 - Construye un modelo basado en la muestra de entrenamiento y lo usa para clasificar nuevos datos.
- **Regresión:**
 - Modela funciones de valores continuos. Predice nuevos valores o valores faltantes.
- **Aplicaciones mas comunes**
 - Aprobacion de credito
 - Para detectar posibles mercados de un producto.
 - En diagnosticos medicos.

Metodos de clasificacion supervisada

1. Analisis Discriminante Lineal.
2. Metodos No Lineales: Discriminacion Cuadratica, Regresion Logistica, Projection Pursuit.
3. Naive Bayes y Redes Bayesianas
4. Clasificadores basados en reglas
5. Arboles de Decision.
6. K vecinos mas cercanos.
7. Clasificadores basados en estimacion de densidad por kernel. Clasificadores que usan mezclas Gaussianas.
8. Redes Neurales: El perceptron de multicapas, Funciones bases radiales, mapas auto-organizantes de Kohonen.
9. Maquinas de soporte vectorial.

Analisis Discriminante Lineal

Considerar la siguiente muestra de entrenamiento con p atributos y dos clases

Y	X_1	X_2	...	X_p
1	X_{11}	X_{21}	X_{p1}
1	X_{12}	X_{22}	...	X_{p2}
..
1	X_{1n1}	X_{2n1}	...	X_{pn1}
2	$X_{1,n1+1}$	$X_{2,n1+1}$...	$X_{p,n1+1}$
2	$X_{1,n1+2}$	$X_{2,n1+2}$...	$X_{p,n1+2}$
..
2	$X_{1,n1+n2}$	$X_{2,n1+n2}$...	$X_{p,n1+n2}$

Analisis discriminante lineal

Sea \bar{x}_1 el vector de medias de los p atributos en clase 1, y sea \bar{x}_2 el vector de medias correspondiente para la clase 2.

Considerar que μ_1 y μ_2 son los vectores de media de las respectivas clases poblacionales. Asumir que ambas poblaciones tienen la misma matriz de covarianza, i.e. $\Sigma_1 = \Sigma_2 = \Sigma$. Esta es conocida como la propiedad de homocedasticidad.

Por ahora, no necesitamos asumir que el vector aleatorio de predictoras $\mathbf{x} = (x_1, \dots, x_p)$ está normalmente distribuido.

Discriminación lineal está basada en el siguiente hecho: una instancia (objeto) x es asignado a la clase C_1 si

$$D(\mathbf{x}, C_1) < D(\mathbf{x}, C_2) \quad (2.1)$$

donde $D(\mathbf{x}, C_i) = (\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i)$, for $i=1,2$, representa el cuadrado de la distancia ***squared Mahalanobis de x al centro de la clase C_i*** .

Analisis discriminante lineal (cont)

La expresion (2.1) puede ser escrita como

$$(\mu_1 - \mu_2)' \Sigma^{-1} [x - (1/2)(\mu_1 + \mu_2)] > 0 \quad (2.2)$$

Usando la muestra de entrenamiento, μ_i puede ser estimada por \bar{x}_i y Σ es estimada por S , la matriz de covarianza combinada, la cual es calculada por

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

donde, S_1 y S_2 representan las matrices de covarianza muestrales del vector de predictoras en cada clase. Luego, la version muestral de (2.2) esta dada por

$$(\bar{X}_1 - \bar{X}_2)' S^{-1} [x - (1/2)(\bar{X}_1 + \bar{X}_2)] > 0 \quad (2.3)$$

El lado izquierdo de la expresion (2.3) es llamado la funcion discriminante **lineal**.

Ejemplo: Notas en un curso

En una clase de Estadística Aplicada I consistente de 32 estudiantes, se toman la siguientes variables

E1: Nota del estudiante en el examen 1 (0-100)

E2: Nota del estudiante en el examen 2 (0-100)

PF: Promedio Final del estudiante en la clase (0-100)

Nota: Asume los valores P: si el estudiante pasa la clase y F si fracasa en la clase.

Los primeros 5 datos son:

	E1	E2	PF	Nota
1	96	100	100	p
2	96	94	99	p
3	100	91	97	p
4	93	96	97	p
5	90	94	95	p

Se quiere predecir la Nota de un estudiante que tomará la clase en un próximo semestre, bajo las mismas condiciones actuales (mismo profesor, mismo texto, estudiantes de similar nivel en la clase, etc).

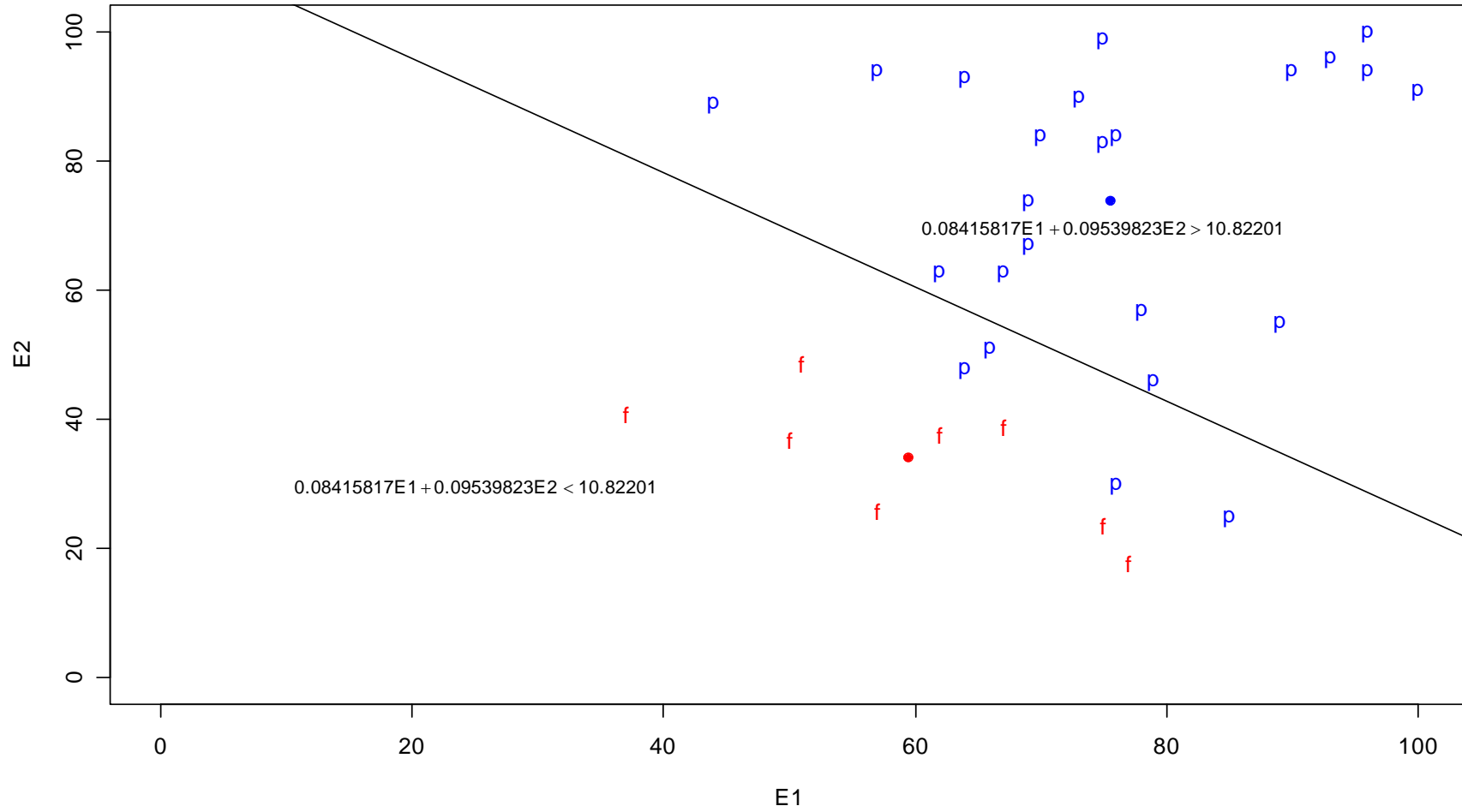
Ejemplo(cont)

```
> #Hallando las medias de E1 y E2 en ambos grupos
> medp=apply(eje1dis[eje1dis[,4]=="p",1:2],2,mean)
> medp
  E1    E2
75.54167 73.75000
> medf=apply(eje1dis[eje1dis[,4]=="f",1:2],2,mean)
> medf
  E1    E2
59.5 34.0
> #Hallando las matrices de covarianzas en ambos grupos
> cov1
      E1    E2
E1 187.56341 34.22826
E2 34.22826 507.06522
> cov2
      E1    E2
E1 183.42857 -97.28571
E2 -97.28571 106.28571
```

Ejemplo(cont)

```
> #Calculando los coeficientes de la funcion discriminante
> covcomb=((npasan-1)*cov1+(nfracasan-1)*cov2)/(npasan+nfracasan-
2)
> coeflda<-(medp-medf)%*%solve(covcomb)
> coeflda
      E1      E2
[1,] 0.08415817 0.09539823
> #Calculando el termino independiente
> indlda<-0.5*(medp-medf)%*%solve(covcomb)%*%(medp+medf)
> indlda
      [,1]
[1,] 10.82201
```

Analisis discriminante Lineal para datos de examenes



LDA(Fisher, 1936)

Fisher obtuvo la función discriminante lineal de la ecuación (2.3) pero siguiendo otro camino. Él trató de encontrar una combinación lineal de los atributos x 's que separe las clases C_1 y C_2 lo más posible asumiendo que hay homogeneidad de matrices de covariancias ($\Sigma_1 = \Sigma_2 = \Sigma$). Más específicamente, si $y = d'x$, entonces, la distancia cuadrada entre las medias de y en cada clase dividida por la varianza de y está dado por

$$\frac{(d' \mu_1 - d' \mu_2)^2}{d' \Sigma d} \quad (2.4).$$

Esta razón es maximizada cuando $d = \Sigma^{-1}(\mu_1 - \mu_2)$. Este resultado es obtenido aplicando la desigualdad de Cauchy-Schwartz (ver Rao, C. R. *Linear Statistical Inference and its applications*, p. 60).

LDA (cont)

- El numerador es llamado la suma of cuadrados entre grupos (BSS), y el denominador es llamado la suma of cuadrados dentro de grupos (WSS). Un estimado del valor d es $S^{-1}(\bar{x}_1 - \bar{x}_2)$.

- Fisher asigno un objeto x a la clase C1 si

$y = d'x = (\bar{x}_1 - \bar{x}_2)' S^{-1} x$ esta mas cerca a $\bar{y}_1 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_1$ que a \bar{y}_2 . El punto medio entre \bar{y}_1 y \bar{y}_2 es

$$\frac{(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2)}{2}$$

- Notar que y es mas cerca a \bar{y}_1 si $y > \frac{\bar{y}_1 + \bar{y}_2}{2}$, esto da la ecuacion (2.3).

Tests for homogeneity of covariance matrices(homocedasticity)

- The most well know test for cheking homocedasticity (homogeneity of covariance matrices) is the Bartlett test. This test is a modification of the likelihood ratio test, however it is subject to the assumption of multivariate normal distribution. It makes use of a Chi-Square distribution. Bartlett est is available in SAS. The Mardia test is one of several test to check multivariate normality.
- Other alternative is to extent the Levene's test for comparing the variance of several univariate populations.
- Some statistcal packages like SPSS use the Box'M test to check homocedasticity).

Mardia's test for multivariate Normality (1970)

Consideremos que \mathbf{x}'_j ($j=1, \dots, n$) representan las observaciones en la muestra de entrenamiento correspondiente a una clase particular C. Si se consideran p variables predictoras entonces cada \mathbf{x}_j es un vector columna p -dimensional. Deseamos probar que el vector aleatorio $\mathbf{X}=(X_1, \dots, X_p)$ se distribuye en forma normal multivariada en C. Mardia basa su prueba en las medidas de asimetría y kurtosis, cuyas estimaciones basadas en la muestra de entrenamiento están definidas como :

y

$$b_1 = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n \{(\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_k - \bar{\mathbf{x}})\}^3$$

respectivamente.

$$b_2 = \frac{1}{n} \sum_{j=1}^n \{(\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})\}^2$$

Mardia's test (cont)

Si la hipótesis nula H_0 : \mathbf{x} es normal multivariada en la clase C es cierta entonces se puede mostrar que para n grande

$$(n/6)b_1 \sim \chi_d^2$$

con $d=(p/6)(p+1)(p+2)$ grados de libertad, y

$$[b_2 - p(p+2)]/\sqrt{(8/n)p(p+2)} \sim N(0,1)$$

- La prueba de Hawkins (Technometrics, 1981) permite probar simultáneamente normalidad multivariada y homocedasticidad. No aparece en ningún programa estadístico.

Example: Bupa

```
> mardia(bupa,1)
mard1= 420.9489
pvalue for m3= 0
mard2= 15.91613
p-value for m4= 0
There is not statistical evidence for normality
> mardia(bupa,2)
mard1= 1178.14
pvalue for m3= 0
mard2= 37.50413
p-value for m4= 0
There is not statistical evidence for normality
```

Clasificación supervisada desde un punto de vista Bayesiano

Suponga que conocemos de antemano la probabilidad a priori π_i ($i=1, 2, \dots, G$) de que un objeto pertenezca a la clase C_i . Si no se conoce información adicional entonces un objeto cualquiera será clasificado como perteneciente a la clase C_i si

$$\pi_i > \pi_j \text{ para } i=1, 2, \dots, G, j \neq i \quad (3.1)$$

Sin embargo, usualmente alguna información adicional es conocida, tal como un vector \mathbf{x} de mediciones hechas en el objeto a ser clasificado. En este caso, comparamos la probabilidad de pertenecer a cada clase para un objeto con vector de mediciones \mathbf{x} , y el objeto es clasificado como de clase C_i si

$$P(C_i/\mathbf{x}) > P(C_j/\mathbf{x}) \text{ para todo } j \neq i \quad (3.2)$$

Esta regla de decisión es llamada la Regla de **minimo error**.

Esto es equivalente a escribir, $i = \operatorname{argmax}_k P(C_k/\mathbf{x})$ para todo k en $1, 2, \dots, G$.

Clasificación Bayesiana

Las probabilidades $P(C_i/x)$ son llamadas probabilidades posteriores. Desafortunadamente, estas probabilidades raras veces son conocidas y deben ser estimadas. Esto ocurre en regresión logística, clasificadores por vecinos más cercanos, y redes neurales.

Una formulación más conveniente de la regla anterior puede ser obtenida aplicando el teorema de Bayes, que afirma que

$$P(C_i / \mathbf{x}) = \frac{f(\mathbf{x} / C_i)\pi_i}{f(\mathbf{x})} \quad (3.3)$$

donde, $f(\mathbf{x} / C_i)$, es la densidad condicional de la clase C_i . Por lo tanto un objeto será clasificado en la clase C_i si

$$f(\mathbf{x} / C_i)\pi_i > f(\mathbf{x} / C_j)\pi_j \quad (3.4)$$

para todo $i \neq j$. Esto es, $i = \operatorname{argmax}_k f(\mathbf{x} / C_k)\pi_k$

Si las densidades condicionales de clase, $f(\mathbf{x} / C_i)$ son conocidas entonces el problema de clasificacion esta resuelto explicitamente, como ocurre en el analisis discriminante lineal y cuadratico.

Pero frecuentemente las $f(\mathbf{x} / C_i)$ no son conocidas y ellas deben ser estimadas usando la muestra de entrenamiento. Este es el caso de los clasificadores k -nn, los clasificadores basados en estimacion de densidad por kernel y en aquellos basados en mezclas gaussianas.

Analisis discriminante lineal como un clasificador Bayesiano

Supongamos que tenemos dos clases C_1 y C_2 que siguen una distribución normal multivariada, $Np(\mathbf{u}_1, \Sigma_1)$ y $Np(\mathbf{u}_2, \Sigma_2)$ respectivamente y que además tienen igual matriz de covarianza $\Sigma_1 = \Sigma_2 = \Sigma$. Luego la ecuación (3.4) puede ser escrita como

$$\frac{\exp[-1/2(\mathbf{x} - \mathbf{u}_1)' \Sigma^{-1} (\mathbf{x} - \mathbf{u}_1)]}{\exp[-1/2(\mathbf{x} - \mathbf{u}_2)' \Sigma^{-1} (\mathbf{x} - \mathbf{u}_2)]} > \frac{\pi_2}{\pi_1} \quad (3.5)$$

Tomando logaritmos en ambos lados se obtiene

$$-1/2[(\mathbf{x} - \mathbf{u}_1)' \Sigma^{-1} (\mathbf{x} - \mathbf{u}_1) - (\mathbf{x} - \mathbf{u}_2)' \Sigma^{-1} (\mathbf{x} - \mathbf{u}_2)] > Ln\left(\frac{\pi_2}{\pi_1}\right) \quad (3.6)$$

Despues de algunas simplificaciones resulta

$$(\mathbf{u}_1 - \mathbf{u}_2)' \Sigma^{-1} (\mathbf{x} - \frac{1}{2}(\mathbf{u}_1 + \mathbf{u}_2)) > \ln\left(\frac{\pi_2}{\pi_1}\right) \quad (3.7)$$

Si estimamos los parametros poblacionales, esta desigualdad es similar a la que aparece en (2.2), excepto por el termino del lado derecho. Si ademas consideramos que las probabilidades a priori son iguales ($\pi_1 = \pi_2$) entonces si se obtendria la misma expresion (2.2).

Ejemplo(cont)

```
> # LDA con priors iguales
> lda1<-lda(Nota~E1+E2,eje1dis,prior=c(.5,.5))
> lda1
> plda1=predict(lda1,eje1dis[,-c(3,4)])$class
> error1=sum(plda1!=eje1dis[,4])
> error1
[1] 4
> #LDA con priors ponderadas
> lda2<-lda(Nota~E1+E2,eje1dis)
Prior probabilities of groups:
  f  p
0.25 0.75
> plda2=predict(lda2,eje1dis[,-c(3,4)])$class
> error2=sum(plda2!=eje1dis[,4])
> error2
[1] 2
```


LDA para mas de dos clases

Para G clases, el LDA asigna un objeto con vector de atributos \mathbf{x} a la clase i tal que

$$i = \operatorname{argmax}_k \bar{\mathbf{x}}_k' \mathbf{S}^{-1} \mathbf{x} - (1/2) \bar{\mathbf{x}}_k' \mathbf{S}^{-1} \bar{\mathbf{x}}_k + \ln(\pi_k)$$

Para todo k en $1, 2, \dots, G$. Al igual que antes el lado derecho, llamada **funcion discriminante**, es estimada usando la muestra de entrenamiento.

Ejemplo: El conjunto de datos Vehicle

```
Library(MASS)
ldaveh=lda(vehicle[,1:18],vehicle[,19])
predict(ldaveh)$posterior
predict(ldaveh)$class
# Estimating the error rate
mean(vehicle[,19]!=predict(ldaveh)$class)
[1] 0.2021277
```

Es estimado que 20.21% de las instancias estan mal clasificadas.

La tasa de error de mala clasificación

Dado un clasificador d , su tasa de error de mala clasificación $R(d)$ es la probabilidad de que d clasifique incorrectamente un objeto de una muestra (muestra de prueba) obtenida posteriormente a la muestra de entrenamiento.

También es llamado el error verdadero.

Es un valor desconocido que necesita ser estimado.

Metodos para estimar la taza de error de mala clasificacion

- i) **Error por resubstitution or error aparente.** (Smith, 1947).
Este es simplemente la proporcion de instancias en la muestra de entrenamiento que son incorrectamente clasificadas por la regla de clasificacion. En general es un estimador demasiado optimista y puede conducir a conclusiones erroneas cuando el numero de instancias no es demasidado grande comparado con el numero de atributos. Este estimador tiene un sesgo grande.

- ii) **Estimacion dejando uno afuera (LOO).** (Lachenbruch, 1965).
En este caso una instancia es omitida de la muestra de entrenamiento. Luego, el clasificador es construido y se obtiene la prediccion para la instancia omitida. Se registra si la instancia fue correcta o incorrectamente clasificada. El proceso es repetido para todas las instancias y la estimacion del error de mala clasificacion sera la proporcion de instancias clasificadas inconrrectamente. Este estimador tiene poco sesgo pero su varianza tiende a ser grande.

Examples of LOO

```
> ldaexa=lda(eje1dis[,1:2],eje1dis[,4],CV=TRUE)
> mean(eje1dis[,4]!=ldaexa$class)
[1] 0.0625
> ldaveh=lda(vehicle[,1:18],vehicle[,19],CV=TRUE)
> mean(vehicle[,19]!=ldaveh$class)
[1] 0.2210402
```

Metodos para estimar la tasa de error de mala clasificacion

iii) Validacion cruzada. (Stone, 1974) En este caso la muestra de entrenamiento es dividida al azar en v partes ($v=10$ es lo mas usado). Luego, el clasificador es construido usando todas las partes menos una. La parte omitida es considerada como la muestra de prueba y se hallan las predicciones de cada una de sus instancias. La tasa de error de mala clasificacion CV es hallada sumando el numero de malas clasificaciones en cada parte y dividiendo el total por el numero de instancias en la muestra de entrenamiento. El estimado CV tiene poco sesgo pero una alta variabilidad. Para reducir dicha variabilidad se repite la estimacion varias veces.

La estimacion de la varianza del estimador CV es un “hard problem” (Bengio and Grandvalet, 2004).

Ejemplo:

```
crossval(eje1dis,method="lda",repet=20)
```

```
[1] 0.096875
```

```
crossval(vehicle,method="lda",repet=10)
```

```
[1] 0.2206856
```

iv) El metodo de retencion. Aqui se extrae al azar un porcentaje de los datos (70%) y es considerado como la muestra de entrenamiento y los restantes datos son considerados como la muestra de prueba. El clasificador es evaluado en la muestra de prueba. El experimento es repetido varias veces y luego se promedia la tasa de error en las muestras de prueba.

v) Bootstrapping. (Efron, 1983). En este metodo se generan de la muestra de entrenamiento varias muestras con reemplazo. Las que luego se usan para construir el clasificador. La idea es reducir el sesgo del error aparente. Es casi insesgado pero tiene una varianza grande. Su costo computacional es alto.

Existen varias variantes de este metodo.