UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Unidad Académica de los Ciclos Profesional y de Posgrado del CCH
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

ANALISIS DISCRIMINANTE CON COSTOS ASOCIADOS A VARIABLES

T E S I S

Que para obtener el grado de:
Maestra en Estadística e Investigación de Operaciones
Presenta la Actuaria
Patricia I. Romero Mares

México, D. F. Marzo 1989

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

INTRODUCCION

El Análisis Discriminante ha tenido aplicación en una amplia gama de áreas del conocimiento, como son la Biología, la Pedagogía, la Medicina, en las cuales se pretende diferenciar poblaciones bajo estudio. Ha sido largamente estudiado bajo diversos enfoques, empezando con el enfoque propuesto por Fisher, en los años treinta, con su función Discriminante Lineal, pasando después al enfoque provisto por la Teoría de Decisiones, hasta el enfoque puramente bayesiano, en el caso en que se desconocen los parámetros de las poblaciones y a ellos se les supone una distribución.

La bibliografía sobre el tema de Análisis Discriminante es muy amplia y se tienen resueltos muchos de los problemas que surgen en la aplicación de estos métodos; sin embargo, se ha estudiado muy poco el problema de discriminación cuando es necesario considerar los costos de medición de las variables involucradas. Este es un problema práctico y que aparece casi en cualquier investigación. El investigador se enfrenta a la decisión de si medir o no una variable que puede tener un gran potencial discriminatorio pero que a la vez pudiera ser muy cara de observar. En cuanto al costo de medición, éste representa no solamente el costo monetario de obtención de la información en sí, sino también la dificultad práctica de efectuar la observación.

El Análisis Discriminante Clásico, que utiliza la Función Discriminante de Fisher, no toma en cuenta el costo de las variables involucradas, y tal vez una primera

aproximación a la solución del problema de costos sea el aplicar algún método de selección de variables. Existen diversos métodos para selección de variables que se encuentran implantados en casi cualquiera de los paquetes estadísticos.

El propósito de este trabajo es el dar a conocer un método nuevo para efectuar Análisis Discriminante; esto es, para la asignación de individuos a poblaciones, cuando resulta muy importante el considerar el costo de medición de las observaciones que le son hechas a cada individuo antes de ser asignados.

El trabajo se divide en tres capítulos:

El primero presenta una visión general del Análisis Discriminante bajo el enfoque de la Teoría de Decisiones. Se presentan las soluciones óptimas para el caso general, y como caso particular, se presentan las soluciones para el caso normal. Someramente se comentan los problemas que aparecen cuando las poblaciones no están completamente determinadas, asimismo se mencionan algunos procedimientos para la selección de las variables más importantes en el proceso de discriminación.

El capítulo dos trata sobre procedimientos secuenciales ya existentes. Entre estos procedimientos se comentan los de Wald (1948) y los de Fu (1968). Se enfatiza la complejidad numérica que aparece al tratar de implantar soluciones óptimas. Dicha complejidad numérica, que llega a convertir en no factible un procedimiento, dá lugar a la búsqueda de soluciones subóptimas. Estas también se comentan.

Finalmente, el capítulo tres muestra el procedimiento propuesto en este trabajo. Se trata con detalle el algoritmo respectivo, y los problemas surgidos de su implantación en una microcomputadora. Se proporcionan los resultados que se obtuvieron al probar el método con los datos clásicos de Iris de Fisher, al igual que con los datos resultantes, a través de simulación, de transformaciones realizadas al modelo normal subyacente en los datos de Fisher. Lo anterior, con el objeto de evaluar comparativamente el método propuesto.

CAPITULO

1

Análisis Discriminante

1.1 Introducción

En muchas áreas de investigación es frecuente encontrarse con el problema de que un objeto o individuo debe ser asignado a una de varias poblaciones, es decir, se desea categorizar un objeto con base en un perfil de sus características.

Como ejemplos de tales situaciones se tiene:

- Diagnóstico de personas que padecen o no cierta enfermedad (diabetes, por ejemplo), basado en diferentes pruebas clínicas como presión sanguínea, cantidad de colesterol en la sangre, etc.

- Pertenencia de un individuo a una entre ciertas clases de neurosis determinadas, con base en medidas como estado de ansiedad, sentimiento de culpa, etc.

Debe notarse en los ejemplos que las poblaciones o categorías están definidas de antemano y lo que se desea es asignar un nuevo individuo a una de ellas.

Cada población está caracterizada por una distribución de probabilidad de las mediciones y a un individuo se le considera como una observación aleatoria de la población a la que pertenece. En algunos casos las distribuciones de probabilidad son totalmente conocidas, mientras que en otros se puede conocer la forma de la

distribución pero se deben estimar los parámetros a partir de una muestra de cada población. Existen también situaciones en que no se puede suponer una forma de distribución y entonces se utilizan técnicas no paramétricas, pero estos casos no serán tratados en el presente trabajo.

El Análisis Discriminante tiene varios objetivos:

1. Probar hipótesis de que no hay diferencia entre las poblaciones consideradas.

2. Caracterizar, a través de las mediciones, las diferencias entre las poblaciones.

3. Asignar un nuevo individuo a uno de varios grupos conocidos.

En el presente trabajo se tratará solamente con este último objetivo, es decir, el problema es el asignar uno o varios individuos "nuevos" a una de varias poblaciones conocidas.

1.2 Planteamiento Formal del Problema

El Análisis Discriminante ha sido estudiado largamente y bajo muy diversos enfoques, como expresó el Dr. B. Wagle en la discusión del artículo de Hills (1966):

"El objetivo del Análisis Discriminante se ha desarrollado a través de tres etapas. La primera fué la etapa Fisheriana usando un enfoque intuitivo y desarrollando la teoría de las funciones discriminantes lineales. A ésta siguió la etapa probabilística considerada por Welch, Rao y otros. La tercera etapa fué la etapa Waldiana basada en los principios de la teoría de decisiones estadísticas. Estas etapas suponen básicamente poblaciones normales multivariadas cuando tratan con problemas numéricos y reemplazan a los parámetros desconocidos por estimadores muestrales...".

El enfoque que se tratará en este trabajo es el de teoría de decisiones y se mencionará el enfoque bayesiano en el caso de poblaciones no completamente especificadas.

1.2.1 Caso General

Para el caso de discriminación que nos ocupa, el enfoque de Teoría de Decisiones es el siguiente:

Sea Θ el espacio de estados de la naturaleza compuesto por p poblaciones

$$\Theta = \{\Pi_1, \Pi_2, \ldots, \Pi_p\}.$$

Sea $D = \{d_1, d_2, \ldots, d_p\}$ el espacio de decisiones, donde d_i representa la decisión de asignar el individuo a la población Π_i.

Se define una función de pérdida, $L(d, \Pi)$, que simboliza la pérdida en que se incurre al tomar la decisión d cuando el individuo pertenece realmente a la población Π. Nótese que $L: D \times Theta \rightarrow \Re$.

a. Caso sin observaciones.

Si no se tiene información a priori acerca de la distribución de frecuencias de las Π_i, se utiliza el criterio Minimax, buscando aquella decisión $d^* \in D$ que minimice la máxima pérdida.

d^* es minimax si

$$\max_i L(d^*, \Pi_i) \leq \max_i L(d, \Pi_i) \quad \forall d.$$

Si se dispone de información a priori acerca de la distribución de frecuencias de Π, ésta se incorpora a través de q_i que es la probabilidad de pertenencia a la población Π_i, $i = 1, \ldots, p$. Se calcula la pérdida esperada con respecto a la densidad a priori para cada decisión d_j,

$$E_q\left[L(d_j,\Pi)\right]=\sum_{i=1}^{p}L(d_j,\Pi_i)q_i. \qquad (1.1)$$

Se sabe que la decisión que minimiza la pérdida esperada es la decisión de Bayes (Wald (1950)).

d^{**} es decisión de Bayes si

$$\sum_{i=1}^{p}L(d^{**},\Pi_i)q_i \leq \sum_{i=1}^{p}L(d,\Pi_i)q_i \quad \forall d.$$

b. Caso con observaciones.

Ahora, suponga que se tiene disponible un vector aleatorio X con m componentes $X = (X_1, X_2, \ldots, X_m)$, con función de densidad $f(X|\Pi_i)$ completamente especificada, $i = 1, \ldots, p$.

El problema cambia al de elegir una decisión $d \in D$ después de haber observado X. Para esto es necesario encontrar una función de decisión $d(X)$, que para cada valor de X le asigne un valor $d \in D$. Nótese que $d: \Omega_X \rightarrow D$, $d(X) = d_j \Rightarrow$ el individuo se asigna a la población Π_j. Ω_X es el espacio muestral de las observaciones.

Si se elige una decisión d, la pérdida depende de X a través de $L(d(X), \Pi)$.

En promedio, se tiene una pérdida esperada o riesgo, condicionada a Π de la

siguiente forma:

$$R(d,\Pi) = E_f\left[L(d(X),\Pi)\right] = \int_{\Omega_x} L(d(X),\Pi) f(X|\Pi) dX$$

En este caso, si no se conocen las probabilidades a priori de pertenencia a cada población, se utiliza el criterio Minimax, buscando d^* tal que

$$\max_i R(d^*,\Pi_i) \leq \max_i R(d,\Pi_i) \quad \forall d.$$

Si se conocen las probabilidades a priori, q_i, se busca la función de decisión de Bayes, d^{**}, que minimice el riesgo de Bayes (Wald(1950)),

$$\begin{aligned} B(d) &= E_q\left[R(d,\Pi)\right] \\ &= \sum_{i=1}^{p} \left\{ \int_{\Omega_X} L\left[d(X),\Pi_i\right] f(X|\Pi_i) dX \right\} q_i \\ &= \int_{\Omega_X} \left\{ \sum_{i=1}^{p} L\left[d(X),\Pi_i\right] q_{i|X} \right\} f(X) dX, \end{aligned} \tag{1.2}$$

donde $f(X) = \sum_{i=1}^{p} q_i f(X|\Pi_i)$ y $q_{i|X}$ es la probabilidad posterior de pertenencia a la población Π_i, y, por el Teorema de Bayes, es:

$$f(X|\Pi_i) q_i = q_{i|X} f(X).$$

d^{**} es función de decisión Bayes si:

$$E_q\left[R(d^{**},\Pi)\right] \leq E_q\left[R(d,\Pi)\right] \quad \forall d.$$

De (1.2) se puede demostrar que para que d minimice el Riesgo de Bayes, $B(d)$, es necesario y suficiente que d minimice

$$\begin{aligned} R_X(d) &= \sum_{i=1}^{p} L\left[d(X), \Pi_i\right] q_{i|X} \\ &\propto \sum_{i=1}^{p} L\left[d(X), \Pi_i\right] q_i f(X|\Pi_i) \end{aligned} \tag{1.3}$$

$R_X(d)$ es el riesgo a posteriori de Bayes y depende de d y de X.

Una forma de calcular la función de decisión $d(X)$ que minimice el riesgo a posteriori de Bayes es determinando una partición $B = \{B_1, B_2, \ldots, B_p\}$ del espacio muestral Ω_X, donde la decisión es asignar el individuo a la población Π_j si X cae en la región B_j, (Anderson(1958)).

Se selecciona $d(X)$ que haga mínimo el riesgo de Bayes a posteriori para cada X, definiendo así las regiones $B_1, \ldots, B_p$.

Anderson(1958) demuestra que el procedimiento de clasificación es como sigue:

Sea $L(d_j, \Pi_i) = C(j|i)$ el costo de asignar a Π_j cuando en realidad el individuo pertenece a Π_i.

Las regiones de clasificación quedan definidas como:

$$B_k : \left\{ X \middle| \sum_{\substack{i=1 \\ i \neq k}}^{p} q_i f(X|\Pi_i) C(k|i) < \sum_{\substack{i=1 \\ i \neq j}}^{p} q_i f(X|\Pi_i) C(j|i), \quad j = 1, \ldots, m, \quad j \neq k \right\} \tag{1.4}$$

y se asigna el individuo a Π_k si su correspondiente vector de mediciones X cae en B_k, $k = 1, \ldots, p$.

Si los costos de clasificación son de la forma:

$$C(j|i) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases} \tag{1.5}$$

entonces

$$B_k : \left\{ X | \sum_{\substack{i=1 \\ i \neq k}}^{p} q_i f(X|\Pi_i) < \sum_{\substack{i=1 \\ i \neq j}}^{p} q_i f(X|\Pi_i), \quad j = 1, \ldots, m, \quad j \neq k \right\}$$

restando $\sum_{\substack{i=1 \\ i \neq k,j}}^{p} q_i f(X|\Pi_i)$ en ambos lados, tenemos que

$$\begin{aligned} B_k &: \left\{ X | q_j f(X|\Pi_j) < q_k f(X|\Pi_k) \quad \forall j \neq k \right\} \\ &= \left\{ X | q_{k|X} > q_{j|X} \quad \forall j \neq k \right\}, \quad k = 1, \ldots, p. \end{aligned} \tag{1.6}$$

Se observa que B_k está definida para los puntos X en los cuales Π_k es la población más probable a posteriori.

Sea

$$P(j|i, B) = \int_{B_j} f(X|\Pi_i) dX,$$

la probabilidad de clasificar en Π_j cuando la verdadera población es Π_i, dada la partición B. Note que

$$P(i|i, B) = \int_{B_i} f(X|\Pi_i) dX,$$

es la probabilidad de clasificar correctamente en la población Π_i, dada la partición B.

El riesgo condicional a que $X \sim f(X|\Pi_i)$ es

$$R_{\Pi_i}(d) = \int_{\Omega_X} L\left[d(X), \Pi_i\right] f(X|\Pi_i) dX.$$

Para costos de clasificación de la forma (1.5) se tiene que

$$R_{\Pi_i}(d) = \sum_{\substack{j=1 \\ j \neq k}}^{p} P(j|i, B) = 1 - P(i|i, B). \tag{1.7}$$

Otra forma de expresar el riesgo de Bayes, $B(d)$, es sustituyendo (1.7) en (1.2) quedando:

$$\begin{aligned} B(d) &= \sum_{i=1}^{p} q_i R_{\Pi_i}(d) \\ &= \sum_{i=1}^{p} q_i \left[1 - P(i|i, B)\right] \\ &= 1 - \sum_{i=1}^{p} q_i P(i|i, B). \end{aligned} \tag{1.8}$$

Se observa de (1.8) que en el caso de tener función de pérdida de la forma (1.5), el riesgo de Bayes es la probabilidad de clasificación equivocada.

1.2.2 Caso Normal

Suponga $p = 2$ poblaciones normales $N\left(\underline{\mu}^{(i)}, \Sigma_i\right) \quad i = 1, 2$, con parámetros conocidos.

La i-ésima densidad es:

$$f(X|\Pi_i) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left(X-\underline{\mu}^{(i)}\right)'\Sigma_i^{-1}\left(X-\underline{\mu}^{(i)}\right)\right] \quad i=1,2.$$

La regla de clasificación está dada en (1.4) e indica que se asigne el individuo a Π_1 si

$$q_1 f(X|\Pi_1)C(2|1) > q_2 f(X|\Pi_2)C(1|2)$$

que es equivalente a

$$\frac{f(X|\Pi_1)}{f(X|\Pi_2)} > \frac{C(1|2)}{C(2|1)}\frac{q_2}{q_1} \tag{1.9}$$

entonces, en el caso de densidades normales, la regla de clasificación queda como:

$$\frac{|\Sigma_1|^{-1/2} exp\left[-\frac{1}{2}\left(X-\underline{\mu}^{(1)}\right)'\Sigma_1^{-1}\left(X-\underline{\mu}^{(1)}\right)\right]}{|\Sigma_2|^{-1/2} exp\left[-\frac{1}{2}\left(X-\underline{\mu}^{(2)}\right)'\Sigma_2^{-1}\left(X-\underline{\mu}^{(2)}\right)\right]} > \frac{C(1|2)}{C(2|1)}\frac{q_2}{q_1}.$$

Ya que el logaritmo es una función monótona creciente, la regla de clasificación, en términos de los logaritmos, queda: asigne el individuo en Π_1 si

$$-\frac{1}{2}\left[\left(X-\underline{\mu}^{(1)}\right)'\Sigma_1^{-1}\left(X-\underline{\mu}^{(1)}\right) - \left(X-\underline{\mu}^{(2)}\right)'\Sigma_2^{-1}\left(X-\underline{\mu}^{(2)}\right)\right] > log\left[\frac{|\Sigma_1|^{1/2}}{|\Sigma_2|^{1/2}}\frac{C(1|2)}{C(2|1)}\frac{q_2}{q_1}\right] \tag{1.10}$$

Matrices de Covarianza Iguales.

Si las matrices de covarianza son iguales, es decir $\Sigma_1 = \Sigma_2$, se simplifican las fórmulas y se tienen los siguientes resultados.

Expandiendo los términos del lado izquierdo y rearregiándolos, la desigualdad (1.10) queda como

$$X'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)-\frac{1}{2}\left(\underline{\mu}^{(1)}+\underline{\mu}^{(2)}\right)'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)>log\left[\frac{C(1|2)}{C(2|1)}\frac{q_2}{q_1}\right]$$

Por lo tanto, las regiones de clasificación están dadas por:

$$B_1:\left\{X|X'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)-\frac{1}{2}\left(\underline{\mu}^{(1)}+\underline{\mu}^{(2)}\right)'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)>log\left[\frac{C(1|2)}{C(2|1)}\frac{q_2}{q_1}\right]\right\}$$

y

$$B_2:\left\{X|X'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)-\frac{1}{2}\left(\underline{\mu}^{(1)}+\underline{\mu}^{(2)}\right)'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)<log\left[\frac{C(1|2)}{C(2|1)}\frac{q_2}{q_1}\right]\right\}.$$

Estas regiones son las mejores en el sentido que minimizan el riesgo esperado. Nótese que el primer término del lado izquierdo de estas dos desigualdades es la bien conocida función discriminante de Fisher que es una combinación lineal del vector de observaciones X.

Si los costos de clasificación equivocada son iguales y las probabilidades a priori de pertenencia a cada población son iguales, las regiones se pueden reexpresar en la forma:

$$B_1:\left\{X|X'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)\geq\frac{1}{2}\left(\underline{\mu}^{(1)}+\underline{\mu}^{(2)}\right)'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)\right\}$$

y (1.11)

$$B_2:\left\{X|X'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)<\frac{1}{2}\left(\underline{\mu}^{(1)}+\underline{\mu}^{(2)}\right)'\Sigma^{-1}\left(\underline{\mu}^{(1)}-\underline{\mu}^{(2)}\right)\right\}$$

En el caso en que no se conocen las probabilidades a priori, se utiliza el criterio Minimax, haciendo iguales las pérdidas esperadas debidas a clasificación equivocada para las dos poblaciones.

Sea

$$U = X'\Sigma^{-1}\left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right) - \frac{1}{2}\left(\underline{\mu}^{(1)} + \underline{\mu}^{(2)}\right)'\Sigma^{-1}\left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right) \qquad (1.12).$$

Bajo el supuesto de que X es una observación aleatoria distribuida, primero como $N\left(\underline{\mu}^{(1)}, \Sigma\right)$ y después como $N\left(\underline{\mu}^{(2)}, \Sigma\right)$. Nótese que U es equivalente a la Función Discriminante Lineal de Fisher. Cuando X se distribuye como $N\left(\underline{\mu}^{(1)}, \Sigma\right)$, $U \sim N(1/2\alpha, \alpha)$ donde $\alpha = \left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right)'\Sigma^{-1}\left(\underline{\mu}^{(1)} - \underline{\mu}^{(2)}\right)$.

Si X se distribuye como $N\left(\underline{\mu}^{(2)}, \Sigma\right)$ entonces $U \sim N(-1/2\alpha, \alpha)$.

La probabilidad de mala clasificación si la observación es de la población Π_1 es:

$$P(2|1) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2\alpha}(z-\frac{1}{2}\alpha)^2} dz$$

y la probabilidad de mala clasificación si la observación es de la población Π_2 es:

$$P(1|2) = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2\alpha}(z+\frac{1}{2}\alpha)^2} dz$$

Para encontrar la solución Minimax, escogemos c tal que:

$$C(1|2)\int_{(c+\frac{1}{2}\alpha)/\sqrt{\alpha}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = C(2|1)\int_{-\infty}^{(c-\frac{1}{2}\alpha)/\sqrt{\alpha}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy.$$

Para el caso en que se tienen p poblaciones con distribución $N\left(\underline{\mu}^{(i)}, \Sigma\right)$, $i = 1, \ldots, p$, sea

$$u_{jk}(X) = log\left[\frac{f(X|\Pi_j)}{f(X|\Pi_k)}\right] \qquad (1.13)$$
$$= \left[X - \frac{1}{2}\left(\underline{\mu}^{(j)} + \underline{\mu}^{(k)}\right)\right]' \Sigma^{-1}\left(\underline{\mu}^{(j)} - \underline{\mu}^{(k)}\right)$$

. Suponga además, que los costos de mala clasificación son iguales, siguiendo la regla de clasificación dada por (1.6)

$$B_j : \left\{X|u_{jk}(X) > log\frac{q_k}{q_j}, \quad k = 1, \ldots, p, \quad k \neq j\right\} \qquad j = 1, \ldots, p$$

Si no se conocen las probabilidades a priori $q_1, \ldots, q_p$, el procedimiento Minimax define las regiones de la siguiente manera:

$$B_j : \left\{X|u_{jk}(X) > C_j - C_k, \quad k = 1, \ldots, p \quad k \neq j\right\} \quad j = 1, \ldots, p$$

con C_i, $i = 1, \ldots, p$ determinadas de tal manera que $P(i|i, B)$, $i = 1, \ldots, p$ sean iguales, (Anderson, 1958).

1.3 Solución cuando las Poblaciones no están Completamente Especificadas

Suponga ahora (que puede ser el caso más común), que no se conocen completamente las densidades de las poblaciones, es decir, se sabe que la densidad de la población Π_i es $f(X|\theta_i)$ pero se desconoce el vector de parámetros θ_i.

En el tratamiento de este problema, existen básicamente dos enfoques que se tratarán en forma ilustrativa. El nombre de los dos enfoques está tomado del artículo de Aitchison et. al.(1977).

1.3.1 Enfoque Estimativo

Básicamente, lo que se hace es estimar la función de densidad $f(X|\theta_i)$, a través de reemplazar los parámetros desconocidos por estimadores de ellos.

a) Método de Inserción.

Se sustituyen todos los parámetros desconocidos por estimadores de ellos basados en muestras $x_1^{(i)}, \ldots, x_{n_i}^{(i)}$ de tamaño n_i de cada población Π_i, $i = 1, \ldots, p$.

Así,

$$\hat{f}(X|\theta_i) = f(X|\hat{\theta}_i)$$

La región definida en (1.4) queda estimada como

$$\hat{B}_k : \left\{ X | \sum_{\substack{i=1 \\ i \neq k}}^{p} q_i f(X|\hat{\theta}_i) C(k|i) < \sum_{\substack{i=1 \\ i \neq j}}^{p} q_i f(X|\hat{\theta}_i) C(j|i), \quad j = 1, \ldots, m, \quad j \neq k \right\},$$

$k = 1, \ldots, p.$

En el caso normal con solo dos poblaciones, las regiones definidas en (1.11) se convierten en:

$$\hat{B}_1 : \left\{ X | X' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)} \right) \geq \frac{1}{2} \left(\bar{x}^{(1)} + \bar{x}^{(2)} \right)' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)} \right) \right\}$$

y

$$\hat{B}_2 : \left\{ X | X' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)} \right) < \frac{1}{2} \left(\bar{x}^{(1)} + \bar{x}^{(2)} \right)' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)} \right) \right\}$$

donde $\hat{\mu}^{(1)} = \bar{x}^{(1)}$, $\hat{\mu}^{(2)} = \bar{x}^{(2)}$ y $S = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_j^{(i)} - \bar{x}^{(i)})(x_j^{(i)} - \bar{x}^{(i)})'$.

Es necesario mencionar que estas regiones estimadas no son óptimas en el sentido de que minimicen el riesgo esperado, ya que están estimando los parámetros de las densidades, pero se espera que se parezcan mucho a las regiones óptimas cuando n_1 y n_2 son grandes.

Sea

$$V = X' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)} \right) - \frac{1}{2} \left(\bar{x}^{(1)} + \bar{x}^{(2)} \right)' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)} \right)$$

nótese que V es igual a U definida en (1.12) al reemplazar los parámetros por sus estimadores. El primer término de V es la función discriminante de Fisher basada en dos muestras. El problema ahora es encontrar la distribución de V y sus características, para poder medir su eficiencia en la discriminación.

Se ha demostrado (Anderson, 1958), que si $n_1 \to \infty$ y $n_2 \to \infty$, la distribución límite de V es la misma que la de U,definida en (1.12), lo que implica que para muestras grandes usaremos V como si conociéramos las densidades completamente.

En el caso de p poblaciones normales, se sustituyen estimadores de los parámetros de $u_{jk}(X)$ definidos en (1.13), obteniéndose:

$$v_{ij}(X) = \left[X - \frac{1}{2}\left(\bar{x}^{(i)} + \bar{x}^{(j)}\right)\right]' S^{-1}\left(\bar{x}^{(i)} - \bar{x}^{(j)}\right)$$

Si $n_i \to \infty$ la distribución de v_{ij} se aproxima a la de u_{ij} por lo que se usa v_{ij} de la misma manera que u_{ij}.

b) Método de razón de verosimilitud.

Suponga que se tienen una muestra de tamaño n_1 de la población Π_1 con distribución $N\left(\underline{\mu}^{(1)}, \Sigma\right)$ y una muestra de tamaño n_2 de la población Π_2 que se distribuye como $N\left(\underline{\mu}^{(2)}, \Sigma\right)$, y se tiene una nueva observación x que se quiere asignar a alguna de las dos poblaciones.

Se tienen las siguientes hipótesis:

$$H_0 : x, x_1^{(1)}, \ldots, x_{n_1}^{(1)} \sim N\left(\underline{\mu}^{(1)}, \Sigma\right)$$
$$x_1^{(2)}, \ldots, x_{n_2}^{(2)} \sim N\left(\underline{\mu}^{(2)}, \Sigma\right)$$

vs.

$$H_1 : \; x_1^{(1)}, \ldots, x_{n_1}^{(1)} \sim N\left(\underline{\mu}^{(1)}, \Sigma\right)$$
$$x, x_1^{(2)}, \ldots, x_{n_2}^{(2)} \sim N\left(\underline{\mu}^{(2)}, \Sigma\right)$$

con $\underline{\mu}^{(1)}, \underline{\mu}^{(2)}, \Sigma$ desconocidos.

Bajo H_0, los estimadores máximo verosímiles de $\underline{\mu}^{(1)}$, $\underline{\mu}^{(2)}$ y Σ son

$$\hat{\underline{\mu}}_0^{(1)} = \frac{\left(n_1 \overline{x}^{(1)} + x\right)}{(n_1 + 1)}$$

$$\hat{\underline{\mu}}_0^{(2)} = \overline{x}^{(2)}$$

$$\hat{\Sigma}_0 = \frac{1}{n_1 + n_2 + 1}\left[\sum_{\alpha=1}^{n_1}\left(x_\alpha^{(1)} - \hat{\underline{\mu}}_0^{(1)}\right)\left(x_\alpha^{(1)} - \hat{\underline{\mu}}_0^{(1)}\right)' + \left(x - \hat{\underline{\mu}}_0^{(1)}\right)\left(x - \hat{\underline{\mu}}_0^{(1)}\right)' + \sum_{\alpha=1}^{n_2}\left(x_\alpha^{(2)} - \hat{\underline{\mu}}_0^{(2)}\right)\left(x_\alpha^{(2)} - \hat{\underline{\mu}}_0^{(2)}\right)'\right]$$

$$= \frac{1}{n_1 + n_2 + 1}\left[C + \frac{n_1}{n_2 + 1}\left(x - \overline{x}^{(1)}\right)\left(x - \overline{x}^{(1)}\right)'\right]$$

donde $C = \sum_{i=1}^{2}\sum_{\alpha=1}^{n_i}\left(x_\alpha^{(i)} - \overline{x}^{(i)}\right)\left(x_\alpha^{(i)} - \overline{x}^{(i)}\right)'$.

Bajo H_1, los estimadores máximo verosímiles son:

$$\underline{\hat{\mu}}_1^{(1)} = \overline{x}^{(1)}$$

$$\underline{\hat{\mu}}_1^{(2)} = \frac{\left(n_2\overline{x}^{(2)} + x\right)}{(n_2+1)}$$

$$\hat{\Sigma}_1 = \frac{1}{n_1+n_2+1}\left[C + \frac{n_2}{n_2+1}\left(x-\overline{x}^{(2)}\right)\left(x-\overline{x}^{(2)}\right)'\right]$$

La razón de verosimilitud es la potencia $\frac{(n_1+n_2+1)}{2}$ de

$$\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_0|} = \frac{|C + \frac{n_2}{n_2+1}\left(x-\overline{x}^{(2)}\right)\left(x-\overline{x}^{(2)}\right)'|}{|C + \frac{n_1}{n_1+1}\left(x-\overline{x}^{(1)}\right)\left(x-\overline{x}^{(1)}\right)'|}$$

que se puede escribir como

$$\frac{1 + \frac{n_2}{n_2+1}\left(x-\overline{x}^{(2)}\right)' C^{-1}\left(x-\overline{x}^{(2)}\right)}{1 + \frac{n_1}{n_1+1}\left(x-\overline{x}^{(1)}\right)' C^{-1}\left(x-\overline{x}^{(1)}\right)}$$

La región de clasificación en Π_1 consiste de los puntos para los cuales la razón de verosimilitudes es mayor que un número dado, dependiendo del nivel de significancia de la prueba.

1.3.2 Enfoque Predictivo

Suponga que para la población Π_i, se dispone de una muestra aleatoria $x_1^{(i)}, \ldots, x_n^{(i)}$. Los parámetros, θ_i, de la población Π_i son desconocidos.

Dada la función a priori $r(\theta_i)$ para los parámetros desconocidos, se actualiza ésta tomando en cuenta la muestra observada en cada población, obteniéndose la densidad a posteriori de la siguiente forma:

$$r(\theta_i|x_1^{(i)},\ldots,x_n^{(i)}) = \frac{f(x_1^{(i)},\ldots,x_n^{(i)}|\theta_i)r(\theta_i)}{\int f(x_1^{(i)},\ldots,x_n^{(i)}|\theta_i)r(\theta_i)d\theta_i}$$

y de aquí se calcula la densidad predictiva para X como sigue:

$$f(X|x_1^{(i)},\ldots,x_n^{(i)},\Pi_i) = \int f(X|\theta_i)r(\theta_i|x_1^{(i)},\ldots,x_n^{(i)})d\theta_i.$$

La densidad $f(X|x_1^{(i)},\ldots,x_n^{(i)},\Pi_i)$ es la densidad predictiva para X de la población Π_i y será usada para la clasificación.

Geisser (1964) trata el problema para el caso de k poblaciones normales univariadas Π_i, $N(\theta_i,\sigma_i^2)$, $i = 1,\ldots,k$, suponiendo que se dispone de los estimadores usuales de θ_i y σ_i^2:

$$\hat{\theta}_i = \bar{x}^{(i)} = \frac{1}{n_i}\sum_{\alpha}^{n_i} x_\alpha^{(i)}$$

$$\hat{\sigma}_i^2 = s_i^2 = \frac{1}{n_i - 1}\sum_{\alpha}^{n_i}\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)^2$$

basados en muestras de tamaño n_i, $i = 1,\ldots,k$.

Suponiendo una a priori no informativa para los parámetros,

$$g(\theta_i,\sigma_i)d\theta_i d\sigma_i \propto \frac{d\theta_i d\sigma_i}{\sigma_i}$$

calcula la densidad posterior de θ_i y σ_i

$$r(\theta_i, \sigma_i | \bar{x}_i, s_i) \propto f(\bar{x}_i, s_i | \theta_i, \sigma_i) g(\theta_i, \sigma_i)$$

y de aquí calcula la densidad predictiva para una observación z, condicional a la población Π_i,

$$f(z|\bar{x}^{(i)}, s_i, \Pi_i) = \int\int f(z|\theta_i, \sigma_i) r(\theta_i, \sigma_i | \bar{x}^{(i)}, s_i) d\theta_i d\sigma_i$$

$$= \left(\frac{n_i}{(n_i^2-1)\pi}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}n_i)}{\Gamma\{\frac{1}{2}(n_i-1)\}s_i} \left(1 + \frac{n_i\left(\bar{x}^{(i)} - z\right)^2}{(n_i^2-1)s_i^2}\right)^{-\frac{1}{2}n_i}.$$

Una vez teniendo la densidad predictiva, se sustituye en (1.4) en lugar de $f(X|\Pi_i)$ y se sigue el procedimiento de asignación ya visto.

Geisser, en el mismo artículo, trata también el caso multivariado, donde supone una observación $z' = (z_1, \ldots, z_p)$ con probabilidad a priori q_i de pertenencia a Π_i, $N\left(\underline{\mu}^{(i)}, \Sigma_i\right)$, y se tienen, como antes, estimadores $\bar{x}_i' = (\bar{x}_{1i}, \ldots, \bar{x}_{pi})$ y s_i de los parámetros. Discute nueve casos diferentes al variar suposiciones acerca de Σ y $\underline{\mu}$.

Press (1972) trata el caso de k poblaciones normales p-variadas

$$\Pi_j = N(\theta_j, \Sigma_j) \quad j = 1, \ldots, k,$$

con parámetros desconocidos y distribución a priori de los parámetros no informativa; encuentra que la distribución predictiva para clasificar z en Π_j está dada por la

densidad t-Student multivariada:

$$f(z|x_1^{(j)},\ldots,x_n^{(j)},\Pi_j) = \frac{k_j}{\left[1+\frac{n_j}{n_j^2-1}\left(z-\overline{x}^{(j)}\right)' S_j^{-1}\left(z-\overline{x}^{(j)}\right)\right]^{\frac{n_j}{2}}}$$

donde k_j es una constante que no depende de z, dada por

$$k_j = \left[\frac{n_j}{(n_j+1)\pi}\right]^{p/2} \frac{\Gamma\left(\frac{n_j}{2}\right) q_j}{\Gamma\left(\frac{n_j-p}{2}\right) |(n_j-1)s_j|^{1/2}}$$

y q_j es la probabilidad a priori de clasificar z en Π_j.

Aitchison et. al. (1977) compararon el método estimativo con el predictivo, recomendando el uso del método predictivo. Aunque, como en todo problema Bayesiano, existe una cierta dificultad en decidir qué tipo de distribución a priori utilizar y aún en el caso en que se decida utilizar una no informativa, a veces es difícil de obtener.

1.4 Selección de Variables

En cualquier problema de clasificación existe un número ilimitado de mediciones que se pueden tomar a los objetos por clasificar; sin embargo, no es conveniente trabajar con un número grande de variables (mediciones) por las siguientes razones:

1. Costo. Si es muy caro (monetariamente ó en tiempo) el realizar las mediciones, es mejor tener pocas. Este es un punto muy importante en cualquier análisis práctico.

2. Simplicidad. Si se considera un número pequeñó de variables, es mucho más fácil la interpretación de los resultados y aún más, el análisis resulta más sencillo.

3. Eliminar redundancia.

4. Tasa de mala clasificación. Varios autores han estudiado el fenómeno de que para un conjunto de datos dado, al aumentar el número de mediciones la tasa de mala clasificación se reduce al principio pero después empieza a aumentar.

   Una explicación informal es la siguiente (Hand(1981)): conforme el número de variables, m, aumenta, el número de parámetros que definen la superficie de decisión aumenta y se clasifica al conjunto de datos con menor tasa de

mala clasificación. Sin embargo, conforme m aumenta, cada vez menos de la probabilidad asociada con cada población cae en regiones de baja densidad; esto significa que el conjunto de datos es más y más poco densamente distribuido, y los elementos del conjunto de datos se vuelven cada vez menos representativos de la forma de la distribución de la población. La consecuencia es que la superficie de decisión se puede ajustar mejor al conjunto de datos dado, cuando se incrementa m, pero esta misma superficie de decisión se ajusta menos bien a nuevas muestras, es decir, la verdadera tasa de error aumenta.

Por estas razones se han desarrollado métodos para seleccionar las m' "mejores variables" del conjunto original de m. A continuación se mencionarán algunos de ellos.

1. Métodos de búsqueda exhaustiva. Se consideran cada uno de los $\binom{m}{m'}$ conjuntos posibles; y se evalúan por algún método y se escoge el mejor. Estos procedimientos se pueden aplicar solamente cuando m es pequeño.

2. Métodos de búsqueda acelerada. Se consideran todos los conjuntos de variables, pero no se evalúa a todos explícitamente.

3. Métodos subóptimos por pasos (stepwise). Permiten una búsqueda rápida pero con el inconveniente de que no se garantiza el que se encuentre la mejor

solución. Son los más utilizados en paquetes estadísticos y por esto, se explicarán un poco más.

a. Selección secuencial hacia adelante (forward). Se selecciona la variable que en cada paso discrimine mejor, es decir, la que produce el mayor valor del criterio, en presencia de las ya seleccionadas anteriormente. Este método tiene dos desventajas: no toma en cuenta las relaciones entre las variables que aún no entran y, segundo, una vez que una variable se selecciona no hay forma de eliminarla, ya que adiciones subsecuentes pueden hacerla innecesaria.

b. Eliminación secuencial hacia atrás (backward). El proceso empieza con todas las variables y se elimina la variable que discrimine menos, es decir, la que contribuya menos al criterio. Este paso se repite sucesivamente. Tiene las desventajas de que las relaciones entre las variables eliminadas no son tomadas en cuenta y una vez que una variable es eliminada no puede ser reincorporada.

c. Selección por pasos (Stepwise). Es una combinación de los dos métodos anteriores, es decir, se selecciona una variable para entrar, con cierto criterio, pero además se analizan todas las variables que ya entraron para ver la posibilidad de eliminar alguna de ellas.

Existen varios criterios que se pueden utilizar para los métodos descritos arriba, algunos de ellos son medidas de distancia entre los grupos, como son:

- La Λ de Wilks.
- La D^2 de Mahalanobis.
- La V de Rao.

En cuanto a una regla de paro, Rao dá una estadística para probar la significancia de la contribución a la discriminación de m' variables, para el caso de dos poblaciones:

$$F = \frac{(n - m' - 1)n_1 n_2 (D_m^2 - D_{m'}^2)}{(m - m')\left[n(n-2) + n_1 n_2 D_{m'}^2\right]}$$

donde n es el tamaño de muestra total, D_m^2 es la D^2 de Mahalanobis evaluada en m variables y $D_{m'}^2$ es lo mismo evaluada en el subconjunto de m' variables. Esta estadística tiene una distribución F con $(m - m')$ y $(n - m - 1)$ grados de libertad.

En conclusión, al utilizar alguno de los procedimientos de selección descritos, se encuentran las "mejores" variables a medir. "Mejores" dependiendo del criterio usado. Se debe notar que estos procedimientos no trabajan en ningun momento con los costos de medición asociados a las variables, por lo que el "mejor" subconjunto de variables encontrado para la discriminación, no necesariamente es el "mejor" al considerar costos.

En el siguiente capítulo se trata el problema de discriminación con un enfoque

secuencial para tratar de incorporar el costo de medición en la selección de las variables a medir.

CAPITULO

2

Discriminación Secuencial

2.1 Introducción

Un experimento secuencial puede ser definido como aquel en el que el curso del experimento depende de alguna forma de los resultados obtenidos hasta ese momento.

Se deben considerar los procedimientos secuenciales cuando las mediciones, por su naturaleza, se toman en forma secuencial, es decir, una a la vez, y sobre todo cuando el costo de la observación es alto, ya sea monetariamente, en tiempo, en complejidad o en riesgo.

En un procedimiento secuencial se hace un balance entre la información aportada por la medición y el costo de tomarla. Es importante mencionar que el costo podría ser una medida de preferencia para incluir alguna característica en el estudio y no necesariamente es una calificación negativa para la característica.

Un equilibrio entre el error (de clasificación equivocada) y el número de características a medir, se puede obtener utilizando un procedimiento secuencial y terminándolo cuando se alcance una precisión de clasificación deseada.

Si las mediciones se van a tomar secuencialmente, es importante el orden en que se van a medir las características. Se esperaría que las características tuvieran un orden tal, que las mediciones tomadas en ese orden llevaran a una pronta decisión.

Fu (1968) trata algunos métodos para seleccionar y ordenar las características a medir, evaluando la "bondad" de las características con criterios como el de divergencia, que es una medida de distancia entre dos poblaciones.

En el caso de asignación a p poblaciones, considerado en este trabajo, un procedimiento de decisión secuencial está formado por un espacio de estados de la naturaleza $\Theta = \{\Pi_1, \Pi_2, \ldots, \Pi_p\}$ que representa las poblaciones consideradas, un espacio de decisiones $D = \{d_1, d_2, \ldots, d_p\}$ y una función de pérdida $L : D \times \Omega \rightarrow \Re$.

Además se observa secuencialmente un vector aleatorio $X = (X_1, \ldots, X_m)$, cuya densidad condicional a la población Π_i es $f(X|\Pi_i)$.

Un procedimiento de decisión secuencial consta de dos partes:

a) Plan de muestreo. Donde se especifica si se va a tomar una decisión en D sin observaciones, o si no es el caso, se especifica para cualquier posible conjunto de valores $X_1 = x_1, \ldots, X_n = x_n$ $(n \geq 1)$, si el muestreo se para y se toma una decisión en D sin más observaciones, ó si se observa el valor de X_{n+1}.

b) Regla de decisión. Es la que especifica la decisión $d_0 \in D$ que se va a escoger si no se toman observaciones, y, si por lo menos se toma una observación, especifica la función de decisión $d_n(x_1, \ldots, x_n) \in D$ $(n \geq 1)$ que se va a escoger para cada posible conjunto de valores observados $X_1 = x_1, \ldots, X_n = x_n$ después de los cuales el muestreo debió haber terminado.

A continuación, se tratarán diferentes procedimientos secuenciales, listando sus ventajas y desventajas, hasta llegar a los procedimientos óptimos para solucionar el problema de asignación.

2.2 Solución al Problema vía Procedimientos Secuenciales

2.2.1 Prueba de Wald

En el caso de tener dos poblaciones, se puede utilizar la prueba secuencial de razón de verosimilitudes (Sequential Probability Ratio Test, SPRT por sus siglas en inglés) de Wald, que es análoga a la prueba de hipótesis simple vs. simple de Neyman-Pearson para un procedimiento no secuencial.

La prueba de Wald es como sigue:

Se tienen las hipótesis:

H_1 : el individuo pertenece a la población Π_1.

H_2 : el individuo pertenece a la población Π_2.

Se fijan tanto la probabilidad del error tipo I ($P[\text{rechazar } H_1|H_1] = \alpha$) como la probabilidad del error tipo II ($P[\text{no rechazar } H_1|H_2] = \beta$).

En la etapa n-ésima, es decir, después de haber tomado la n-ésima medición,

se calcula la razón de verosimilitudes λ_n,

$$\lambda_n = \frac{f(x_1, \ldots, x_n|H_1)}{f(x_1, \ldots, x_n|H_2)}.$$

La decisión a tomar es una de las tres siguientes:

Continúe tomando observaciones mientras $B < \lambda_n < A$.

Pare y acepte H_1 si $\lambda_n \geq A$.

Pare y acepte H_2 si $\lambda_n \leq B$.

Donde $A \doteq \frac{1-\beta}{\alpha}$ $\quad B \doteq \frac{\beta}{1-\alpha}$.

VENTAJAS:

- Wald, Wolfowitz (1948) demostraron que para probabilidades de error α y β fijos, requiere en promedio un menor número de observaciones que cualquier otra prueba con las mismas α y β.

- No necesita suponer independencia en las observaciones.

- Con probabilidad 1 llega a una decisión.

DESVENTAJAS:

- El orden en que se van a tomar las observaciones está especificado de antemano.

- No toma en cuenta el costo de observación de las características.

- El número promedio de observaciones puede ser extremadamente grande si se escogen α y β muy pequeños. Este es un inconveniente grave del método ya que generalmente se tiene un número máximo de características a medir.

- No considera las probabilidades a priori.

Wald sugirió una corrección a su procedimiento para truncar el proceso secuencial en el paso m. Si no se ha tomado una decisión en el paso m, acepte H_1 si $\lambda_m > 1$ ó acepte H_2 si $\lambda_m < 1$.

2.2.2 Modificación a la Prueba de Wald

Fu hace una crítica al truncamiento abrupto del procedimiento en el paso m, diciendo que es ineficiente porque si el valor de la razón secuencial de verosimilitudes es grande y el número de mediciones está cercano a m, un número pequeño de mediciones adicionales no permitirán, en general, gran oportunidad de rechazar cualquier hipótesis no importando por ésto el que las mediciones se efectúen o no.

Fu estudia una modificación a la prueba de Wald que, en términos generales, consiste en variar los límites de paro en función de la etapa del proceso (n).

Sea $g_1(n)$ y $g_2(n)$ funciones de n, monótonas no creciente y no decreciente respectivamente. El procedimiento modificado es:

Continúe tomando observaciones mientras

$$e^{g_2(n)} < \lambda_n < e^{g_1(n)} \qquad n = 1, 2, \ldots$$

Si $\lambda_n \geq e^{g_1(n)}$ acepte H_1.

Si $\lambda_n \leq e^{g_2(n)}$ acepte H_2.

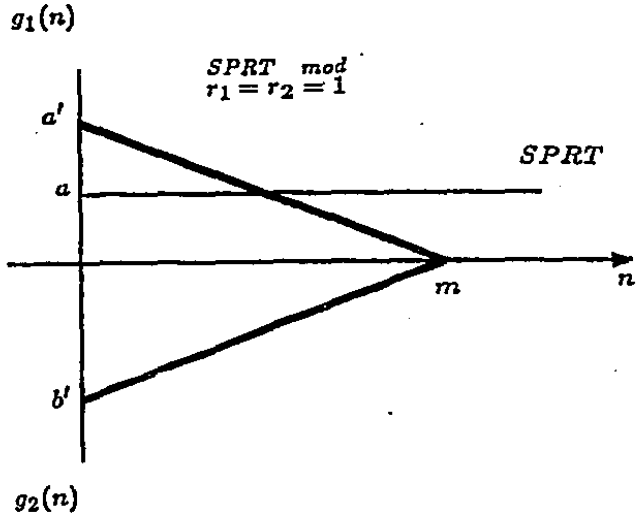
Fu considera las funciones:

$$g_1(n) = a'(1 - \frac{n}{m})^{r_1}$$

$$g_2(n) = -b'(1 - \frac{n}{m})^{r_2}$$

con $0 < r_1, r_2 \leq 1, \quad a' > 0, \quad b' > 0.$

En la siguiente gráfica se presentan las regiones de "continuación" para la prueba de Wald y para la prueba modificada.



Se nota que en la prueba modificada la región de continuación se hace más pequeña conforme n aumenta, hasta que en $n = m$ ya no existe región de "continuación", forzando una decisión.

Las relaciones de la prueba modificada con la de Wald son las siguientes:

- Fu demuestra que la prueba modificada requiere de un número esperado menor de mediciones que la prueba de Wald.

- La potencia de la prueba modificada es menor que la de Wald.

Sin embargo, al construir apropiadamente los límites de paro que varían con el

tiempo, se puede lograr lo siguiente:

- El proceso de clasificación siempre termina en a lo más un número m preespecificado de mediciones.

- El número esperado de mediciones es controlable y generalmente es menor que el requerido por la prueba de Wald.

- Es posible, ajustando los puntos iniciales de los límites de paro, alcanzar probabilidades de error tan pequeñas como las de la prueba de Wald.

2.2.3 Prueba de Wald Generalizada

La prueba SPRT de Wald se ha generalizado a más de dos poblaciones (Generalized Sequential Probability Ratio Test, GSPRT). Fu (1968) comenta el siguiente procedimiento: Suponga que hay p hipótesis, donde la hipótesis H_i es: el individuo pertenece a la población Π_i, $i = 1, \ldots, p$.

En la etapa n-ésima, después de haber observado $X = (X_1, \ldots, X_n)$, la razón secuencial generalizada de verosimilitudes para cada hipótesis se define como:

$$U_n(X|H_i) = \frac{f(X|H_i)}{\left[\prod_{k=1}^{p} f(X|H_k)\right]^{1/p}} \qquad i = 1, \ldots, p.$$

La regla de decisión es: Compare $U_n(X|H_i)$ con el límite de paro $A(H_i)$ de la hipótesis H_i, y elimine las H_i que cumplan con

$$U_n(X|H_i) < A(H_i) \qquad i = 1, \ldots, p,$$

donde

$$A(H_i) = \frac{1 - e_{ii}}{[\prod_{k=1}^{p}(1 - e_{ik})]^{1/p}} \qquad i = 1, \ldots, p,$$

y e_{ik} es la probabilidad de aceptar H_i dado que H_k es cierta.

Después de rechazar la hipótesis H_i, el número total de hipótesis se reduce y se calcula un nuevo conjunto de razones secuenciales generalizadas de verosimilitudes. Las hipótesis se rechazan secuencialmente hasta que queda una, la cual se acepta. Para $p = 2$, esta prueba GSPRT es equivalente a la SPRT.

Ghosh (1970) trata otro procedimiento GSPRT llamado la Solución de Armitage, que es como sigue:

Suponga que hay p hipótesis. Sea $I_{jl}(n)$ las desigualdades:

$$-a_{jl} < Z_n(j,l) = ln\left[\frac{f(x_1, \ldots, x_n; \theta_l)}{f(x_1, \ldots, x_n; \theta_j)}\right] < a_l j$$

para $j \neq l = 0, 1, \ldots, p-1, \quad n \geq 1$.

Donde a_{jl} son $k(k-1)$ constantes positivas que dependen de los niveles de significancia fijados. $I_{jl}(n)$ representa la región de continuación de la prueba SPRT para probar $H_j : \theta = \theta_j$ vs. $H_l : \theta = \theta_l$.

El procedimiento es el siguiente:

Observe x_i sucesivamente, en la etapa n, $n \geq 1$:

i) Acepte H_j para cualquier $j \neq l$, si se violan todas las $k-1$ desigualdades inferiores en $I_{j0}, \ldots, I_{jj-1}, I_{jj+1}, \ldots, I_{jk-1}$.

ii) Acepte H_l si se violan todas las $k-1$ desigualdades superiores en $I_{0l}, \ldots, I_{l-1l}, I_{l+1l}, \ldots, I_{k-1l}$.

iii) En otro caso, continúe observando x_{n+1}.

Este procedimiento para $p = 2$ es equivalente a la prueba SPRT.

2.3 Solución Optima

Como se vió en el capítulo 1 sección 1.2.1, en un problema de decisión con p poblaciones y una formulación no secuencial, en el sentido de incluir las m variables disponibles, la decisión d^* Bayes óptima es aquella que minimiza el riesgo de Bayes $B(d)$, definido como:

$$\begin{aligned} B(d) &= \int_{\Omega_X} \sum_{i=1}^{p} L(d, \Pi_i) f(X|\Pi_i) q_i dX \\ &= \sum_{i=1}^{p} q_i R_{\Pi_i}(d) \end{aligned}$$

donde

$$R_{\Pi_i}(d) = \int_{\Omega_X} L(d, \Pi_i) f(X|\Pi_i) dx$$

es el riesgo condicional a la población Π_i.

d^* es decisión Bayes si

$$B(d^*) = \min_i B(d_i).$$

$L(d_j, \Pi_i)$ es la pérdida incurrida al asignar el individuo a la población Π_j cuando en realidad pertenece a Π_i.

En el caso de una función de pérdida de la forma

$$L(d_j, \Pi_i) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

La decisión Bayes es $d^* = d_i$ si

$$q_i f(X|\Pi_i) \geq q_j f(X|\Pi_j) \qquad \forall j = 1, \ldots, p.$$

Ahora considerando el procedimiento secuencial; tomando las observaciones una a la vez, cada etapa es un problema de decisión entre terminar el muestreo con la selección de la decisión terminal (asignar) ó continuar tomando una observación más.

El procedimiento óptimo se encuentra aplicando el principio de optimalidad de Bellman(1957), que dice:

"Una estrategia óptima tiene la propiedad de que cualesquiera que hayan sido el estado y la decisión iniciales, las decisiones posteriores deben constituir una estrategia óptima respecto al estado resultante de la primera decisión". Esto significa que en cada etapa del proceso secuencial, las decisiones siguientes a esa etapa deben ser una estrategia óptima. Por esto, para determinar la mejor decisión en una etapa (continuar el proceso o no), es necesario conocer la mejor decisión en el futuro.

Por lo tanto, para encontrar el procedimiento óptimo es necesario trabajar para atrás (backward) en el tiempo, considerando todas las posibilidades del futuro, para encontrar la mejor decisión en el presente. Esto se analiza utilizando programación dinámica, considerando la última etapa como el punto de partida y trabajando hacia atrás.

Lo que resta de este capítulo son aportaciones que ha hecho Fu(1968).

Sean:

m el número máximo de variables a observar,

$C_j(x)$ el costo de las observaciones $x_1, \ldots, x_j$,

$d_j(x)$ la función de decisión basada en las observaciones $x_1, \ldots, x_j$,

S el plan de muestreo, $S = \{S_1, S_2, \ldots, S_m\}$, $\cup_{j=1}^{m} S_j = \Omega_x$. Donde cada S_j es el conjunto de paro en la etapa j, es decir, el muestreo se termina y se toma la decisión $d_j(x)$.

Y sean las q_i, $i = 1, \ldots, p$ las probabilidades a priori de pertenencia.

El riesgo promedio asociado a todo el procedimiento secuencial es:

$$\begin{aligned} R(q,S,d) &= \sum_{i=1}^{p} q_i \sum_{j=1}^{m} \int_{S_j} \left[C_j(x) + L(d_j(x), \Pi_i)\right] f(x|\Pi_i)dx \\ &= \sum_{j=1}^{m} \int_{S_j} \sum_{i=1}^{p} \left[C_j(x) + L(d_j(x), \Pi_i)\right] q_i f(x|\Pi_i)dx \\ &= \sum_{j=1}^{m} \int_{S_j} \sum_{i=1}^{p} C_j(x) q_i f(x|\Pi_i)dx + \sum_{j=1}^{m} \int_{S_j} \sum_{i=1}^{p} L(d_j(x), \Pi_i) q_i f(x|\Pi_i)dx \\ &= \sum_{j=1}^{m} \int_{S_j} C_j(x) f(x)dx + \sum_{j=1}^{m} \int_{S_j} \sum_{i=1}^{p} L(d_j(x), \Pi_i) q_{i|x} dF(x). \end{aligned}$$

El problema es encontrar el procedimiento para minimizar este riesgo y las

dificultades están en encontrar los conjuntos de paro S_j y las decisiones $d_j(x)$ que lleven a un proceso óptimo. El procedimiento que minimiza este riesgo se encuentra utilizando programación dinámica. Para ver esto con mayor detalle se puede consultar a DeGroot (1970).

2.3.1 Procedimiento Optimo considerando un orden predeterminado en las observaciones

Sea $\rho_n(x_1, x_2, \ldots, x_n)$ el mínimo riesgo promedio de todo el procedimiento secuencial, habiendo observado $x_1, x_2, \ldots, x_n$.

Sea $C(x_1, x_2, \ldots, x_n)$ el costo de tomar una observación más en la n-ésima etapa.

Sea $R(x_1, x_2, \ldots, x_n; d_i)$ el riesgo promedio de tomar la decisión d_i, es decir, asignar a la población Π_i habiendo observado $x_1, \ldots, x_n$. Si el proceso termina en esta etapa, el riesgo promedio es $\min_i R(x_1, x_2, \ldots, x_n; d_i)$ tomando la decisión Bayes, o sea, asignando a la población Π que corresponda al $\min_i R(x_1, x_2, \ldots, x_n; d_i)$.

Si el proceso continúa a la etapa $n+1$, el riesgo esperado es

$$C(x_1, x_2, \ldots, x_n) + \int \rho_{n+1}(x_1, x_2, \ldots, x_n, x_{n+1}) dF(x_{n+1}|x_1, \ldots, x_n)$$

donde la integral es sobre la región de valores posibles de x_{n+1}.

En la última etapa, habiendo observado $x_1, x_2, \ldots, x_m$, el riesgo de tomar una decisión terminal resulta ser igual al riesgo de tomar una observación más, es decir,

$$\rho_m(x_1, x_2, \ldots, x_m) = \min_i R(x_1, x_2, \ldots, x_m; d_i)$$

De aquí trabajando hacia atrás, en la etapa $m-1$, habiendo observado $x_1, x_2, \ldots, x_{m-1}$, el mínimo riesgo esperado es

$$\rho_{m-1}(x_1, \ldots, x_{m-1}) = min \begin{cases} \text{cont:} & C(x_1, x_2, \ldots, x_{m-1}) + \\ & \int \rho_m(x_1, \ldots, x_m) dF(x_m | x_1, \ldots, x_{m-1}) \\ \text{parar:} & \min_i R(x_1, \ldots, x_{m-1}; d_i) \end{cases}$$

En la etapa $m-2$,

$$\rho_{m-2}(x_1, \ldots, x_{m-2}) = min \begin{cases} \text{cont:} & C(x_1, x_2, \ldots, x_{m-2}) + \\ & \int \rho_{m-1}(x_1, \ldots, x_{m-1}) dF(x_{m-1} | x_1, \ldots, x_{m-2}) \\ \text{parar:} & \min_i R(x_1, \ldots, x_{m-2}; d_i) \end{cases}$$

Así sucesivamente, hasta que en la primera etapa:

$$\rho_1(x_1) = min \begin{cases} \text{cont:} & C(x_1) + \int \rho_2(x_1, x_2) dF(x_2 | x_1) \\ \text{parar:} & \min_i R(x_1; d_i) \end{cases}$$

La ecuación funcional, en la etapa n, $n = 1, 2, \ldots m-1$ es la siguiente:

$$\rho_n(x_1, \ldots, x_n) = min \begin{cases} \text{cont:} & C(x_1, x_2, \ldots, x_n) + \\ & \int \rho_{n+1}(x_1, \ldots, x_{n+1}) dF(x_{n+1} | x_1, \ldots, x_n) \\ \text{parar:} & \min_i R(x_1, \ldots, x_n; d_i) \end{cases}$$

Lo que se está haciendo en cada paso es comparar los riesgos de continuación y de paro. El riesgo de continuar considera el costo de tomar la observación x_{n+1}

aumentado a la esperanza del riesgo a partir de la etapa $n+1$ hasta la última etapa, con respecto a los posibles valores que pueda tomar x_{n+1}.

VENTAJAS:

- Es un procedimiento óptimo en el sentido de que minimiza el riesgo esperado.

DESVENTAJAS:

- En el caso de variables aleatorias continuas, al parecer, es computacionalmente imposible de implantar; ya que el cálculo del riesgo de continuación en cada etapa implica conocer todos los posibles valores que se pueden tener en el futuro. Aún con variables discretas, el requerimiento de espacio de memoria para un algoritmo computacional, podría ser enorme.

Por ejemplo, suponga el caso de tres variables disponibles cuyos valores posibles son $0, 1$.

Los posibles valores de las tres variables son:

$$(0,0,0),\quad (0,0,1),\quad (0,1,0),\quad (0,1,1)$$

$$(1,0,0),\quad (1,0,1),\quad (1,1,0),\quad (1,1,1)$$

En la última etapa (la tercera), habría que calcular 8 funciones de riesgo (2^3).

En la etapa 2, se calculan 4 funciones de riesgo. En la etapa 1, se calculan 2 funciones de riesgo. Así, en total, se calculan $2^3 + 2^2 + 2 = 14$ funciones de riesgo.

En el caso de tener m variables que toman k valores cada una, el número de funciones de riesgo a calcular son $k + k^2 + k^3 + \cdots + k^m$. Con esto se hace notar que en el caso discreto el número de cálculos puede llegar a ser enorme.

2.3.2 Procedimiento Optimo sin un orden predeterminado de las observaciones

Este es un procedimiento más general que el anterior, ya que tiene la capacidad adicional de seleccionar la mejor característica para la siguiente medición.

Sean:

$F_m = \{f_1, \ldots, f_m\}$ el conjunto de m características que se pueden medir, y $F_{t_n} = \{f_{t_1}, \ldots, f_{t_n}\}$ $n = 1, \ldots, m$, una secuencia particular de n características ya medidas en la etapa n del proceso secuencial. El conjunto de características disponibles para mediciones posteriores a la n-ésima etapa es $F_n = F_m - F_{t_n}$.

La característica f_{t_i}, es cualquiera de los elementos de F_m, cuya medición se representa por la variable aleatoria x_i.

En forma similar al procedimiento anterior, sean:

$\rho_n(x_1,\ldots,x_n|F_{t_n})$ el mínimo riesgo esperado de todo el proceso secuencial, habiendo observado las variables $x_1,\ldots,x_n$ correspondientes a la secuencia de características F_{t_n} seleccionadas,

$C(x_1,\ldots,x_n|F_{t_n})$ el costo de tomar la siguiente observación en la etapa n-ésima cuando se seleccionó F_{t_n},

$R(x_1,\ldots,x_n;d_i|F_{t_n})$ el riesgo esperado de tomar la decisión terminal d_i con base en las mediciones $x_1,\ldots,x_n$ de la secuencia de características F_{t_n}, y

$F(x_{n+1};f_{t_{n+1}}|x_1,\ldots,x_n;F_{t_n})$ la distribución condicional de x_{n+1} cuando se selecciona la característica $f_{t_{n+1}}$ dado $x_1,\ldots,x_n$ correspondientes a F_{t_n}.

La ecuación funcional es:

$$\rho_n(x_1,\ldots,x_n|F_{t_n}) = min\begin{cases}\text{cont:} & C(x_1,x_2,\ldots,x_n|F_{t_n})+ \\ & \int \rho_{n+1}(x_1,\ldots,x_{n+1}|F_{t_n},f_{t_{n+1}}) \\ & \quad dF(x_{n+1},f_{t_{n+1}}|x_1,\ldots,x_n;F_{t_n}) \\ \text{parar:} & \min_i R(x_1,\ldots,x_n;d_i|F_{t_n})\end{cases}$$

Con la condición en la última etapa:

$$\rho_m(x_1,\ldots,x_m;F_{t_n}) = \min_i R(x_1,\ldots,x_m;d_i|F_{t_n}).$$

El procedimiento es el mismo que en la sección 2.3.1, con la diferencia que en éste se selecciona automáticamente la mejor secuencia de características a medir,

pudiendo llevar a una decisión terminal en forma más rápida. Tiene los mismos problemas en cuanto a factibilidad computacional que el procedimiento en 2.3.1.

Para salvar las inconveniencias computacionales del método óptimo, Fu esencialmente supone independencia de las observaciones, es decir, $F(x_{n+1}|x_1,\ldots,x_n) = F(x_{n+1})$, o bien supone dependencia markoviana de primer orden, es decir, $F(x_{n+1}|x_1,\ldots,x_n) = F(x_{n+1}|x_n)$ y trabaja categorizando las variables.

Fu estudia un ejemplo de reconocimiento de los caracteres D, J, P y V (4 clases) con 8 variables, suponiendo independencia, discretizándolas en 5 intervalos y suponiendo una distribución multinomial de las poblaciones. Realiza los siguientes experimentos:

Experimento 1.

- Costo igual para las 8 variables.
- Función de pérdida $L(\Pi_i, d_j) = A \quad i \neq j$
- Probabilidades a priori iguales para las 4 clases, esto es, iguales a 0.25
- Categorización: $E_1 = (0,5]$, $E_2 = (5,7]$, $E_3 = (7,8]$, $E_4 = (8,9]$, $E_5 = (9,20]$.

Experimento 2.

- Igual que el anterior excepto que la categorización ahora es $E_1 = (0,6]$, $E_2 = (6,7]$, $E_3 = (7,8]$, $E_4 = (8,11]$, $E_5 = (11,20]$.

Experimento 3.

- Igual al 2 excepto que la función de pérdida para la clasificación equivocada de la clase 3 es cuatro veces la de las otras clases.

Experimento 4.

- Igual al 2 excepto que los costos de medición varían de 0.01 a 0.08.

Compara los resultados de los experimentos 1 y 2 con el método no secuencial, concluyendo que el método utilizando programación dinámica requiere menor número de mediciones que el no secuencial, obteniéndose un mismo porcentaje de clasificaciones correctas.

Al comparar los resultados del experimento 1 y el 2 (que tiene diferente categorización), se observa que se llegan a resultados diferentes, que Fu no menciona, pero que es un punto muy importante, ya que la categorización de las variables, además de llevar a una pérdida de información, lleva a conclusiones diferentes.

La conclusión para el experimento 3, es que al utilizar una función de pérdida

no simétrica, se puede alcanzar un 100% de clasificación correcta en la clase 3 a costa de causar más errores en la clasificación de las otras clases.

Para el experimento 4, al variar los costos de medición de las variables, el número total de mediciones requeridas es menor que en los experimentos anteriores pero con un porcentaje menor de clasificaciones correctas.

2.4 Soluciones Subóptimas

2.4.1 Soluciones de Fu

Fu propone dos soluciones subóptimas para simplificar el problema que presenta la búsqueda de soluciones óptimas en cuanto a dificultad computacional.

En esencia, las dos soluciones subóptimas son la misma, en cuanto a que su característica principal es que son un procedimiento para atrás (backward) pero considerando que la siguiente etapa hacia adelante es la última (Fu le llama "one-stage ahead truncation"), al final de la cual se debe tomar una decisión de asignación.

Las soluciones presentadas son:

1. Solución subóptima con suposición de independencia en las mediciones.

2. Solución subóptima con suposición de dependencia markoviana de primer orden en las mediciones de cada población.

La observación que hace Fu, es que estos procedimientos subóptimos son un "compromiso" y en general no se puede determinar de antemano qué tanto se pierde al utilizarlos. El compromiso consiste en perder optimalidad a cambio de tener factibilidad computacional.

2.4.2 Método de Raiffa

El método propuesto por Raiffa (1961), en esencia tiene la misma idea que la propuesta en el capítulo 3 de este trabajo, solamente que enfocada al Análisis Dicriminante Discreto, considerando variables dicotómicas. El método de Raiffa se describe en el libro de Goldstein y Dillon (1978). Esta descripción es muy breve ya que como los autores explican, es un método impráctico y fué incluido únicamente con la finalidad de dar una visión completa de los métodos secuenciales para el problema de selección de variables.

El método de Raiffa considera dos poblaciones, w_1 y w_2, con costos debidos a clasificación incorrecta a_1 y a_2; probabilidades a priori conocidas para cada población y n variables dicotómicas a medir con ciertos costos de observación asociados.

Al no tener parametrizado el problema, la restricción fuerte en este método es que Raiffa supone conocidas las probabilidades de las 2^n celdas, que, en un problema práctico es casi imposible que se dé. Aún suponiendo conocidas estas probabilidades, la cantidad de cálculos que se deben efectuar es enorme, por lo que se consideró a este método impráctico, sin embargo, como lo señalan Goldstein y Dillon, las ideas son interesantes y potencialmente útiles.

Como se vió a lo largo de este capítulo, los procedimientos secuenciales son una opción muy recomendable para utilizar en problemas de clasificación. Se mostraron

las ventajas y desventajas de cada uno de los métodos tratados. La prueba de Wald, no es recomendable, fundamentalmente por no tomar en cuenta el costo de las mediciones. Las soluciones óptimas no son computacionalmente implantables en el caso continuo. De esto surge la necesidad de considerar métodos subóptimos que sí puedan implantarse y que no decaigan en efectividad.

Las soluciones subóptimas que presenta Fu tienen restricciones muy fuertes; la suposición de independencia de las mediciones o de dependencia markoviana son muy restrictivas. La solución dada por Raiffa no se puede implantar en la práctica. De aquí surge entonces la necesidad de probar con otro tipo de métodos que no sean tan restrictivos. En el siguiente capítulo se presenta uno de estos métodos.

CAPITULO

3

Método Propuesto

3.1 Introducción

En el capítulo anterior se trataron las soluciones óptimas con procedimientos secuenciales al problema de Discriminación. Sin embargo, dichos procedimientos tienen la desventaja de no ser implantables. Es por esto que surge la idea de reconsiderar el problema proponiendo algún otro método que permita ser utilizado en la práctica.

Fu dá algunas soluciones que se pueden llevar a la práctica al suponer cierta estructura en los datos. Pero dicha estructura es muy restrictiva, ya que en general no se puede suponer independencia en las observaciones; Fu, consciente de esto, supone dependencia markoviana de primer orden como un compromiso con la realidad, aunque aún sigue siendo muy restrictiva esta suposición. Incluso la discretización que utiliza Fu, no lleva a una solución óptima; y diferentes discretizaciones llevan a diferentes soluciones.

Por lo anterior, se ve la necesidad de utilizar otro método que no tenga las fallas descritas anteriormente. Un método que, sobre todo, tome en cuenta la estructura real de dependencia de las observaciones.

3.2 Características del Método Propuesto.

El método propuesto tiene las siguientes características:

- Trabaja directamente con $f(x_{n+1}|x_n,\ldots,x_1;\Pi_i)$, es decir, respeta la estructura de dependencia de las variables.

- Las funciones de densidad para cada población, $f(x|\Pi_i)$, están completamente especificadas.

- Es un procedimiento hacia adelante (forward).

- Es secuencial con un límite de m variables para observar.

- No supone un orden preestablecido para la inclusión de las variables, sino que va considerando en cada paso cuál es la mejor variable a medir.

El método hace esencialmente lo siguiente:

1. Del paso 0 al $(m-1)$ se compara el riesgo de tomar una decisión de asignación; es decir, asignar el individuo a una de las p poblaciones, contra el riesgo de tomar alguna otra variable que no ha sido medida. En el paso m, en el peor de los casos, el individuo debe asignarse.

2. Si la decisión es continuar en el paso j $(j = 0, 1, \ldots, m-1)$ se mide la observación de la variable con menor pérdida esperada; se calculan las probabilidades posteriores de pertenencia a cada población y la distribución condicional de x_j, dada la variable x_l que se midió, para todas las variables x_j que no se han incluido. Se regresa al inciso 1.

Una explicación más extensa del método propuesto es la siguiente:

Para facilitar la presentación de las fórmulas, se supondrá que los costos debidos a clasificación equivocada son de la forma $C(j|i) = 1$ si $i \neq j$.

En el paso 0, es decir, sin medir todavía ninguna variable, el riesgo a minimizar está definido en (1.1) y es:

$$E_q\left[L(d_j, \Pi)\right] = \sum_{i=1}^{p} L(d_j, \Pi_i) q_i.$$

Entonces, se tiene que encontrar la decisión d_j que tenga mínimo riesgo esperado. El riesgo o pérdida esperada para la decisión d_j, $j = 1, \ldots, p$ es:

$$\begin{aligned} B(d_j) &= \sum_{i=1}^{p} C(j|i) q_i \\ &= \sum_{\substack{i=1 \\ i \neq j}}^{p} q_i \\ &= 1 - q_j. \end{aligned}$$

Por lo tanto:

$$\min_j B(d_j) = 1 - \max_i q_i.$$

Ahora, si el riesgo mínimo de tomar una decisión sin más observaciones es menor que el costo de observación para cada una de las variables, eso quiere decir que "no conviene" medir ninguna de las variables y que la decisión es asignar el individuo a la población con mayor probabilidad apriori de pertenencia. En caso contrario, conviene medir alguna de las variables, justamente aquella que tenga menor pérdida esperada.

Para cada una de las variables, su pérdida esperada o riesgo es:

$$L(x_j) = c_j + B(d),$$

donde $B(d)$ está definido en (1.7) y es el riesgo de Bayes considerando solamente la variable x_j, en este caso, siendo c_j el costo de observar la variable x_j; por lo tanto, se tiene:

$$L(x_j) = c_j + 1 - \sum_{i=1}^{p} q_i P(i|i, B),$$

donde

$$P(i|i, B) = \int_{B_i} f(x_j|\Pi_i) dx_j.$$

La variable que se medirá en la siguiente etapa es aquella que tenga mínimo riesgo esperado. Es importante hacer notar que la escala en que se deben medir los costos de observación de las variables es entre 0 y $1 - \max_i q_i$.

De aquí, se compara el riesgo de tomar una decisión sin más observaciones con el mínimo riesgo esperado de tomar una observación más. Si el riesgo de tomar una decisión de asignación es menor que el riesgo mínimo de observar una variable, entonces no conviene medir y se toma la decisión. En caso contrario, se observa la

variable con mínimo riesgo esperado y se calculan tanto la probabilidad posterior de pertenencia a cada población como las densidades de cada población condicionadas al haber observado esa variable. Se repite todo el proceso, suponiendo ahora que las probabilidades posteriores son las a priori para el siguiente paso.

La probabilidad posterior de pertenencia a cada población, suponiendo que se observó la variable x_j, es:

$$q_{i|x_j} = \frac{q_i f(x_j|\Pi_i)}{\sum_{k=1}^{p} q_k f(x_j|\Pi_k)}.$$

Nótese que si ahora consideramos a esta probabilidad posterior como a priori, $q_i' \equiv q_{i|x_j}$ y calculamos la posterior condicionada a x_k,

$$q'_{i|x_k} = \frac{q_i' f(x_k|x_j,\Pi_i)}{\sum_{l=1}^{p} q_l' f(x_k|x_j,\Pi_l)},$$

sustituyendo q_i' por su valor

$$q_{i|x_j,x_k} = \frac{\frac{q_i f(x_j|\Pi_i) f(x_k|x_j,\Pi_i)}{\sum_{s=1}^{p} q_s f(x_j|\Pi_s)}}{\frac{\sum_{l=1}^{p} q_l f(x_j|\Pi_l) f(x_k|x_j,\Pi_l)}{\sum_{s=1}^{p} q_s f(x_j|\Pi_s)}}$$

$$= \frac{q_i f(x_j|\Pi_i) f(x_k|x_j,\Pi_i)}{\sum_{l=1}^{p} q_l f(x_j|\Pi_l) f(x_k|x_j,\Pi_l)}.$$

$$= \frac{q_i f(x_j,x_k|\Pi_i)}{\sum_{l=1}^{p} q_l f(x_j,x_k|\Pi_l)}$$

Esto indica que en cada paso en efecto, se está reconstruyendo la posterior.

3.3 Algoritmo

El método propuesto fué implantado en una microcomputadora. Para elaborar el programa de cómputo se construyó el algoritmo que describe los pasos a seguir. El algoritmo es el siguiente:

0. Se define $A = \{x_j; j = 1, \ldots, m\}$ como el conjunto de las variables disponibles, se dan los costos de observación para cada variable, c_j, $j = 1, \ldots, m$ y se dan las probabilidades a priori de pertenencia a cada población, $q_i, i = 1, \ldots, p$.

1. Se evalúa el riesgo de tomar una decisión de asignación sin más observaciones,

$$S = \min_i B(d_i) = 1 - \max_i q_i, \quad i = 1, \ldots, p.$$

Denote por Π_I a la población asociada a S.

2. Si $S \leq \min_j\{c_j \mid x_j \in A\}$ no conviene medir ninguna variable y el individuo se asigna a la población Π_I. Ir al paso 9.

3. Si $A = \emptyset$ se asigna a la población Π_I. Ir al paso 9.

4. Se calcula la pérdida esperada para cada variable $x_j \in A$, $L(x_j)$:

$$L(x_j) = c_j + \left[1 - \sum_{i=1}^{p} q_i P(i|i, B)\right],$$

en donde

$$P(i|i,B) = \int_{B_i} f(x_j|\Pi_i)dx_j$$

es la probabilidad de asignar correctamente a la población Π_i bajo la partición óptima B para la variable x_j.

5. Sea $M = \min_j L(x_j)$ y denote por x_l a la variable asociada a M.

6. Si $S \leq M$ el objeto se asigna a la población Π_J. Ir al paso 9.

7. Se toma la observación x_l. Se actualizan tanto el conjunto A como las probabilidades a priori:

$$A \equiv A - \{x_l\},$$

$$q_i \equiv q_{i|x_l} \quad i = 1, \ldots, p,$$

donde $q_{i|x_l}$ es la probabilidad posterior de pertenencia a la población Π_i. Además, se actualiza la densidad

$$f(x|\Pi_i) \equiv f(x|x_l, \Pi_i) \quad i = 1, \ldots, p.$$

8. Ir al paso 1.

9. Fin.

3.4 Programa de Cómputo

El algoritmo descrito anteriormente se programó en lenguaje Turbo Pascal versión 3.0 para microcomputadora.

La implantación considera:

a. p poblaciones normales m-variadas completamente especificadas.

b. Costos de clasificación de la forma

$$C(j|i) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

Los datos de entrada son:

- Probabilidades a priori de pertenencia a cada población, $q_i, \quad i = 1, \ldots, p.$
- Los costos de medición de cada variable, $c_j, \quad j = 1, \ldots, m.$
- Los vectores de medias y matrices de covarianzas de las poblaciones,

$$\underline{\mu}^{(i)}, \quad V^{(i)}, \quad i = 1, \ldots, p.$$

- El número de individuos de cada población a ser asignados, n_i, pudiéndose

leer los datos de un archivo en disco, o generarse aleatoriamente según las normales especificadas.

Los resultados que se obtienen son:

- La matriz de clasificación.
- El porcentaje de individuos bien clasificados tanto por población como globalmente.
- La matriz de orden, donde se observa la frecuencia y el orden de observación para cada variable.
- Costo total de las n asignaciones, $n = \sum_{i=1}^{p} n_i$. Este costo es la suma de las malas clasificaciones y el costo de las variables observadas.
- El promedio de variables observadas por individuo para cada población y globalmente.

3.4.1 Detalles de cómputo.

Uno de los puntos más interesantes en cuanto a programación fué el cálculo de las regiones involucradas en la probabilidad de asignación correcta para cada población;

ésta se define, para una x_l específica como:

$$P(i|i,B) = \int_{B_i} f(x_l|\Pi_i)dx_l,$$

donde B_i es la región del espacio muestral, en este caso los reales, tal que $P(i|i,B)$ es máxima.

Para calcular las regiones $B_i, \quad i = 1,\ldots,p$, para cada una de las variables $x_j, \quad j = 1,\ldots,p$, se procedió de la siguiente forma:

0. Sea x una de las variables $x_j, \quad j = 1,\ldots,p$.

1. Para cada par de poblaciones Π_i y Π_j, se calculan los puntos de intersección, es decir, los puntos en los que $q_i f(x|\Pi_i) = q_j f(x|\Pi_j)$. Si las varianzas son iguales existe un solo punto de intersección z_1; si no es el caso, entonces existen dos puntos z_1, z_2.

2. Estos puntos de intersección forman una partición de $\Re$:, $(-\infty, z_1], (z_1, \infty)$ en el caso de un solo punto ó bien $(-\infty, z_1], (z_1, z_2], (z_2, \infty)$ en el caso de dos puntos.

3. En cada una de estas regiones, se definen

$$B_{ij}^{(k)} = \left\{x|q_i f(x|\Pi_i) \geq q_j f(x|\Pi_j) \quad j \neq i\right\} \quad k = 1,2.$$

El conjunto $B_{ij}^{(k)}$ representa la región de $\Re$ en donde la población Π_i "domina" a la población Π_j. En el caso de un solo punto de intersección entre estas dos poblaciones, $k = 1$; si hay dos puntos de intersección $k = 1,2$.

4. Se calcula

$$B_i = \left(\bigcap_{j \neq i} B_{ij}^{(1)}\right) \cup \left(\bigcap_{j \neq i} B_{ij}^{(2)}\right)$$

para cada población $i = 1, \ldots, p$ y con esto se calcula $P(i|i, B)$.

Ejemplo del cálculo de las regiones B

Tres poblaciones normales con varianzas iguales

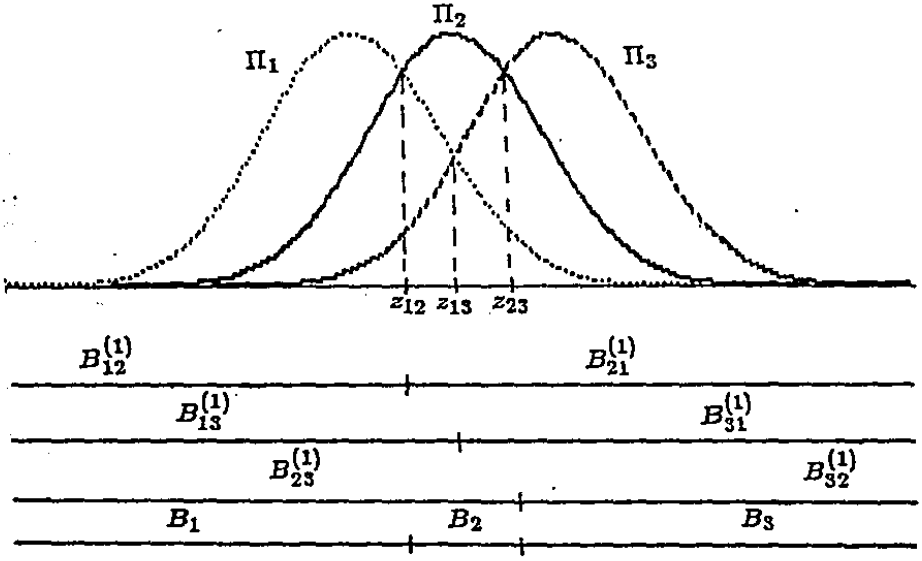


Figura 3.1

En la figura 3.1 se ilustra el procedimiento; z_{ij} es el punto de intersección de las poblaciones Π_i y Π_j. $B_{ij}^{(1)}$ es la región en que la población Π_i domina a la Π_j, y B_i es la región de la población Π_i que interesa.

Es necesario indicar que para evitar problemas numéricos se tuvo que definir numéricamente $(-\infty, \infty)$. Para esto se "calculan" $(-\infty_i, \infty_i)$ para cada población Π_i

y cada variable x_j, como

$$(-\infty_i, \infty_i) \equiv (\mu_j^{(i)} - 6\sigma_{jj}^{(i)}, \mu_j^{(i)} + 6\sigma_{jj}^{(i)}),$$

y se definió

$$(-\infty, \infty) \equiv (\min_i\{-\infty_i\}, \max_i\{\infty_i\}).$$

Otro punto importante que se tuvo que incluir en el programa fué el que al observar una variable, si su valor estaba fuera del intervalo $(-\infty_i, \infty_i)$ para alguna población Π_i, esta población se descartaba para la asignación del individuo. Esto se hizo para evitar problemas numéricos, ya que el cálculo de la densidad para un valor fuera del intervalo considerado es prácticamente cero.

Podría suceder que una observación estuviera fuera de rango para todas las poblaciones, en este caso, el programa no puede asignarla a ninguna de ellas y se asigna a una población hipotética Π_0. En los ejemplos se verán algunos de estos casos.

El cálculo de las densidades de cada población condicionadas a haber observado alguna variable, se hace como sigue:

Suponga que la población Π_i tiene un vector de medias $\underline{\mu}$ y una matriz de covarianzas V, originalmente de dimensiones m y $m \times m$ respectivamente. Sea $r = m$.

1. Si se observó la variable x_l entonces $r \equiv r - 1$.

V se puede descomponer en

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

donde V_{22} es la varianza de la variable x_l, V_{11} es la matriz de varianzas y covarianzas original eliminando el renglón y columna l-ésimos, V_{12} es el vector de covarianzas de la variable x_l con todas las demás y $V_{21} = V_{12}^t$.

2. Se elimina la componente l-ésima del vector de medias $\underline{\mu}$, quedando ahora de dimensión r.

3. $\underline{\mu} \equiv \underline{\mu} + V_{12}V_{22}^{-1}(x_l - \mu_l)$.

4. Se elimina el renglón y columna l-ésimos de la matriz de covarianzas, quedando ahora una matriz de $r \times r$.

5. $V \equiv V_{11} - V_{12}V_{22}^{-1}V_{21}$.

El cálculo de la probabilidad posterior de pertenencia a cada población y de las densidades condicionales de cada población es el punto más importante del método ya que de esta manera se está respetando la estructura de dependencia de las variables.

3.5 Ejemplos

Los ejemplos que se presentan a continuación fueron estudiados para probar el método. La asignación de los costos es completamente artificial y al final se tendrá el porcentaje de mala clasificación, ya que conocemos de antemano la población a la que pertenece cada individuo.

Todas las simulaciones efectuadas se realizaron considerando $q_i = 1/p$, $i = 1, 2, \ldots, p$.

Para cada uno de los ejemplos se ilustra el comportamiento marginal que tiene cada una de las variables para dar una idea del "traslape" que tienen las poblaciones. Las Gráficas se encuentran en el Apéndice G.

Las tablas de resultados, situadas en el Apéndice T, contienen la siguiente información:

- La matriz de clasificación.
- La columna % representa el porcentaje de individuos bien clasificados en cada población, siendo el último renglón el porcentaje global.
- La columna Prom es el promedio de variables observadas por individuo; este promedio se dá por población y en el último renglón, el global.

- La columna Costo muestra el costo de observación considerado en cada ejemplo.

- La matriz de Orden de Entrada proporciona la información acerca del orden en que se incluyeron las variables en el análisis.

- El Costo Total representa el costo de las n asignaciones y es la suma de los costos de observación de las variables medidas y el total de individuos mal clasificados.

3.5.1 Datos de Fisher

El primer ejemplo con que se probó el método es el problema tratado por Fisher de los datos del género Iris. Este es un conjunto de datos clásico, que se ha utilizado en innumerables ocasiones en la bibliografía.

Se tienen tres poblaciones de Iris:

Π_1 Setosa

Π_2 Versicolor

Π_3 Virgínica

y cuatro mediciones realizadas en cada individuo:

x_1 longitud del sépalo

x_2 ancho del sépalo

x_3 longitud del pétalo

x_4 ancho del pétalo

Para este caso, supondremos que los vectores de medias y la matriz de covarianza muestrales son los verdaderos parámetros. Los vectores de medias para cada población son:

$$\underline{\mu}^{(1)} = \begin{pmatrix} 50.06 \\ 34.28 \\ 14.62 \\ 2.46 \end{pmatrix} \qquad \underline{\mu}^{(2)} = \begin{pmatrix} 59.36 \\ 27.70 \\ 42.60 \\ 13.26 \end{pmatrix} \qquad \underline{\mu}^{(3)} = \begin{pmatrix} 65.88 \\ 29.74 \\ 55.52 \\ 20.26 \end{pmatrix}$$

La matriz de varianzas-covarianzas común es:

$$V = \begin{pmatrix} 26.501 & 9.272 & 16.751 & 3.840 \\ 9.272 & 11.539 & 5.524 & 3.271 \\ 16.751 & 5.524 & 18.519 & 4.267 \\ 3.840 & 3.271 & 4.267 & 4.188 \end{pmatrix} \tag{3.1}$$

Se tiene la información de 150 plantas, 50 de cada población, que se asignarán utilizando el método propuesto.

La gráfica 1 representa el comportamiento marginal de cada variable. Se puede notar que las variables x_1 y x_2 tienen muy "traslapadas" a las tres poblaciones, las

variables x_3 y x_4 distinguen más entre las poblaciones, siendo x_4 la variable que separa más a las tres poblaciones. Este comportamiento se verá reflejado en los resultados de las simulaciones presentados más adelante.

RESULTADOS.

Los resultados obtenidos para estos datos se resumen en el siguiente cuadro:

TABLA	COSTO TOTAL DE LAS 4 OBSERVACIONES	OBJETIVO
1	0	Comparar Método propuesto con Discriminante Clásico
2	0.07	Comparar Resultados del Método propuesto al asignar diferentes Costos a las Observaciones
3	0.14	Igual a Tabla 2
4	0.28	Igual a Tabla 2

Tabla 1.

La Tabla 1 hace una comparación de Discriminante Clásico con el Método Propuesto; se consideró un costo de observación para cada variable igual a cero.

Se observa que el método propuesto tiene una matriz de clasificación igual a la del método clásico, por consiguiente con el mismo costo total. Sin embargo, el método propuesto mide en promedio menos variables; esto sucede por la restricción del intervalo $(-\infty, \infty)$ que se tuvo que considerar debido a la capacidad de la computadora

en que se implantó. Se observa en la matriz de orden de entrada de las variables que el método propuesto consideró a la variable x_4 como la más informativa, es decir, con menor pérdida esperada, y es por esto que entra siempre en primer lugar; aquí se observa lo que se comentó en las gráfica 1 de comportamiento marginal.

Las Tablas 2, 3 y 4 muestran los resultados obtenidos aplicando el método propuesto, con diferentes asignaciones de costos. En las tres tablas se consideran tres casos diferentes:

a) costos iguales para las cuatro variables.

b) x_4 cuatro veces más cara que las demás.

c) x_4 la mitad de cara que las demás.

Tabla 2.

La Tabla 2 considera un costo de las cuatro mediciones igual a 0.07. Se observa que tanto en el caso $2a$) como en el $2c$) la variable x_4 es la primera en medirse; si se castiga a esta variable con más costo (caso $2b$)), la primera en medirse es x_3, lo que está de acuerdo con la gráfica 1 en donde se observa que la segunda variable que distingue mejor a las poblaciones es x_3.

En cuanto a la población Π_1, no se tiene ningún problema en la asignación de

sus individuos. Alcanza el 100% de bien clasificados en los tres casos tratados, esto es debido a que está más separada de las poblaciones Π_2 y Π_3 en las variables x_3 y x_4, como se observa en la gráfica 1 y de hecho solo se necesita medir una variable para asignar a los individuos de Π_1 sin error.

La población Π_2 tiene un 96% de bien clasificados en los tres casos, con un promedio de variables observadas de 1.16 en los casos $2a$) y $2c$) donde se observa primero la variable x_4, y sube este promedio a 1.44 en el caso $2b$) cuando se observa primero la variable x_3. En conclusión, si se observa en primer lugar una variable que no es la mejor discriminadora, se tendrán que observar más variables para la asignación, con un resultado similar, como es lógico.

En cuanto a la población Π_3, el caso $2a$) y $2c$) resultaron equivalentes con 94% de bien clasificados y un promedio de 1.06 variables observadas por asignación. El caso $2b$) requiere un promedio mayor de variables observadas con la ventaja, por otro lado, de alcanzar el 96% de bien clasificados.

Al observar las estadísticas globales, esto es, de todas las poblaciones, parece ser que el caso $2b$) es el mejor, con mayor % de bien clasificados que los casos $2a$) y $2c$) pero con un promedio de variables observadas mayor.

Considerando ahora el costo total en los tres casos, el menor es para el caso $2c$) siguiendo el $2b$) y por último el $2a$). Es interesante mencionar aquí la comparación

de costos del método propuesto con el método de discriminación clásico. El costo total del discriminante clásico es mucho mayor que el del método propuesto, ya que con un método no secuencial se miden todas las variables, con un aumento correspondiente en costos.

Tabla 3.

La Tabla 3 considera un costo de las cuatro mediciones igual a 0.14, el doble que en la Tabla 2. Las conclusiones son equivalentes a las de la Tabla 2, excepto en dos puntos:

-En el caso $3b$) el número de mediciones de la variable x_4 es 17 solamente, ya que esta variable cuesta el doble que en el caso $2b$) donde se observa 28 veces; además se tiene mayor participación de las variables x_1 y x_2.

-Se incrementa el % de bien clasificados para la población Π_3, pasando de 96% en el caso $2b$) a 98% en este caso.

Al comparar los resultados globales de esta tabla con los de la Tabla 2 se tiene un % de bien clasificados equivalente en los tres casos estudiados, pero se observa una ligera disminución del promedio de variables observadas en la Tabla 3.

En cuanto al costo total, el mejor caso fué el $3b$) siguiéndole el $3c$) y por último

el $3a$). La comparación con el costo total en el caso no secuencial indica una diferencia grande.

Tabla 4.

La Tabla 4 considera costos de las cuatro mediciones igual a 0.28, el doble que en la Tabla 3.

Se observa una disminución del % de bien clasificados para la población Π_3 y un aumento para la población Π_2, en los casos $4a$) y $4c$) comparados con los casos $3a$) y $3c$), de tal manera que en forma global el % de bien clasificados se mantiene equivalente al de la Tabla 3. Pero al ser ahora las variables más caras, disminuye el promedio de variables observadas. En cuanto a costo total sucede lo mismo que en la Tabla 3; y hay una gran diferencia al comparar los costos totales en los tres casos de esta tabla con el costo total del caso no secuencial.

En conclusión, el método propuesto funciona bien, reconstruyendo la matriz de clasificación calculada con el discriminante clásico. Al considerar costos de medición diferentes de cero para las variables, se observan en promedio menos variables con un % de bien clasificados aceptable y una diferencia (que llega a ser muy grande) del costo total de asignaciones del método propuesto con el costo total del discriminante clásico. En las tablas 2, 3 y 4, los casos a) y c) resultaron equivalentes, ya que la variable x_4 tiene la menor pérdida esperada; esto es, es lo mismo considerar que su

costo es igual al costo de las demás variables, o que es más barata que las otras; estos dos casos se diferencian solo en el costo total.

3.5.2 Variantes a los datos de Fisher

Como se observó en la Gráfica 1, los datos de Fisher se comportan bastante bien, en el sentido de que no hay demasiado traslape entre las poblaciones. Sobre todo la población Π_1 está bastante alejada de las otras dos y la asignación de sus individuos no presenta mayor problema, como se observa en las tablas 1, 2, 3 y 4.

El objetivo ahora, es probar el método propuesto con datos que tengan más traslape entre las poblaciones. La primera idea fué el transformar los datos de Fisher de tal manera que se confunda más la población Π_1 con las otras dos. Para esto se transformó al vector de medias de la población Π_1 de la siguiente manera:

$$\underline{\mu}^{(1)} \equiv \frac{\underline{\mu}^{(1)} + k\left(\underline{\mu}^{(2)} + \underline{\mu}^{(3)}\right)}{1+2k}$$

y se consideraron dos casos: $k = 1$ y $k = 2$.

Para tratar de hacer el traslape más evidente, y estudiar el comportamiento del método con matrices de covarianza diferentes, se empleó una rotación de las poblaciones con la matriz

$$P(\theta,\alpha) = \begin{pmatrix} cos\,\theta & -\text{sen}\,\theta & 0 & 0 \\ \text{sen}\,\theta & cos\,\theta & 0 & 0 \\ 0 & 0 & cos\,\alpha & \text{sen}\,\alpha \\ 0 & 0 & -\text{sen}\,\alpha & cos\,\alpha \end{pmatrix}$$

considerando tres rotaciones:

$$P_1 = P(0°, 0°)$$

$$P_2 = P(45°, 45°)$$

$$P_3 = P(90°, 90°)$$

con las que se transformó a las matrices de covarianza de la siguiente forma:

$$V_1 = P_1 V P_1'$$

$$V_2 = P_2 V P_2'$$

$$V_3 = P_3 V P_3'$$

donde V está especificada en (3.1), y se generaron aleatoriamente los 50 casos de cada población.

Los casos tratados se resumen en el siguiente cuadro:

TABLA	COSTO TOTAL DE LAS 4 OBSERVACIONES	K	OBJETIVO
5	0	1	Comparar Discriminante Cuadrático con Lineal
6	0.28	1	Comparar Resultados del Método propuesto al asignar diferentes Costos a las Observaciones
7	0.28	2	igual a Tabla 6

Caso $k = 1$.

Los vectores de medias de cada población resultaron ser:

$$\underline{\mu}^{(1)} = \begin{pmatrix} 58.43 \\ 30.57 \\ 37.58 \\ 11.99 \end{pmatrix} \qquad \underline{\mu}^{(2)} = \begin{pmatrix} 59.36 \\ 27.70 \\ 42.60 \\ 13.26 \end{pmatrix} \qquad \underline{\mu}^{(3)} = \begin{pmatrix} 65.88 \\ 29.74 \\ 55.52 \\ 20.26 \end{pmatrix}$$

y las matrices de covarianzas:

$$V_1 = \begin{pmatrix} 26.501 & 9.272 & 16.751 & 3.840 \\ 9.272 & 11.539 & 5.524 & 3.271 \\ 16.751 & 5.524 & 18.519 & 4.267 \\ 3.840 & 3.271 & 4.267 & 4.188 \end{pmatrix}$$

$$V_2 = \begin{pmatrix} 9.745 & 7.478 & 5.896 & -5.327 \\ 7.478 & 28.283 & 14.688 & -7.579 \\ 5.896 & 14.688 & 15.615 & -7.163 \\ -5.327 & -7.579 & -7.163 & 7.084 \end{pmatrix}$$

$$V_3 = \begin{pmatrix} 11.539 & -9.272 & -3.271 & 5.524 \\ -9.272 & 26.501 & 3.84 & -16.751 \\ -3.271 & 3.84 & 4.188 & -4.267 \\ 5.524 & -16.751 & -4.267 & 18.519 \end{pmatrix}$$

La gráfica 2 representa el comportamiento marginal; se observa que se logró un traslape mayor entre las poblaciones que el que tienen los datos originales de Fisher (gráfica 1).

Como una forma de ilustrar lo que hace el método propuesto, se considera el caso de asignar un individuo, que se sabe es de la población Π_2, cuyo vector de observaciones es $(60.789, 26.21, 44.49, 13.556)$. Se considerará que el costo de observación de las variables es cero.

En la gráfica 2 se observa que la variable más discriminatoria es x_3, que resulta ser, con el método, la de menor pérdida esperada; por lo tanto, el programa decide

observar x_3. Dada esta observación, las probabilidades posteriores de pertenencia resultan ser:

$$q_1 = .2209 \quad q_2 = .779 \quad q_3 = .00000082$$

En la gráfica 3 se observa el comportamiento marginal de las densidades multiplicadas por las probabilidades posteriores de pertenencia a las poblaciones y condicionadas a haber observado la variable x_3. Nótese que la población Π_3 ya no aparece porque su probabilidad posterior de pertenencia es muy pequeña. Se observa también que la población que domina es Π_2.

En el siguiente paso del procedimiento, la variable que tiene menor pérdida esperada es x_1, por lo que se observa su valor. De aquí las probabilidades posteriores de pertenencia a cada población se convierten en:

$$q_1 = .1091 \quad q_2 = .8909 \quad q_3 = 2.58 \times 10^{-1}.$$

La gráfica 4 presenta las densidades condicionales a haber observado x_1 multiplicadas por su respectiva q. Se observa que la población Π_2 domina aún más a Π_1, y Π_3 ya está fuera de consideración.

El procedimiento decide continuar observando, cuando quedan solamente x_2 y x_4 por medir. Al realizar los cálculos, la variable con menor pérdida esperada resulta ser x_4, por lo que se observa su valor. Dado el valor de x_4, las probabilidades posteriores

resultan ser:

$$q_1 = .1409 \quad q_2 = .859 \quad q_3 = 6.54 \times 10^{-15}.$$

La gráfica 5 representa el comportamiento marginal de la variable x_2 al considerar las probabilidades posteriores arriba descritas.

Ya que la decisión de parar el procedimiento y asignar el individuo a alguna de las poblaciones todavía no se dá, queda solamente la variable x_2 por medir y el programa decide medirla, resultando ahora las probabilidades posteriores de pertenencia iguales a:

$$q_1 = .05783 \quad q_2 = .9421 \quad q_3 = 5.75 \times 10^{-20}.$$

Para este ejemplo, se tuvieron que medir las cuatro variables para poder tomar la decisión final, el asignar el individuo a la población Π_2 que es la que tuvo mayor probabilidad posterior de pertenencia. Esto no sucede en general, sobre todo cuando se consideran costos de observación de las variables diferentes de cero.

RESULTADOS

Tabla 5.

La Tabla 5 presenta una comparación entre tomar las matrices de covarianza diferentes, como lo son en realidad, o suponerlas iguales. En la tabla, $V = 1/3(V_1 + V_2 + V_3)$. Los costos de medición de las cuatro variables son cero.

En el caso $5a$) se observa que la transformación realizada lleva a más traslape entre Π_1 y Π_2, pero no se alcanzó a traslapar Π_3 con las otras. Esto se ve en la matriz de clasificación donde Π_3 tiene un 96% de bien clasificados y dos casos que no se pudieron clasificar.

La matriz de clasificación mejora al considerar matrices de covarianza diferentes, caso $5a$), que al considerarlas iguales, caso $5b$), excepto para Π_3 donde se mantiene el % de bien clasificados en ambos casos. Además el promedio de variables observadas por población y globalmente es menor en $5a$); por consecuencia el costo total es menor en $5a$) que en $5b$). Se observa entonces, que vale la pena considerar el discriminante cuadrático.

Tabla 6

La Tabla 6 hace una comparación para diferentes costos de observación de las variables y matrices de covarianza diferentes.

Se consideraron tres casos para los costos de las cuatro mediciones:

6a) 0.00

6b) 0.28 costos iguales para las cuatro variables.

6c) 0.28, x_3 más cara que las demás.

Se observa que al tener costos de observación diferentes de cero, el promedio de variables observadas disminuye.

El % de bien clasificados es equivalente en 6a) y 6b), sin embargo, en 6b) se "arregla" el problema de los dos casos de la población Π_3 que no se clasifican en 6a). Una explicación a esto es que, debido a que se observan menos variables en 6b), no se observaron las variables que causaban confusión en estos casos en 6a).

Al parecer la variable más informativa es x_3 ya que se mide en primer lugar en 6a) y 6b). Al hacerla más cara en 6c) se observa una disminución en el % de bien clasificados y un aumento en el promedio de variables observadas que en 6b), por consiguiente 6c) tiene un costo total más alto que el de 6b).

Caso $k = 2$.

Los vectores de medias de cada población resultaron ser:

$$\underline{\mu}^{(1)} = \begin{pmatrix} 60.11 \\ 29.83 \\ 42.17 \\ 13.2 \end{pmatrix} \qquad \underline{\mu}^{(2)} = \begin{pmatrix} 59.36 \\ 27.70 \\ 42.60 \\ 13.26 \end{pmatrix} \qquad \underline{\mu}^{(3)} = \begin{pmatrix} 65.88 \\ 29.74 \\ 55.52 \\ 20.26 \end{pmatrix}$$

y las matrices de covarianzas son las mismas que para el caso $k = 1$.

La gráfica 6 muestra el comportamiento marginal de las poblaciones al realizar

esta transfomarción. Se observa que, a diferencia de la gráfica 2, se tiene mucho más traslape entre las poblaciones Π_1 y Π_2.

RESULTADOS

Tabla 7

La tabla 7 hace una comparación del funcionamiento del procedimiento cuando se consideran diferentes costos de medición de las variables. Se toma como base costos cero, y después con la suma de los costos de las cuatro variables igual a 0.28, se consideran dos casos; la variable más discriminadora, x_4, con costo igual al de las otras tres y cuando a esta variable se le castiga con un costo cuatro veces mayor que el de las otras.

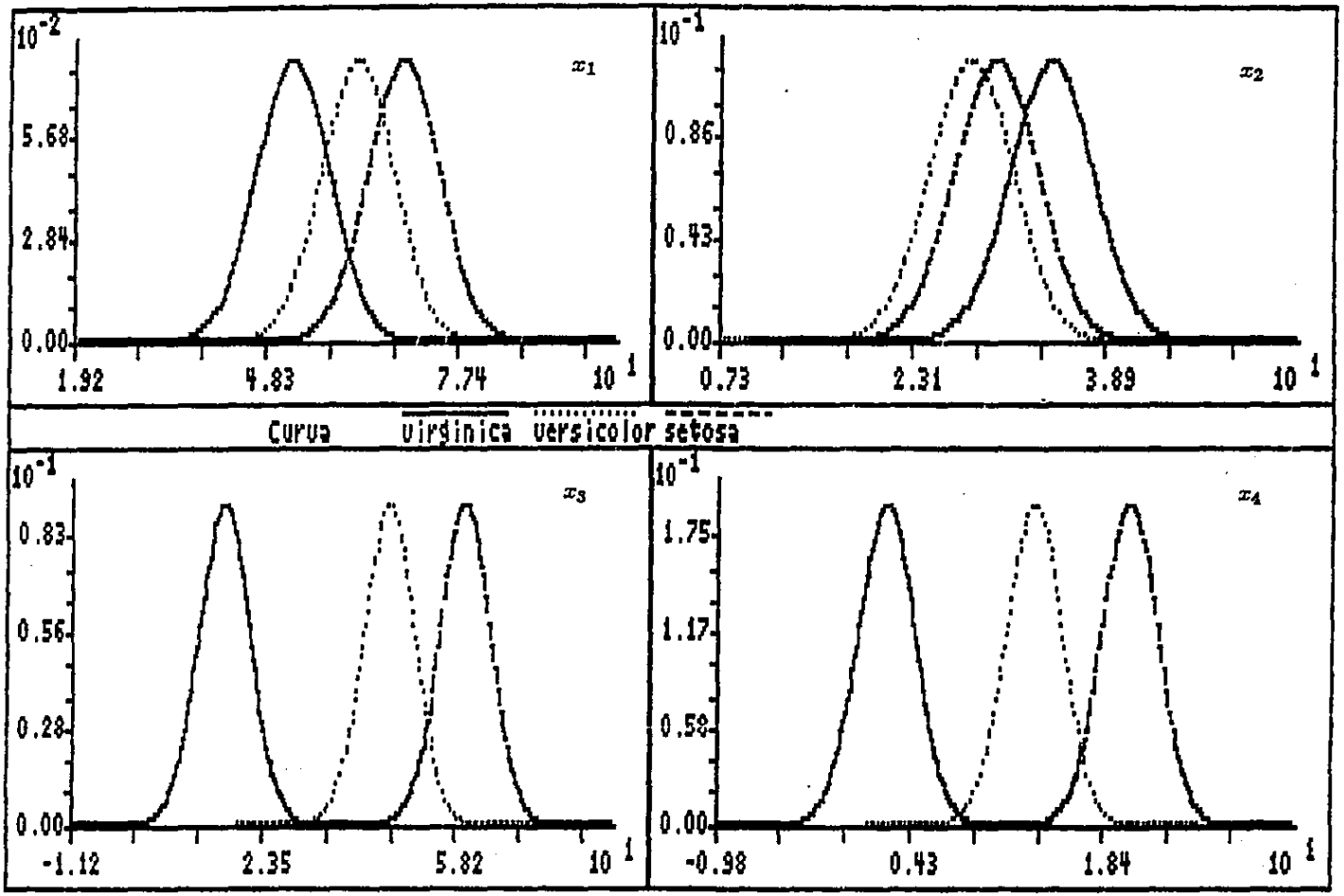
Las conclusiones para esta tabla son las siguientes:

En el caso $7c$), al considerar a la variable x_4 "cara", la primera que entra es x_1, que parece ser la que mejor distingue a la población Π_3 de las otras dos, ya que dá un 100% de bien clasificados. Esta variable no es la mejor variable para discriminar a la población Π_1, por que en este caso, esta población tiene menor porcentaje de bien clasificados al comparar con los casos $7a$) y $7b$).

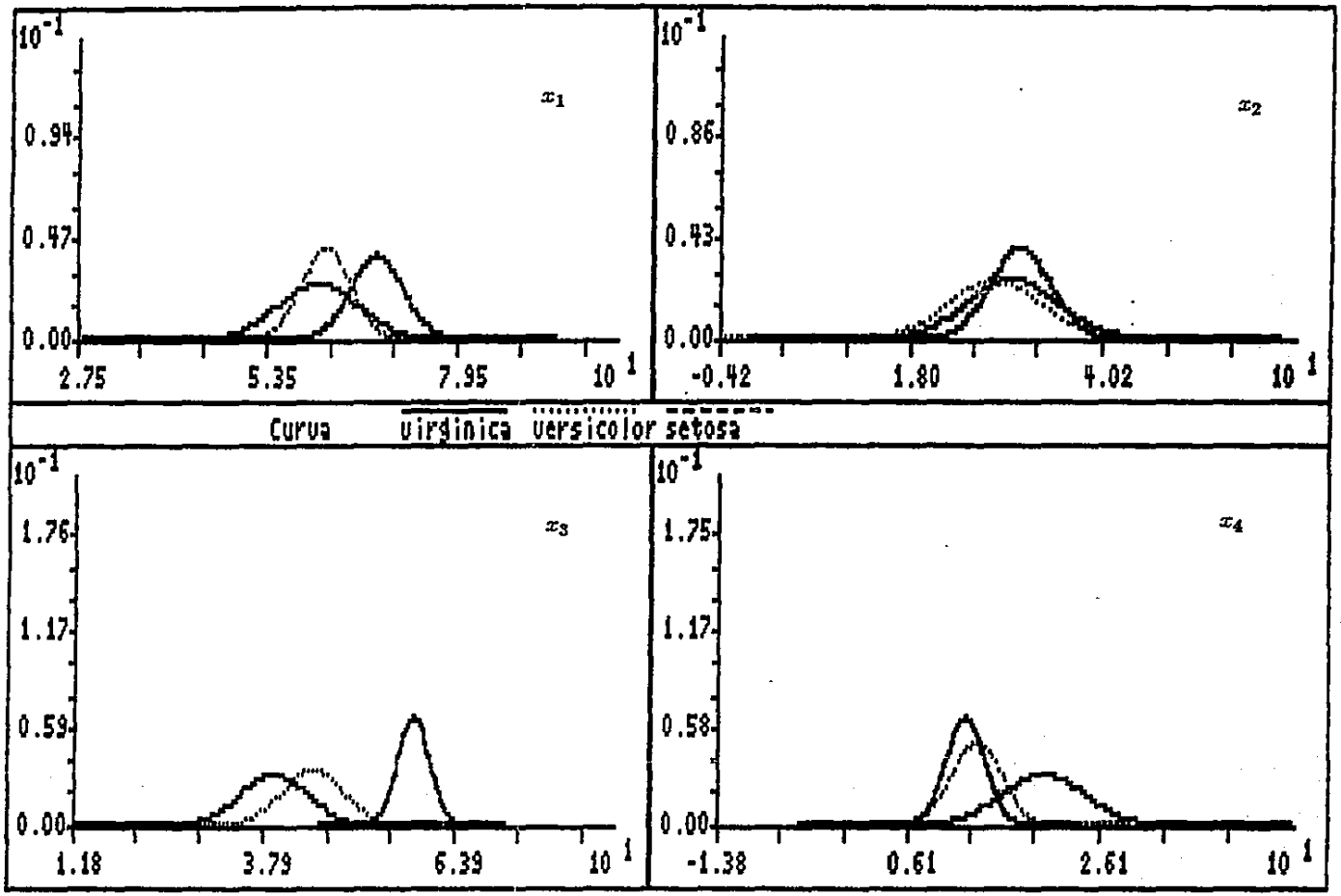
Al comparar esta tabla con la tabla 6, se observa una disminución del % de bien clasificados global en los tres casos, además de un incremento considerable

en el costo total. Lo anterior fué porque se consiguió un traslape mayor entre las poblaciones.

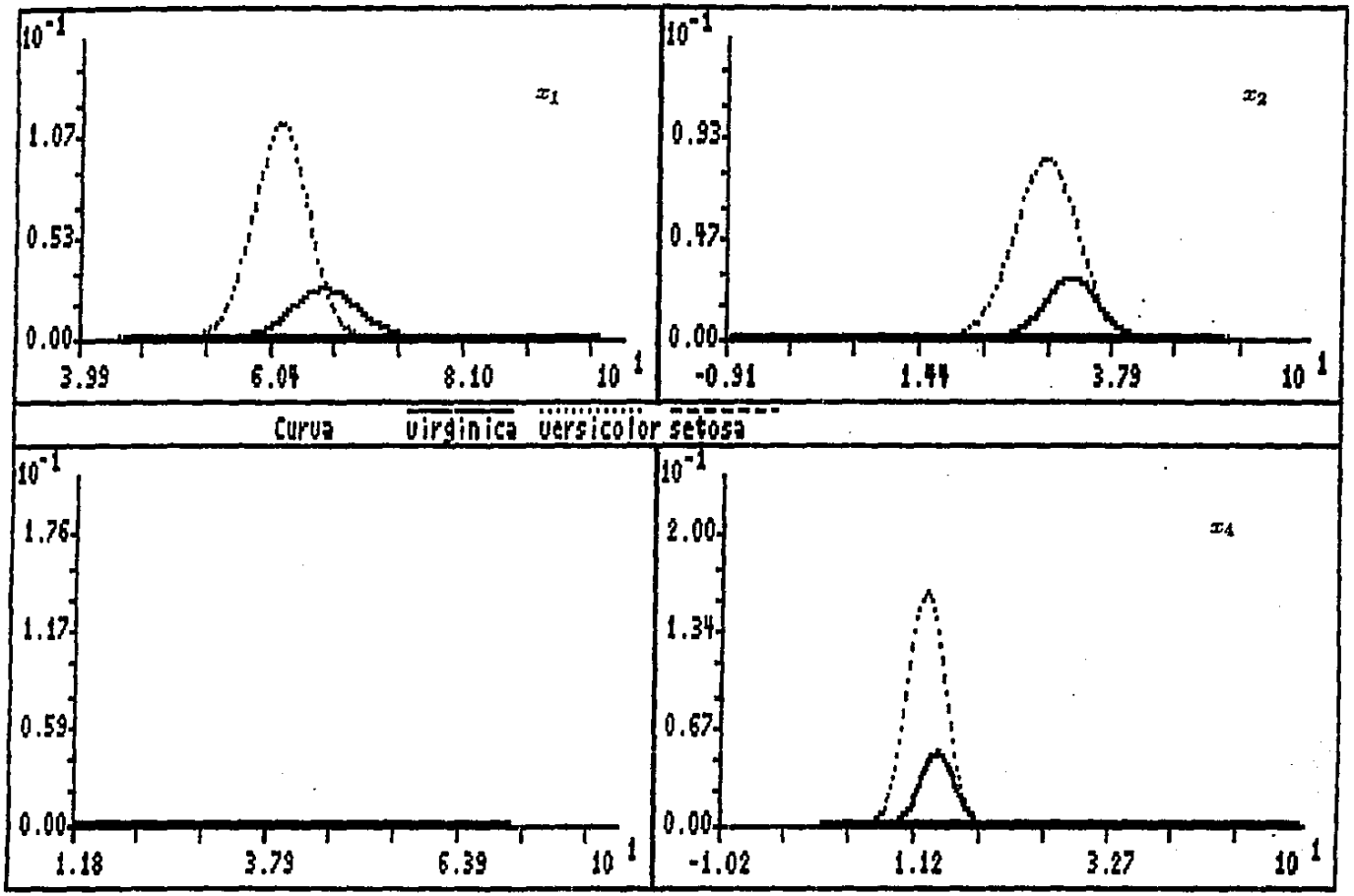
APENDICE G


10^{-2}
5.68
2.84
0.00
1.92
4.83
7.74
10
x_1
10^{-1}
0.86
0.43
0.00
0.73
2.31
3.89
10
x_2
Curva
virginica
versicolor
setosa
10^{-1}
0.83
0.56
0.28
0.00
-1.12
2.35
5.82
10
x_3
10^{-1}
1.75
1.17
0.58
0.00
-0.98
0.43
1.84
10
x_4

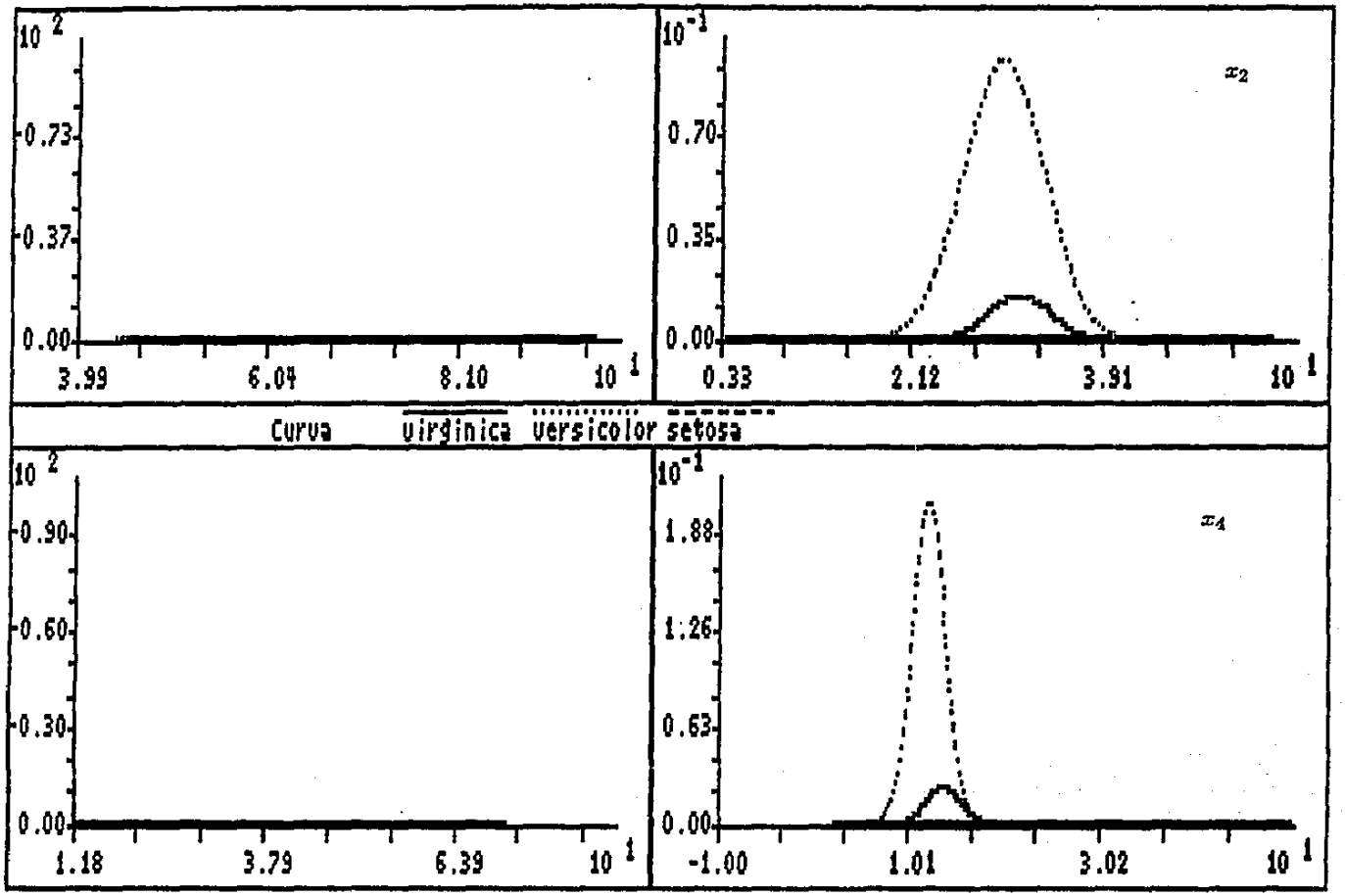
GRAFICA 1


10-1
0.94
0.47
0.00
2.75
5.35
7.95
10 1
x1
10-1
0.86
0.43
0.00
-0.42
1.80
4.02
10 1
x2
Curva virginica versicolor setosa
10-1
1.76
1.17
0.59
0.00
1.18
3.79
6.39
10 1
x3
10-1
1.75
1.17
0.58
0.00
-1.38
0.61
2.61
10 1
x4

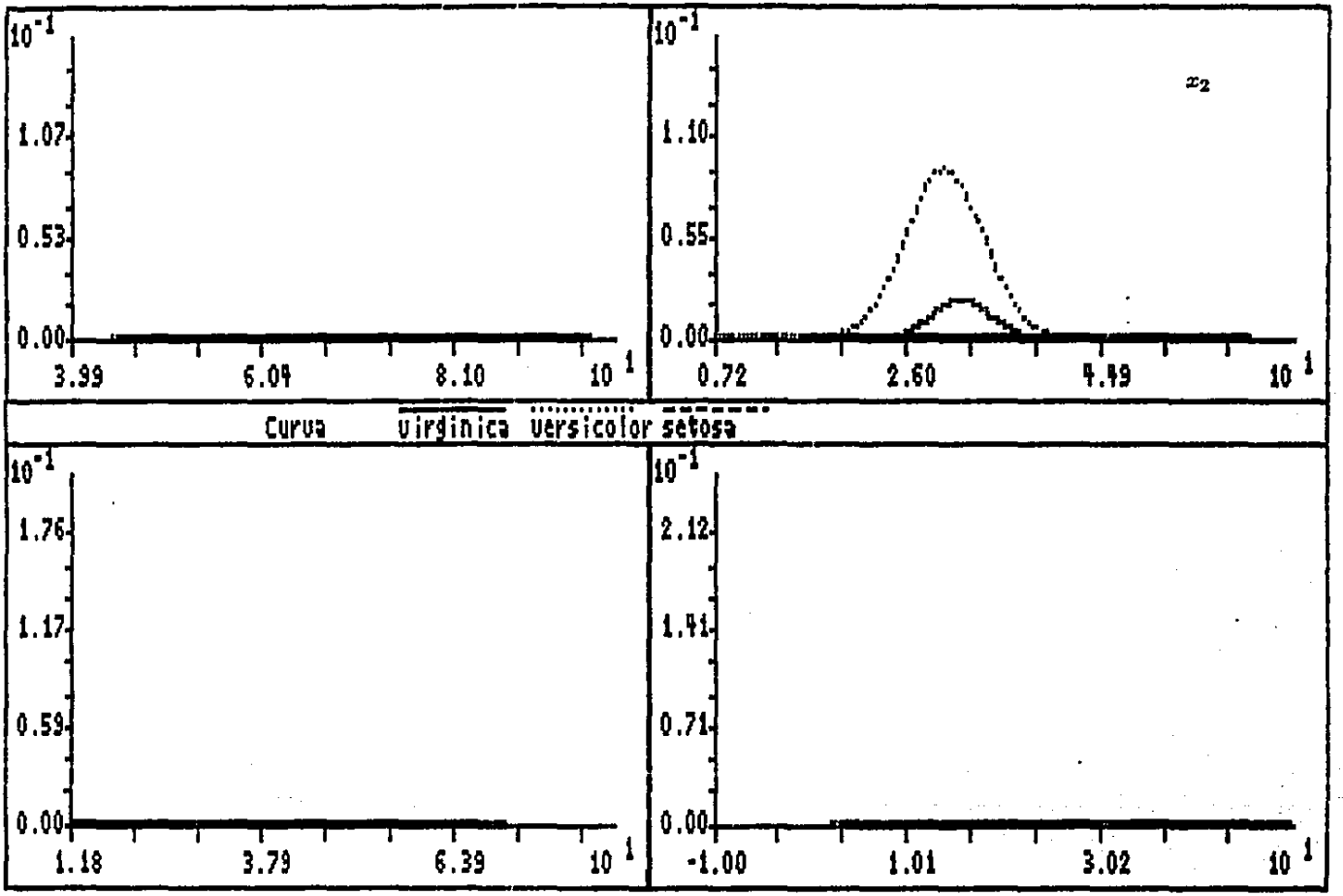
GRAFICA 2


10^{-1}
x_1
1.07
0.53
0.00
3.99
6.04
8.10
10 1
10^{-1}
x_2
0.93
0.47
0.00
-0.91
1.44
3.79
10 1
Curva
virginica
versicolor
setosa
10^{-1}
1.76
1.17
0.59
0.00
1.18
3.79
6.39
10 1
10^{-1}
x_4
2.00
1.34
0.67
0.00
-1.02
1.12
3.27
10 1

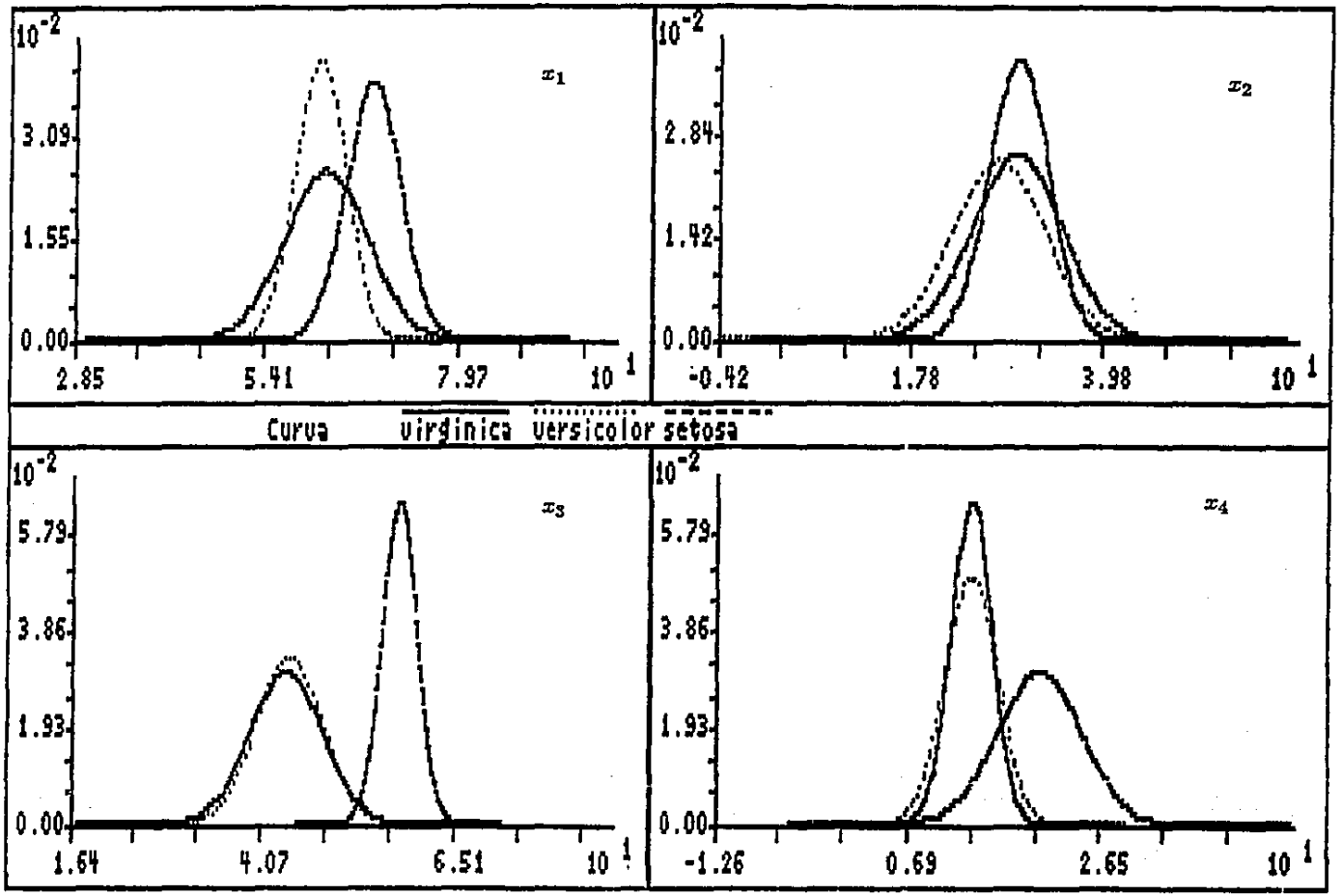
GRAFICA 3


10 2
0.73
0.37
0.00
3.99
6.04
8.10
10 1
10 -1
0.70
0.35
0.00
0.33
2.12
3.91
10 1
x_2
Curva
virginica
versicolor
setosa
10 2
0.90
0.60
0.30
0.00
1.18
3.79
6.39
10 1
10 -1
1.88
1.26
0.63
0.00
-1.00
1.01
3.02
10 1
x_4

GRAFICA 4


10^{-1}
1.07
0.53
0.00
3.99
6.09
8.10
10 1
10^{-1}
1.10
0.55
0.00
x_2
0.72
2.60
4.49
10 1
Curva
virginica
versicolor
setosa
10^{-1}
1.76
1.17
0.59
0.00
1.18
3.79
6.39
10 1
10^{-1}
2.12
1.41
0.71
0.00
-1.00
1.01
3.02
10 1

GRAFICA 5


10^{-2}
3.09
1.55
0.00
2.85
5.41
7.97
10 1
x_1
10^{-2}
2.84
1.42
0.00
-0.42
1.78
3.98
10 1
x_2
Curva virginica versicolor setosa
10^{-2}
5.79
3.86
1.93
0.00
1.64
4.07
6.51
10 1
x_3
10^{-2}
5.79
3.86
1.93
0.00
-1.26
0.69
2.65
10 1
x_4

GRAFICA 6

APENDICE T

TABLA 1

Discriminante Clásico

Costo de las cuatro mediciones = 0.00

Pob	Asignada				Variable			Orden de Entrada			
Real	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
Π_1	50	0	0	100	4	0	x_1	150	0	0	0
Π_2	0	48	2	96	4	0	x_2	150	0	0	0
Π_3	0	1	49	98	4	0	x_3	150	0	0	0
Costo Total: 3				98	4	0	x_4	150	0	0	0

Método Propuesto

Costo de las cuatro mediciones = 0.00

Pob	Asignada				Variable			Orden de Entrada			
Real	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
Π_1	50	0	0	100	2.92	0	x_1	0	0	86	26
Π_2	0	48	2	96	4.00	0	x_2	0	39	39	58
Π_3	0	1	49	98	3.82	0	x_3	0	111	26	8
Costo Total: 3				98	3.58	0	x_4	150	0	0	0

TABLA 2

Costo de las cuatro mediciones = .07

Pob	Asignada				Variable			Orden de Entrada			
Real	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
2*a*: Costos iguales											
Π_1	50	0	0	100	1.00	.0175	x_1	0	0	4	0
Π_2	0	48	2	96	1.16	.0175	x_2	0	0	0	1
Π_3	0	3	47	94	1.06	.0175	x_3	0	6	0	0
Costo Total: 7.82				96.7	1.07	.0175	x_4	150	0	0	0
2*b*: x_4 'cara'											
Π_1	50	0	0	100	1.00	.01	x_1	0	4	12	0
Π_2	0	48	2	96	1.44	.01	x_2	0	0	0	5
Π_3	0	2	48	96	1.54	.01	x_3	150	0	0	0
Costo Total: 6.83				97.3	1.33	.04	x_4	0	28	0	0
2*c*: x_4 'barata'											
Π_1	50	0	0	100	1.00	.02	x_1	0	0	4	0
Π_2	0	48	2	96	1.16	.02	x_2	0	0	0	1
Π_3	0	3	47	94	1.06	.02	x_3	0	6	0	0
Costo Total: 6.72				96.7	1.07	.01	x_4	150	0	0	0
Costo Total caso no secuencial: 13.5											

TABLA 3

Costo de las cuatro mediciones = .14

Pob	Asignada				Variable			Orden de Entrada			
Real	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
3*a*: Costos iguales											
Π_1	50	0	0	100	1.00	.035	x_1	0	0	3	0
Π_2	0	48	2	96	1.12	.035	x_2	0	0	0	0
Π_3	0	3	47	94	1.06	.035	x_3	0	6	0	0
Costo Total: 10.57				96.7	1.06	.035	x_4	150	0	0	0
3*b*: x_4 'cara'											
Π_1	50	0	0	100	1.00	.02	x_1	0	14	7	0
Π_2	0	48	2	96	1.34	.02	x_2	0	0	0	3
Π_3	0	1	49	98	1.48	.02	x_3	150	0	0	0
Costo Total: 7.84				98	1.27	.08	x_4	0	13	4	0
3*c*: x_4 'barata'											
Π_1	50	0	0	100	1.00	.04	x_1	0	0	3	0
Π_2	0	48	2	96	1.12	.04	x_2	0	0	0	0
Π_3	0	3	47	94	1.06	.04	x_3	0	6	0	0
Costo Total: 8.36				96.7	1.06	.02	x_4	150	0	0	0
Costo Total caso no secuencial: 24											

TABLA 4

Costo de las cuatro mediciones = .28

Pob	Asignada				Variable			Orden de Entrada			
Real	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
4*a*: Costos iguales											
Π_1	50	0	0	100	1.00	.07	x_1	0	0	1	0
Π_2	0	49	1	98	1.04	.07	x_2	0	0	0	0
Π_3	0	5	45	90	1.02	.07	x_3	0	2	0	0
Costo Total: 16.71				96	1.02	.07	x_4	150	0	0	0
4*b*: x_4 'cara'											
Π_1	50	0	0	100	1.00	.04	x_1	0	27	0	0
Π_2	0	48	2	96	1.30	.04	x_2	0	0	0	3
Π_3	0	1	49	98	1.42	.04	x_3	150	0	0	0
Costo Total: 11.64				98	1.24	.16	x_4	0	0	9	0
4*c*: x_4 'barata'											
Π_1	50	0	0	100	1.00	.08	x_1	0	0	1	0
Π_2	0	49	1	98	1.04	.08	x_2	0	0	0	0
Π_3	0	5	45	90	1.02	.08	x_3	0	2	0	0
Costo Total: 12.24				96	1.02	.04	x_4	150	0	0	0
Costo Total caso no secuencial: 45											

TABLA 5

5a: V_1, V_2, V_3

Costo de las cuatro mediciones = 0.00

Pob	Asignada					Variable			Orden de Entrada			
Real	Π_0	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
Π_1	0	44	6	0	88	3.86	0	x_1	0	54	19	72
Π_2	0	3	47	0	94	3.94	0	x_2	0	66	26	48
Π_3	2	0	0	48	96	3.86	0	x_3	150	0	0	0
Costo Total: 11					93	3.89	0	x_4	0	29	101	18

5b: V

Costo de las cuatro mediciones = 0.00

Pob	Asignada					Variable			Orden de Entrada			
Real	Π_0	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
Π_1	1	39	10	0	78	3.94	0	x_1	0	0	26	80
Π_2	0	7	43	0	86	4.00	0	x_2	0	90	25	7
Π_3	2	0	0	48	96	2.58	0	x_3	150	0	0	0
Costo Total: 20					87	3.51	0	x_4	0	29	101	18

TABLA 6

Pob	Asignada					Variable			Orden de Entrada			
Real	Π_0	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
6a: Costo de las cuatro mediciones = 0.00												
Π_1	0	44	6	0	88	3.86	0	x_1	0	54	19	72
Π_2	0	3	47	0	94	3.94	0	x_2	0	66	26	48
Π_3	2	0	0	48	96	3.86	0	x_3	150	0	0	0
Costo Total: 11					93	3.89	0	x_4	0	29	101	18
6b: Costo de las cuatro mediciones = 0.28 Costos iguales												
Π_1	0	44	6	0	88	2.18	.07	x_1	0	10	10	4
Π_2	0	3	47	0	94	2.52	.07	x_2	0	66	1	1
Π_3	0	0	0	50	100	1.02	.07	x_3	150	0	0	0
Costo Total: 29.02					94	1.91	.07	x_4	0	17	26	1
6c: Costo de las cuatro mediciones = 0.28 x_3 'cara'												
Π_1	1	43	6	0	86	2.30	.02	x_1	0	33	12	5
Π_2	0	13	36	1	72	2.38	.02	x_2	0	67	18	0
Π_3	2	1	0	47	94	2.08	.22	x_3	0	35	16	2
Costo Total: 41.36					84	2.25	.02	x_4	150	0	0	0

TABLA 7

Pob	Asignada					Variable			Orden de Entrada			
Real	Π_0	Π_1	Π_2	Π_3	%	Prom	Costo		1	2	3	4
7a: Costo de las cuatro mediciones = 0.00												
Π_1	1	44	4	1	88	3.50	0	x_1	0	0	52	59
Π_2	0	19	31	0	62	3.26	0	x_2	0	36	45	49
Π_3	2	2	0	46	92	3.52	0	x_3	0	96	25	2
Costo Total: 29					81	3.43	0	x_4	150	0	0	0
7b: Costo de las cuatro mediciones = 0.28 Costos iguales												
Π_1	0	40	9	1	80	2.26	.07	x_1	0	0	6	11
Π_2	0	19	30	1	60	2.24	.07	x_2	0	34	9	1
Π_3	2	2	0	46	92	1.64	.07	x_3	0	77	19	0
Costo Total: 55.49					77	2.04	.07	x_4	150	0	0	0
7c: Costo de las cuatro mediciones = 0.28 x_4 'cara'												
Π_1	1	36	13	0	72	2.68	.04	x_1	150	0	0	0
Π_2	0	11	39	0	78	2.62	.04	x_2	0	23	36	7
Π_3	0	0	0	50	100	2.26	.04	x_3	0	106	23	0
Costo Total: 44.08					83	2.52	.16	x_4	0	17	12	4

CONCLUSIONES

Se probó el método propuesto bajo los supuestos de normalidad y matrices de covarianza tanto iguales como diferentes, en casos en que las poblaciones tenían distintos grados de dificultad para diferenciarse. Lo anterior lleva a las siguientes observaciones:

1. En cuanto al método.

- Tomando en cuenta los costos de medición de las variables, este procedimiento lleva a decisiones más rápidas y más baratas que el método discriminante clásico

- Cuando en el método se consideran costos de medición iguales a cero para todas las variables, los resultados obtenidos son comparables a los del Discriminante Lineal o Cuadrático, con matrices de covarianza iguales o distintas, respectivamente.

- Es un procedimiento secuencial que respetando la estructura multivariada de las observaciones, considera de manera adecuada tanto el costo como la capacidad discriminatoria de las componentes de las observaciones.

2. En cuanto a la implantación del método.

- Se tiene el programa funcionando bajo los supuestos explicados en el capítulo 3. La implantación se desarrolló para la gama de computadoras IBM-PC y compatibles, requiriendo una configuración mínima de 128 Kb de memoria y una unidad de diskettes.

- La implantación, sin embargo, tiene algunas restricciones:

  - La precisión de la máquina implicó la necesidad de asignar individuos "lejanos" a las poblaciones a una población hipotética Π_0. Para evitar esto, podría simplemente asignarse a los individuos "lejanos" por medio del método clásico, aunque sería conveniente estudiar la posibilidad de que estos individuos fuesen "aberrantes".

  - Solo se consideraron costos de clasificación equivocada de la forma $C(i|j) = 1, \quad i \neq j$. Esto podría relajarse programando otras funciones de costo.

  - Solo se consideraron funciones de densidad para cada población normales multivariadas con parámetros conocidos.

Como conclusión general, el método propuesto está funcionando con las ventajas y supuestos descritos y está sirviendo de base a futuros desarrollos en esta área. Dichos desarrollos pueden ser el reconsiderar las ideas de Raiffa para el caso de discriminante discreto, tratando de hacer su método factible, o el estudiar el caso

en que los parámetros de las poblaciones no son conocidos, que lleva a la utilización de un enfoque bayesiano.

BIBLIOGRAFIA

Aitchison, J., Habbema, J.D.F., Kay J.W. (1977), A critical comparison of two methods of Statistical Discrimination. App. Stat. 26, No. 1, 15-25.

Anderson, T.W. (1958), An introduction to Multivariate Statistical Analysis. John Wiley & Sons Inc. New York.

Bellman, R. (1957), Dynamic Programming. Princeton Univ. Press, Princeton, New Jersey.

De Groot, M. H. (1970), Optimal Statistical Decisions. Mc. Graw-Hill, New York.

Fu, K. S. (1968), Sequential Methods in pattern recognition and machine learning. Academic Press, London.

Geisser, S. (1964), Posterior odds for multivariate normal classifications. J.R.S.S. B V26, 69-76.

Ghosh, B. K. (1970), Sequential Tests of Statistical Hypotheses. Addison-Wesley Publishing Company.

Goldstein, M., Dillon, W.R. (1978), Discrete Discriminant Analysis. John Wiley & Sons Inc. New York.

Hand, D. J. (1981), Discrimination and classification. John Wiley & Sons Inc. New York.

Hills, M. (1966), Allocation rules and their error rates. J.R.S.S. B V28, 1-31.

O'Reilly, T. F. (1969), Enfoque Decisional de las pruebas de hipótesis. Tesis de Licenciatura. Fac. de Ciencias, UNAM.

Press, S. J. (1972), Applied Multivariate Analysis. Holt, Rinehart & Wilson Inc., Series in Quantitative Methods for Decision-Making.

Raiffa, H. (1961), Statistical Decision Theory Approach to Item Selection for Dichotomous Test and Criterion Variables. Studies in item analysis and prediction. Stanford Univ. Press.

Seber, G.A.F. (1984), Multivariate observations. John Wiley & Sons Inc. New York.

Wald, A., Wolfowitz, J. (1948), Optimum character of the sequential probability ratio test. Ann. Math. Statist. 19, 326-339.

Wald, A. (1950), Statistical Decision Functions. Chelsea Publishing Company. Bronx, New York.

Wetherill, G. B. (1966), Sequential Methods in Statistics. Methuen & Co. LTD, London.