MODELOS DE CLASIFICACION REGULARES

por
JOSE-DOMINGO BERMUDEZ EDO

Tesis
presentada para optar al grado de
Doctor en Matemáticas
por la
Universidad de Valencia

Departamento de Bioestadística
Universidad de Valencia
Mayo 1984

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

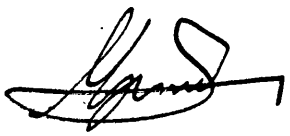
CERTIFICADO

JOSE-MIGUEL BERNARDO HERRANZ, Catedrático de Bioestadística, y Director del Departamento de Bioestadística de la Universidad de Valencia.

CERTIFICA: Que la presente memoria *Modelos de Clasificación Regulares* ha sido realizada bajo mi dirección, en el Departamento de Bioestadística de la Universidad de Valencia, por el Sr. D. José-Domingo Bermúdez Edo, y constituye su tesis para optar al grado de Doctor en Matemáticas.

Y para que conste, en cumplimiento de la legislación vigente, presento ante la Facultad de Matemáticas de la Universidad de Valencia, a de Mayo de 1984.

EL DIRECTOR



J.M. Bernardo Herranz

RESUMEN

En esta memoria se introduce el concepto de Modelo de Clasificación Regular, caracterizándolo a través de un conjunto de propiedades básicas, y se investiga el comportamiento de las funciones de verosimilitud obtenidas con este tipo de modelos. A pesar de que estas funciones de verosimilitud se comportan de una manera muy poco usual, se comprueba que la aproximación asintótica Normal a la distribución final sigue siendo válida.

A continuación se estudia el modelo inferencial Bayesiano para los Modelos de Clasificación Regulares haciendo especial incapié en la búsqueda de las distribuciones de referencia. Los resultados teóricos obtenidos son dificilmente tratables desde un punto de vista computacional. Por tanto se hace necesaria la obtención de aproximaciones que, siendo fáciles de calcular, estén suficientemente cerca de los resultados teóricos. Las aproximaciones propuestas, así como los resultados teóricos cuando son computacionalmente tratables, se ejemplifican utilizando diversos bancos de datos; unos simulados y otros reales.

Como un resultado adicional importante se obtiene una debilitación de los axiomas de regularidad propuestos por Walker(1969) para la normalidad asintótica de la distribución final. También se estudian, como ejemplos de resultados teóricos mas generales, los procesos inferenciales Bayesianos para los modelos Logístico Aditivo, Logístico Multiplicativo y Normal Acumulado.

Palabras Clave: CLASIFICACION PREDICTIVA; DISTRIBUCIONES DE REFERENCIA; INFERENCIA BAYESIANA; MODELOS CLASIFICACION REGULARES; MODELOS CLASIF. TIPO 1; MODELOS CLASIF. TIPO 2; MODELO LOGISTICO; MODELO NORMAL ACUMULADO; NORMALIDAD ASINTOTICA; PARADIGMA DIAGNOSTICO; PREDICCION.

Clasificación AMS (1980):
Primaria 62F15, 62H30 *Secundaria* 62E20, 62E25, 62P10

Clasificación UNESCO: 1209.13, 1209.14

CONTENIDO

PREFACIO

Durante los últimos años, mas concretamente a partir de 1972, existe una clara tendencia en la investigación estadística hacia la sustitución del modelo Normal por el modelo Logístico, o modelos similares, para el estudio de datos de clasificación. Esta tendencia ha sido ampliamente potenciada, dentro de la metodología Clásica, por los trabajos del Prof. J.A. Anderson. Sin embargo, desde una prespectiva Bayesiana, prácticamente los únicos aunque importantes, trabajos publicados hasta el momento se deben al Prof. J. Aitchison y están basados en una aproximación asintótica solo justificada de forma heurística.

El principal motivo que me impulsó a comenzar la investigación presentada en esta memoria fue la impresión, obtenida como consecuencia de los estudios realizados en mi Tesis de Licenciatura, de la existencia de un importante vacío en la investigación actual sobre Clasificación Estadística desde la metodología Bayesiana. Vacio que es necesario cubrir para obtener buenos resultados en ese importante campo de aplicaciones.

Esta tesis ha sido realizada bajo la dirección del Profesor J.M. Bernardo. Quiero expresarle desde aquí mi agradecimiento por su continuo estímulo y por los interesantes comentarios realizados a lo largo de todo el trabajo que han influído de manera importante en la orientación y realización de esta memoria. Así mismo, he de agradecer a los restantes miembros del Departamento de Bioestadística de la Universidad de Valencia, la ayuda prestada mediante sus comentarios y su compañerismo.

CAPITULO 1

PRELIMINARES

1.1 INTRODUCCION

La *Clasificación Estadística* engloba a todos aquellos métodos que permiten relacionar, *en presencia de incertidumbre*, a un objeto o *individuo* con un conjunto finito de categorías, Δ, previamente especificado. Sin pérdida de generalidad, debido a su carácter finito, el conjunto Δ puede considerarse integrado por k *clases* o categorías *exclusivas y exhaustivas*, $\Delta \equiv \{\delta : \delta = 1, \ldots, k\}$. En efecto, si Δ no fuera una partición entonces se consideraría como conjunto de las clases a la partición maximal no trivial, con respecto a la relación de inclusión, incluída en el álgebra generada por el conjunto Δ.

Todo individuo de la población puede ser codificado mediante un conjunto de *indicadores* que lo definen. El vector aleatorio correspondiente al conjunto de indicadores de un individuo, x, se denomina *vector representante*; en el se resume toda la información conocida sobre el individuo. En general, se dispone de un *banco de datos*, $D \equiv \{(x_i, \delta_i),\ i = 1, \ldots, n\}$, donde se recogen los

vectores representantes de n individuos cuya clasificación correcta es conocida.

Debido a la presencia de incertidumbre, la respuesta completa a todo problema de Clasificación Estadística debe consistir en una distribución de probabilidades, $p(\delta|x,D)$. Esta distribución, denominada *distribución de clasificación*, describe las probabilidades de que un individuo con vector representante x, pertenezca a cada una de las k poblaciones, una vez conocida la información proporcionada por el banco de datos D.

En numerosas aplicaciones, esta clasificación probabilística solo representa un paso intermedio en el planteamiento general. Así por ejemplo, la *diagnosis clínica*, clasificación entre diversas enfermedades alternativas, no es un objetivo en si misma sino una etapa previa a la *elección de tratamiento*. La distribución de clasificación, llamada *distribución diagnóstica* en el contexto médico, proporciona las probabilidades necesarias para solucionar el problema de decisión a que da lugar la elección de tratamiento, (Bermúdez, 1981).

Los diferentes métodos estadísticos utilizados en el cálculo de la distribución de clasificación, $p(\delta|x,D)$, pueden catalogarse en dos grandes grupos: unos centran su mayor interés en la modelización de la población de vectores representantes en cada una de las clases, utilizando para ello modelos probabilísticos de la forma $p(x|\delta,\theta)$; los otros estudian direc-

tamente la distribución de clasificación, proponiendo modelos probabilísticos de la forma $p(\delta|x,\theta)$. A estos dos planteamientos se les ha denominado, (Dawid, 1976), *paradigma de muestreo* y *paradigma diagnóstico*, respectivamente.

Sin embargo, el término paradigma no parece muy apropiado en este contexto. En efecto, siguiendo a Kuhn (1970, segunda ed. ampliada), las dos acepciones del término paradigma son : 'Por una parte, el conjunto de todas las ideas, valores y técnicas compartidas por los miembros de una comunidad científica dada. Por otra parte, cada uno de los elementos del conjunto anterior que, empleado como ejemplo o modelo, pueda reemplazar a reglas explícitas...'. Ninguna de estas dos acepciones corresponde a la idea que se pretende transmitir con las expresiones paradigma de muestreo y paradigma diagnóstico. Una alternativa mas realista es *enfoque muestral* y *enfoque clasificatorio*.

El objetivo de esta tesis es la aplicación de la metodología Bayesiana a la Clasificación Estadística, desde un enfoque clasificatorio. Aunque el énfasis se situa en el estudio y desarrollo teórico de los modelos propuestos, no se ha descuidado la obtención de resultados que hagan posible su aplicación en problemas prácticos concretos.

Esta memoria está dividida en seis capítulos y dos apéndices. Cada capítulo, dividido a su vez en varios apartados, comienza con un pequeño resumen de su contenido. El primer capítulo, del que forma parte esta

introducción, termina con un segundo apartado en el que se discuten las diferencias mas importantes entre los dos enfoques antes mencionados, y se introducen algunos conceptos básicos que forman el contexto en el que se enmarca esta memoria.

En el capítulo dos se definen los *Modelos de Clasificación Regulares*, concepto que engloba y generaliza a todos los modelos utilizados hasta el momento desde el enfoque clasificatorio. En el se estudian las propiedades mas importantes de las funciones de verosimilitud proporcionadas por esos modelos.

El capítulo tres trata sobre las aproximaciones asintóticas a la distribución final obtenida a partir de un modelo de clasificación regular. En el se demuestra que, bajo condiciones muy generales, la aproximación Normal asintótica usual es correcta.

El proceso inferencial completo para este tipo de modelos forma el contenido del capítulo cuatro. Se hace especial incapié en la obtención de distribuciones de referencia, y se proponen aproximaciones a los resultados teóricos que hagan viable su utilización rutinaria.

En el capítulo cinco se comparan, mediante ejemplos numéricos, los métodos propuestos en el capítulo cuatro con los métodos alternativos mas importantes.

Por último, en el capítulo seis se valoran y discuten los resultados obtenidos en la tesis, y se comentan algunas áreas de investigación futura.

Esta memoria se completa con dos apéndices. El primero contiene una generalización del resultado de Walker (1969) sobre el comportamiento asintótico de las distribuciones finales, generalización utilizada en el capítulo tres. El segundo recoge los bancos de datos de los ejemplos estudiados en el capítulo cinco y los datos simulados utilizados en el capítulo tres.

A lo largo de todo el trabajo se ha procurado simplificar al máximo la notación, acompañando su introducción con un comentario literario aclaratorio. Las letras latinas se han empleado para representar tanto las cantidades aleatorias muestrales como sus valores observados, mayúsculas en un caso y minúsculas en el otro, reservando las letras griegas para la representación de los parámetros que identifican los modelos. Todas las densidades, al igual que las funciones de probabilidad, se han representado con la misma letra p seguida, entre parentesis, por el nombre de la cantidad aleatoria a la que se refiere. El operador esperanza se ha representado mediante la letra E subindicada, allí donde fuera necesario, por las letras definitorias de los vectores aleatorios correspondientes. El operador varianza se ha representado por VAR, reservando la notación D^2 para representar la matriz de derivadas segundas parciales. Así, $D_t^2(f(t_0))$ representa a la matriz cuyo elemento genérico es

$$\frac{\partial^2}{\partial t_i \partial t_j} f(t)$$

calculada en el punto t_0. De forma análoga, $D_t^1(f(t_0))$ representa la función gradiente correspondiente a la

función vectorial $f(t)$,calculada en el punto t_0.

Para numerar las ecuaciones, definiciones, proposiciones, teoremas y ejemplos se ha utilizado notación decimal. Así, la definición 2.1.1 es la primera definición del apartado 1 del capítulo 2. Del mismo modo, la proposición 3.2.1 es la proposición 1 del apartado 2 del capítulo 3.

1.2 CONCEPTOS PREVIOS

La obtención de la distribución de clasificación, $p(\delta|x,D)$, mediante los enfoques muestral y clasificatorio conlleva hipótesis de independencia diferentes. Así, los métodos propuestos desde un enfoque muestral asumen, explícita o implícitamente, la hipótesis

HM: La familia de modelos probabilísticos que describen la generación de los pares clase-vector representante, (δ,x), viene identificada por dos vectores paramétricos de dimensionalidad finita, $\theta\in\Theta$ y $\psi\in\Psi$, de forma que

$$p(\delta,x|\theta,\psi) = p(x|\delta,\theta,\psi)\ p(\delta|\theta,\psi)$$

con

$$p(x|\delta,\theta,\psi) = p(x|\delta,\theta)$$
$$p(\delta|\theta,\psi) = p(\delta|\psi)$$

Por el contrario, los métodos propuestos desde un enfoque clasificatorio asumen la hipótesis alternativa:

HC: La familia de modelos probabilísticos que describen la generación de los pares clase-vector representante, (δ,x), viene identificada por dos vectores paramétricos de dimensionalidad finita, $\theta\in\Theta$ y $\psi\in\Psi$, de forma que

$$p(\delta,x|\theta,\psi) = p(\delta|x,\theta,\psi)\ p(x|\theta,\psi)$$

con

$$p(\delta|x,\theta,\psi) = p(\delta|x,\theta)$$
$$p(x|\delta,\psi) = p(x|\psi)$$

El enfoque muestral parece ser el mas adecuado en aquellas situaciones en las que se supone una relación causa-efecto entre la clase y el vector representante de cada individuo. Es entonces cuando el modelo

$p(x|\delta,\theta)$ aparece de forma natural. Similarmente, el enfoque clasificatorio será plenamente aconsejable si el vector representante se considera causa de la clase.

Sin embargo, no siempre está clara la relación causa-efecto, o viceversa, entre x y δ. En numerosas aplicaciones el vector representante puede dividirse en dos subvectores, $x=(r,s)$, de forma que r representa a indicadores que pueden ser considerados causantes de la clase δ, mientras que s representa a indicadores causados por δ.

Bernardo (1978), recogiendo una idea formulada por Dawid (1976), propone un compromiso entre los dos enfoques. Así, si los vectores r y s se consideran condicionalmente independientes, la distribución de clasificación puede hallarse mediante el Teorema de Bayes

$$p(\delta|x,D) \propto p(s|\delta,D)\ p(\delta|r,D)$$

situando los cálculos para la obtención de $p(s|\delta,D)$ y $p(\delta|r,D)$ en los enfoques muestral y clasificatorio respectivamente. Bermúdez (1979) generaliza este desarrollo, evitando la hipótesis de independencia entre r y s, pero al precio de necesitar bancos de datos muy grandes.

A pesar de las ventajas apuntadas por Dawid (1976) del enfoque clasificatorio sobre el muestral, no es de esperar que uno de ellos prevalezca sobre el otro. Así, han de ser las condiciones particulares de cada problema concreto las que sugieran el enfoque correcto, (Bermúdez, 1984).

En el contexto de Clasificación Estadística se ha reproducido la controversia existente entre la escuela Clásica y la escuela Bayesiana, dando lugar respectivamente, a las clasificaciones *Estimativa* y *Predictiva* (Aitchison, Habbema and Kay, 1977).

Desde el enfoque muestral, la solución estimativa (Welch, 1939; Wald, 1944; Anderson, 1958) generaliza las técnicas de Análisis Discriminante introducidas mediante argumentos geométricos por Fisher (1936). Utilizando el banco de datos D, proponen obtener un estimador, $\hat{\theta}(D)$, del parámetro desconocido y entonces aproximar $p(x|\delta,D)$ mediante $p(x|\delta,\hat{\theta}(D))$. La distribución de clasificación se obtiene vía Teorema de Bayes como $p(\delta|x,D) \propto \hat{p}(\delta)\ p(x|\delta,\hat{\theta}(D))$. Donde $\hat{p}(\delta)$ es un estimador de la probabilidad $p(\delta)$ obtenido con el mismo banco de datos D, o con un banco alternativo.

Por el contrario, la solución predictiva desde el enfoque muestral (Geisser, 1964; Dunsmore, 1966; Aitchison and Dunsmore, 1975) calcula la distribución predictiva de los vectores representantes en cada una de las clases,

$$p(\theta|D) \propto p(\theta) \prod_{i=1}^{n} p(x_i|\delta_i,\theta)$$

$$p(x|\delta,D) = \int p(x|\delta,\theta)\ p(\theta|D)\ d\theta$$

para obtener la distribución de clasificación mediante una nueva aplicación del Teorema de Bayes:

$$p(\delta|x,D) \propto p(x|\delta,D)\ p(\delta)$$

En cualquier caso, el enfoque muestral necesita la especificación de un modelo multivariante, $p(x|\delta,\theta)$,

que sea realista y que presente soluciones tratables. La única familia multivariante con la que se han obtenido resultados operativos es la familia Normal. Sin embargo, el caracter discreto que generalmente poseen muchas de las componentes del vector representante hacen que el modelo Normal no sea un modelo realista. Como una alternativa a la normalidad multivariante es de destacar el método propuesto por Bernardo (1983).

Desde el enfoque clasificatorio, la solución estimativa (Walker and Duncan, 1967; Cox, 1970; Anderson, 1972) consiste en aproximar $p(\delta|x,D)$ mediante $p(\delta|x,\hat{\theta}(D))$; siendo $\hat{\theta}(D)$ un estimador del parámetro desconocido, θ, obtenido a partir del banco de datos ,D. La solución predictiva (Teather, 1974; Aitchison and Lauder, 1979) se obtiene de nuevo mediante el uso sucesivo de los teoremas de Bayes y de la Probabilidad Total,

$$p(\theta|D) \propto p(\theta) \prod_{i=1}^{n} p(\delta_i|x_i,\theta)$$

$$p(\delta|x,D) = \int p(\delta|x,\theta)\ p(\theta|D)\ d\theta$$

El *modelo Logístico* ha sido el mas utilizado para describir la distribución $p(\delta|x,\theta)$. Este modelo se caracteriza por la linealidad de los logaritmos de los cocientes de las probabilidades, esto es:

$$\text{Log}\ \frac{p(\delta=i|x,\theta)}{p(\delta=k|x,\theta)} = \theta_{i0}+\theta_{i1}x_1+\ldots+\theta_{im}x_m \quad ,i=1,\ldots,k-1$$

lo que da lugar a,

$$p(\delta=i|x,\theta) = \exp(\theta_{i0}+\theta_{i1}x_1+\ldots+\theta_{im}x_m)\ p(\delta=k|x,\theta) \quad (i=1,\ldots,k-1)$$

$$p(\delta=k|x,\theta) = 1/\left(1+\sum_{i=1}^{k-1}\exp(\theta_{i0}+\theta_{i1}x_1+\ldots+\theta_{im}x_m)\right) \qquad (1)$$

Un modelo alternativo utilizado cuando solo existen dos clases, $k=2$, es el *modelo Normal Acumulado* definido como

$$p(\delta=1|x,\theta) = 1-p(\delta=2|x,\theta) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\theta'x} \exp(-y^2/2)\, dy \qquad (2)$$

con $\theta'x = \theta_1x_1+\theta_2x_2+\ldots+\theta_mx_m$

Para una discusión mas detallada sobre la diferencia entre los dos enfoques puede consultarse Bermúdez (1982) y referencias allí citadas.

Los desarrollos obtenidos anteriormente están plenamente justificados si el banco de datos, $D\equiv\{(x_i,\delta_i), i=1,\ldots,n\}$, constituye una muestra aleatoria de la población global (*Muestreo Prospectivo*). En tal caso, la función de verosimilitud es

$$V(\theta,\psi|D) \propto \prod_{i=1}^{n} p(x_i,\delta_i|\theta,\psi),$$

y por tanto, aplicando las hipótesis HM o HC según se adopte el enfoque muestral o el clasificatorio, se obtiene

$$V(\theta|D) \propto \prod_{i=1}^{n} p(x_i|\delta_i,\theta) \quad \delta \quad V(\theta|D) \propto \prod_{i=1}^{n} p(\delta_i|x_i,\theta)$$

respectivamente.

Sin embargo, en muchos problemas concretos la incidencia de las distintas clases en la población global, $p(\delta)$, es muy distinta. En tales situaciones, mediante un muestreo prospectivo sería necesario observar muchos individuos de una de las clases, δ_1, para encontrar unos pocos de una segunda clase, δ_2. Por ello es muy frecuente que el banco de datos, $D\equiv\{ x_{i\delta};\ i=1,\ldots n_\delta,$

$\delta=1,\ldots,k\}$, esté formado por n_δ individuos, $\delta=1,\ldots,k$, seleccionados aleatoriamente entre los que pertenecen a la clase δ (*Muestreo Retrospectivo*). La función de verosimilitud definida por este tipo de muestreo es:

$$V(\theta,\psi|D) \propto \prod_{\delta=1}^{k} \prod_{i=1}^{n_\delta} p(x_{i\delta}|\delta,\theta,\psi)$$

Debido a la hipótesis HM, la función de verosimilitud de θ, en el enfoque muestral es

$$V(\theta|D) \propto \prod_{\delta=1}^{k} \prod_{i=1}^{n_\delta} p(x_{i\delta}|\delta,\theta),$$

que coincide con la obtenida mediante un muestreo prospectivo. La situación se complica al adoptar un enfoque clasificatorio, para el que los datos retrospectivos pueden representar problemas graves. En efecto, por el Teorema de Bayes:

$$p(x_{i\delta}|\delta,\theta,\psi) = \frac{p(\delta|x_{i\delta},\theta,\psi)\ p(x_{i\delta}|\theta,\psi)}{p(\delta|\theta,\psi)} \qquad (3)$$

donde $$p(\delta|\theta,\psi) = \int p(\delta|x_{i\delta},\theta,\psi)\ p(x_{i\delta}|\theta,\psi)\ dx_{i\delta}$$

Por tanto,

$$V(\theta,\psi|D) \propto \prod_{\delta=1}^{k} \left[p(\delta|\theta,\psi)\right]^{-n_\delta} \prod_{i=1}^{n_\delta} p(\delta|x_{i\delta},\theta)\ p(x_i|\psi)$$

Con la hipótesis adicional $p(\delta|\theta,\psi) = p(\delta|\psi)$, la función de verosimilitud de θ coincide con la obtenida mediante un muestreo retrospectivo. Sin embargo, esta nueva hipótesis es, en general, contradictoria con HC.

Los problemas ocasionados por los muestreos retrospectivos son debidos a que estos muestreos no proporcionan información sobre la proporción de cada una de las clases en la población global. Si esta población fuese conocida, quizás a través de un banco de datos alternativo, la expresión (3) podría escribirse como,

$$p(x_{i\delta}|\delta,\theta,\psi) \propto p(\delta|x_{i\delta},\theta,\psi)\ p(x_{i\delta}|\theta,\psi)$$

con lo que, al igual que en los muestreos prospectivos,

$$V(\theta|D) \propto \prod_{\delta=1}^{k} \prod_{i=1}^{n_\delta} p(x_{i\delta}|\delta,\theta).$$

Sin embargo, suponer conocida la distribución $p(\delta)$ implica ciertas restricciones sobre los vectores paramétricos θ y ψ, ($\int p(\delta,x|\theta,\psi)\ dx$, conocida). Posiblemente esta dificultad pueda ser superada introduciendo esas restricciones en la distribución inicial $p(\theta)$.

Un muestreo especial que presenta una solución mas atractiva es el *muestreo retrospectivo sintético* (Mantel, 1973). En lugar de fijar inicialmente los tamaños muestrales n_δ, $\delta=1,\ldots,k$, como hacen los muestreos retrospectivos, este tipo de muestreo supone la existencia de un hipotético banco de datos prospectivo D_p. De los datos pertenecientes a la clase δ que integran el banco D_p, solo se selecciona un porcentaje, q_δ, fijado con antelación. La distribución de las clases en el banco de datos así obtenido, D, es:

$$\pi(\delta|\theta,\psi) = C(\delta)\ p(\delta|\theta,\psi)$$
$$C(\delta) = q_\delta / \sum_{\delta=1}^{k} q_\delta\ p(\delta|\theta,\psi) \qquad \delta=1,\ldots,k \qquad (4)$$

donde $p(\delta|\theta,\psi)$ es la distribución en la población glo-

bal. De igual forma, la distribución de cada dato en el banco de datos D es:

$$\pi(x,\delta|\theta,\psi) = p(x|\delta,\theta,\psi)\ \pi(\delta|\theta,\psi)$$

Utilizando HC y la expresión (4),

$$\begin{aligned}\pi(x,\delta|\theta,\psi) &= C(\delta)\ p(\delta|\theta,\psi)\ p(x|\delta,\theta,\psi)\\ &= C(\delta)\ p(\delta,x|\theta,\psi)\\ &= C(\delta)\ p(\delta|x,\theta)\ p(x|\psi)\end{aligned}$$

Si $p(\delta|x,\theta)$ sigue el modelo logístico definido mediante la expresión (1):

$$\text{Log}\ \frac{\pi(x,\delta=i|\theta,\psi)}{\pi(x,\delta=k|\theta,\psi)} = \text{Log}\ \frac{C(i)}{C(k)} + \text{Log}\ \frac{p(\delta=i|x,\theta)}{p(\delta=k|x,\theta)}$$

$$= \text{Log}\ \frac{q_i}{q_k} + \theta_{i_0} + \theta_{i_1}x_1 + \ldots + \theta_{im}x_m$$

Por tanto, la función de verosimilitud proporcionada por el muestreo retrospectivo sintético es:

$$V(\theta|D) \propto \prod_{\delta=1}^{k}\ \prod_{i=1}^{n_\delta} \pi(\delta|x_{i\delta},\theta)$$

donde, si $p(\delta|x,\theta)$ sigue un modelo logístico, $\pi(\delta|x,\theta)$ tambien sigue un modelo logístico de parámetro θ^*,

$$\theta_i^* = \theta_i + \text{Log}\ \frac{q_i}{q_k} \qquad , i=1,\ldots,k-1$$

$$\theta_{ij}^* = \theta_{ij} \qquad , i=1,\ldots,k-1\ ,\quad j=1,\ldots,m$$

CAPITULO 2

MODELOS DE CLASIFICACION REGULARES

En este capítulo se introduce el concepto de Modelo de Clasificación Regular, justificando su definición y demostrando las propiedades básicas a las que esta da lugar. Se definen los modelos de Tipo 1 y Tipo 2, demostrando que son modelos de clasificación regulares, y se proponen diversos ejemplos concretos. Por último, se estudia el comportamiento asintótico de la función de verosimilitud generada por este tipo de modelos.

2.1 DEFINICION Y PRIMERAS PROPIEDADES

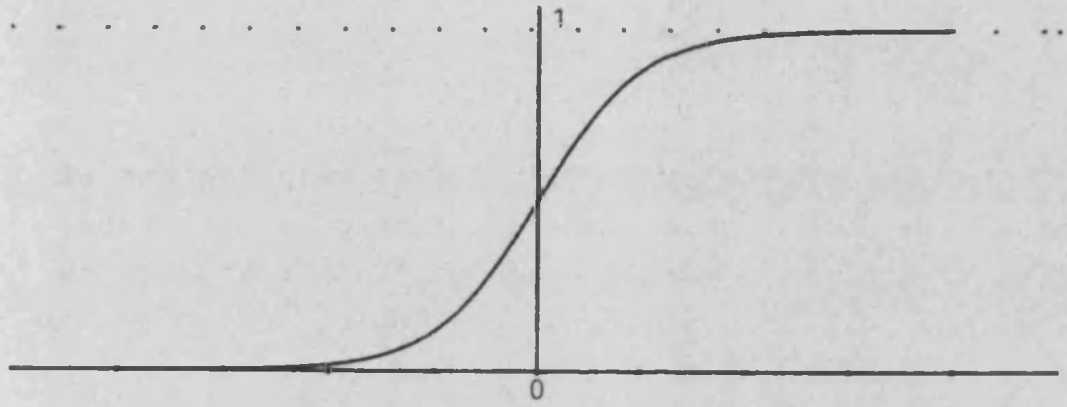
Al modelizar la función de probabilidades $p(\delta|x,\theta)$, parece conveniente dotarla de la suficiente flexibilidad para que, considerada como función de θ e independientemente del valor de x, el rango de valores que pueda tomar coincida con el intervalo abierto (0,1). Con este fin, $p(\delta|x,\theta)$ puede definirse en términos de una función, $F_\delta(.)$, monótona en sus componentes y que tenga como asíntotas horizontales las rectas $y=0$ e $y=1$. Esto es, $p(\delta|x,\theta)=F_\delta(t)$ donde t es una función de x y θ, $t=t(x,\theta)$.

Tanto el modelo logístico, definido por la expresión 1.2.1, como el normal-acumulado, expresión 1.2.2, responden a esta estructura. En efecto, el modelo logístico para dos clases se define como:

$$p(\delta=1|x,\theta) = 1- p(\delta=2|x,\theta) = F_1(t(x,\theta))$$

con $t(x,\theta)= \theta'x = \theta_1x_1+\ldots+\theta_mx_m$, siendo $F_1(t) = e^t/(1+e^t)$. En este caso, $F_1(t)$ (ver fig. 1) es precisamente la función de distribución de una cantidad aleatoria continua cuya función de densidad es $p(t) = e^t/(1+e^t)^2$, i.e. la densidad correspondiente a la distribución logística (Johnson and Kotz; 1970, vol. 2, cap. 22).

El modelo normal-acumulado tiene la misma estructura pero utiliza como función de soporte la función de distribución Normal de media 0 y varianza 1.



- fig. 1 -
Función de distribución Logística

Entre las componentes del vector representante, x, es posible incluir cualquier potencia o producto cruzado de los indicadores e incluso un término constante. Por tanto, Teorema de Taylor, puede suponerse con suficiente generalidad que la función $t(x,\theta)$ es lineal, o por lo menos aproximadamente lineal, con respecto al vector m-dimensional x, con coeficientes θ_i, $i=1,\ldots,m$ independiente de x.

En el caso general, cuando existen k clases, la función de probabilidades $p(\delta|x,\theta)$, $\delta=1,\ldots,k$, está determinada por k números positivos sujetos a la restricción de suma uno, por tanto $p(\delta|x,\theta)$ posee $k-1$ indeterminaciones o *grados de libertad*. Parece lógico introducir en el modelo paramétrico $k-1$ *informaciones* distintas. Una forma de conseguir este objetivo es exigir que la función $t(x,\theta)$ proporcione $k-1$ combinaciones lineales, i.e.

$$t(x,.) : \Theta \equiv R^{mx(k-1)} \rightarrow R^{k-1}, \quad t(x,\theta) = \theta' x$$

donde θ es una matriz $mx(k-1)$ de manera que su columna i-ésima proporciona los coeficientes de la i-ésima combinación lineal de x. θ' es la matriz transpuesta de θ.

De nuevo, esta característica es cumplida por el modelo logístico. Unico modelo paramétrico utilizado hasta el momento, desde un enfoque clasificatorio, en problemas de clasificación en los que el número de clases es superior a dos.

Todas estas consideraciones sugieren la siguiente

DEFINICION 1

El modelo $p(\delta|x,\theta)$ *es un modelo de Clasificación Regular si y solo si* $p(\delta=i|x,\theta) = F_i(t(x,\theta))$. *Siendo* $t(x,\theta)$ *un conjunto de* k *combinaciones lineales,* $t(x,\theta)=\theta'x$ *con* $\theta\in R^{m\times(k-1)}$ *y* $x\in R^m$, *y exigiendo a las* k *funciones* $F_i(.)$ *que cumplan las propiedades:*

P1. $F_i(.)$, $i=1,\ldots,k$, *es la iésima componente de una función vectorial uno a uno, definida de* R^{k-1} *en el simplex* $S^k\equiv\{(a_1,\ldots,a_k)\in R^k: a_i>0, a_1+\ldots+a_k=1\}$.

P2. Regularidad. *Las* k *funciones* $F_i(t)$ *son dos veces continuamente diferenciables. Además, la matriz* $G(t)$ *con elemento genérico* $G_{ij}(t)$ *dado por*

$$G_{ij}(t) = \sum_{\delta=1}^{k} F_\delta(t)\left[\frac{\partial}{\partial t_i} Log(F_\delta(t))\,\frac{\partial}{\partial t_j} Log(F_\delta(t))\right]$$

es definida positiva para todo $t\in R^{k-1}$.

P3. Monotonicidad. *La función* $F_k(t)$ *es monótona decreciente en cada una de sus componentes. La función* $F_i(t)$, $\forall i=1,\ldots,k-1$, *es monótona creciente como función de la componente i-ésima, siendo monótona decreciente con respecto a las*

demás componentes. Además,

P3.1. $\forall \varepsilon>0 \ \exists \rho_{i1}(\varepsilon)>0: F_i(t)<\varepsilon \ \forall t\in R^{k-1}$ *con* $t_i<-\rho_{i1}$

esto es, el límite cuando $t_i\to-\infty$ *de* $F_i(t)$ *es 0.*

P3.2. $\forall \varepsilon>0 \ \exists \rho_{i2}(\varepsilon)>0: F_i(t)>1-\varepsilon \ \forall t\in R^{k-1}$ *con*

$t_i>\rho_{i2}$ *y con* $t_j<-\rho_{i2}$ $j\neq i$. *Esto es, el*

límite cuando $t_i\to+\infty$ *y* $t_j\to-\infty$ *de* $F_i(t)$ *es 1.*

La propiedad P1 es estrictamente necesaria para que el modelo de clasificación regular esté bien definido. Esto es, para garantizar que, una vez fijado el parámetro θ, y el vector x, $p(\delta|x,\theta)$ sea una verdadera función de probabilidad.

La propiedad P2 representa condiciones matemáticas deseables que aseguran un comportamiento *suficientemente regular* de la función de verosimilitud.

La propiedad de Monotonicidad, P3, es realmente la propiedad definitoria de los modelos de clasificación regular. Como consecuencia de esta propiedad, dado cualquier vector representante, x, siempre existe un valor del parámetro que hace *prácticamente imposible* la clase i-ésima, $i<k$, (P3.1). Por el contrario, (P3.2), existe otro valor del parámetro que la hace *prácticamente segura*. Que esos valores se alcancen de forma asintótica garantiza que el recorrido de la función *continua* F_i es todo el intervalo abierto (0,1). Un hecho intuitivamente deseable.

Por otra parte, la monotonicidad de P3 imprime una relación muy estrecha entre cada una de las componentes del vector $t \in R^{k-1}$ y la clase correspondiente. Así, un incremento en t_i implica un aumento de la probabilidad de la clase $\delta=i$, mientras que las demás probabilidades o disminuyen o permanecen inalteradas.

Aunque no aparece de forma explícita en la definición, el comportamiento de la función F_k es similar al exigido por la propiedad P3 a las funciones F_i, $i<k$. En efecto,

PROPOSICION 1

Si F es una función vectorial que cumple P1 y P3, entonces el límite de $F_k(t)$ *cuando,* $\forall$ i<k, t_i *tiende a* $-\infty$ *es 1*

DEMOSTRACION

Dado $\varepsilon>0$ sea $\rho_{k1}(\varepsilon)$ el máximo de los $k-1$ números $\rho_{i1}(\varepsilon/(k-1))$ definidos al igual que en P3.1

Sea $t \in R^{k-1}$ tal que $t_i<-\rho_{k1}(\varepsilon)$ $\forall$ $i=1,\ldots,k-1$. Entonces $t_i<-\rho_{i1}(\varepsilon/(k-1))$, y por tanto (P3.1), $F_i(t)<\varepsilon/(k-1)$ $\forall$ $i=1,\ldots,k-1$. Esto es,

$$\sum_{i=1}^{k-1} F_i(t) < \sum_{i=1}^{k-1} \varepsilon/(k-1) = \varepsilon$$

de donde se deduce,

$$F_k(t) = 1 - \sum_{i=1}^{k-1} F_i(t) > 1-\varepsilon$$

■

PROPOSICION 2

Si F es una función vectorial que cumple P1 y P3 entonces $F_k(t) \to 0$ *cuando* $t_{i_0} \to +\infty$ *para algún* $i_0 = 1,\ldots,k-1$

DEMOSTRACION

Sin pérdida de generalidad supóngase $i_0 = 1$.

Dado $\varepsilon > 0$, sea $\rho_{k2}(\varepsilon)$ el máximo de los $k-1$ números $\rho_{i2}(\varepsilon)$, definidos en P3.2. Sea t tal que $t_1 > \rho_{k2}(\varepsilon) \geq \rho_{12}(\varepsilon)$ y sea $t^* = \text{Max}(|t_i|,\ i=1,\ldots,k-1)$. Por tanto, $t^* \geq t_1 > \rho_{k2}(\varepsilon)$.

Como $t_i > -t^*$, $i=1,\ldots,k-1$, por la monotonicidad de F_k,

$$F_k(t) = F_k(t_1,\ldots,t_{k-1}) \leq F_k(t_1,-t^*,\ldots,-t^*) \leq 1 - F_1(t_1,-t^*\ldots,-t^*)$$

pero $t_1 > \rho_{k2}(\varepsilon) \geq \rho_{12}(\varepsilon)$ mientras que $-t^* < -\rho_{k2}(\varepsilon) \leq -\rho_{12}(\varepsilon)$. Aplicando P3.2, $F_1(t_1,-t^*,\ldots,-t^*) > 1-\varepsilon$, de donde se deduce $F_k(t) \leq 1 - F_1(t_1,-t^*,\ldots,-t^*) < \varepsilon$. ∎

En el caso particular en el que existan solamente dos clases, $k=2$, las tres propiedades se reducen a exigir que $F_1(t) + F_2(t) = 1$ para todo t, y que $F_1(t)$ sea una función de distribución *estrictamente creciente* y dos veces continuamente diferenciables.

Exigir que $F_1(t)$ sea estrictamente creciente es necesario para garantizar la segunda parte de P2. Si $F_1(t)$ fuera constante en cierto intervalo, su derivada sería nula en dicho intervalo, con lo que $G(t) = 0$, en contradicción con P2.

2.2 ALGUNOS MODELOS CONCRETOS

Los modelos de clasificación regulares para dos clases, $k=2$, están íntimamente relacionados con los modelos utilizados en regresión dosis-respuesta con respuesta cuantal (Finney, 1978; cap. 17). Esto es así debido a que en ambos estudios se modeliza la verosimilitud de los datos a través de una función de distribución. De hecho, todos los modelos de clasificación, $p(\delta|x,\theta)$, propuestos hasta el momento, así como sus métodos de estimación, han sido heredados de los estudios sobre regresión dosis-respuesta.

La estructura general de los modelos de clasificación regulares para dos clases es:

$$p(\delta=1|x,\theta) = F(t(x,\theta)), \qquad p(\delta=2|x,\theta) = 1-F(t(x,\theta))$$

siendo $t(x,\theta) = \theta_1 x_1 + \ldots + \theta_m x_m$, y siendo $F(t)$ una función de distribución estrictamente creciente y dos veces continuamente diferenciable.

Algunas de las funciones $F(t)$ mas importantes se presentan en los siguientes ejemplos.

EJEMPLO 1. *Modelo Logístico.*

$$F(t) = \exp(t)/(1+\exp(t))$$

■■

EJEMPLO 2. *Modelo Normal Acumulado.*

$$F(t) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{t} \exp(-y^2/2)\, dy$$

■■

EJEMPLO 3. *Modelo Cauchy Acumulado.*

$$F(t) = 0.5 + 1/\pi \text{ arctg}(t)$$ ■■

La extensión de estos modelos al caso general, $k \geq 2$, no es inmediata. Para tres clases, $k=3$, Lauder (1980) propone algunas extensiones de los modelos Logístico y Normal Acumulado. Sin embargo el único modelo para un número de categorías general, utilizado hasta el momento, es

EJEMPLO 4. *Modelo Logístico Aditivo.*

$$p(\delta=i|x,\theta) = F_i(t(x,\theta)) \quad i=1,\ldots,k$$

con $t(x,\theta)=(t_1,\ldots,t_{k-1})' = \theta'x$. Donde θ es una matriz de parámetros $m \times (k-1)$, x es el vector representante m dimensional.

$$F_i(t) = \exp(t_i)\, F_k(t), \quad i=1,\ldots,k-1$$
$$F_k(t) = \left(1+\sum_{i=1}^{k-1} \exp(t_i)\right)^{-1}$$ ■■

La mayor dificultad que presenta la modelización de $p(\delta|x,\theta)$ viene dada por las restricciones implicadas por su condición de función de probabilidades; en particular la restricción de suma 1. Una forma de evitar esta restricción es modelizando los $k-1$ cocientes $p(\delta=i|x,\theta)/p(\delta=k|x,\theta)$, $i=1,\ldots,k-1$. Esta idea se recoge en la siguiente

DEFINICION 1 *Modelos de Tipo 1.*

Los modelos de clasificación de Tipo 1 son aquellos modelos de clasificación que presentan la siguiente estructura.

$$p(\delta=i|x,\theta) = F_i(t(x,\theta)) \qquad i=1,\ldots,k$$

siendo $t(x,\theta) = (t_1,\ldots,t_{k-1})' = \theta'x$, *mientras que*

$$F_i(t)/F_k(t) = \psi_i(t_i)/(1-\psi_i(t_i)), \quad i=1,\ldots,k-1$$

$$F_k(t) = \left(1+\sum_{i=1}^{k-1} \psi_i(t_i)/(1-\psi_i(t_i))\right)^{-1}$$

Las k-1 funciones, ψ_i, son funciones de distribución univariantes, posiblemente distintas, estrictamente crecientes y dos veces continuamente diferenciables.

El modelo Logístico Aditivo introducido en el ejemplo 4 es un modelo de Tipo 1 con $\psi_i(y) = \exp(y)/(1+\exp(y))$ $i=1,\ldots,k-1$.

Obviamente, los modelos de Tipo 1 cumplen la propiedad P1 de la definición 2.1.1. Las proposiciones siguientes demuestran que estos modelos también cumplen las propiedades P2 y P3. Por tanto, los modelos de Tipo 1 son modelos de Clasificación Regulares.

PROPOSICION 1

Los modelos de Tipo 1 cumplen la propiedad de Monotonicidad que caracteriza a los modelos de clasificación regulares, propiedad P3 en la definición 2.1.1

DEMOSTRACION

Como las funciones $\{\psi_i\}$ son monótonas crecientes, las funciones $\{(1-\psi_i)^{-1}\}$ también lo serán.

Por tanto, la función $F_k(t) = \left(1+\sum_{i=1}^{k-1} \psi_i(t_i)/(1-\psi_i(t_i)\right.$

es monótona decreciente en cada una de sus componentes.

En consecuencia, las funciones

$$F_i(t) = F_k(t)\ \psi_i(t_i)/(1-\psi_i(t_i))$$

también son monótonas decrecientes en todas sus componentes excepto t_i.

Como $F_j(t)$, $j=1,\ldots,k$, $j\neq i$, son monótonas decrecientes respecto a t_i y además $\sum F_j(t)=1$, $F_i(t)$ debe ser monótona creciente con respecto a t_i. Esto demuestra la primera parte de P3.

Por otro lado, si $t_i\downarrow-\infty$ entonces $\psi_i(t_i)\downarrow 0$ y $(1-\psi_i(t_i))\uparrow 1$, luego $\psi_i(t_i)/(1-\psi_i(t_i))\downarrow 0$. Esto es, límite de $F_i(t)$ cuando t_i tiende a $-\infty$ es 0. Propiedad P3.1.

De igual forma, si $t_j\downarrow-\infty$ para todo $j\neq i$, entonces $F_j(t)\downarrow 0$ para todo $j\neq i$. Mientras tanto $F_k(t)\uparrow 1/(1+\psi_i(t_i)/(1-\psi_i(t_i)))$. Si además $t_i\uparrow+\infty$, $F_i(t)=F_k(t)\ \psi_i(t_i)/(1-\psi_i(t_i))\uparrow 1$, ya que $\psi_i(t_i)/(1-\psi_i(t_i))\uparrow+\infty$.

∎

PROPOSICION 2

Los modelos de Tipo 1 cumplen la propiedad de regularidad P2 de la definición 2.1.1

DEMOSTRACION

La primera parte de la propiedad P2, $F_i(t)$ dos veces continuamente diferenciable, se exige explícitamente en la definición de los modelos de Tipo 1. Por tanto solo es necesario demostrar que la matriz

$$G(t) = \sum_{\delta=1}^{k} F_\delta(t)\ D_t^1(\text{Log } F_\delta(t))\ \left(D_t^1(\text{Log } F_\delta(t))\right)'$$

es definida positiva. Ahora bien,

$$a'G(t)a = \sum_{\delta=1}^{k} F_\delta(t) \left[a' D_t^1(\text{Log } F_\delta(t)) \right]^2$$

Como $F_\delta(t)>0 \quad \forall t \in R^{k-1}$ y $\forall \delta=1,\ldots,k$, la matriz $G(t)$ será definida positiva si y solo si $\forall a \in R^{k-1}$, $a \neq 0$, $\exists \delta$ $(=1,\ldots,k-1)$ tal que $a' D_t^1(\text{Log } F_\delta(t)) \neq 0$. Esto es equivalente a decir que la matriz M, matriz cuadrada $(k-1)\times(k-1)$ cuya columna i-ésima es el vector $D_t^1(\text{Log } F_i(t))$, es no singular.

Sea m_{ij} el elemento genérico de la matriz M. Por la regla de la cadena,

$$m_{ii} = \frac{d}{dt_i} \text{Log } F_i(t) = \frac{d}{dt_i} \text{Log } \frac{\psi_i(t_i)}{1-\psi_i(t_i)} +$$

$$+ \frac{d}{dt_i} \text{Log } F_k(t) = \left\{ \frac{1}{\psi_i(1-\psi_i)} + \frac{d}{d\psi_i} \text{Log } F_k \right\} \frac{d\psi_i}{dt_i}$$

$$= \left\{ \frac{1}{\psi_i(1-\psi_i)} - F_k \frac{1}{(1-\psi_i)^2} \right\} \frac{d\psi_i}{dt_i} =$$

$$= \left\{ 1-F_i \right\} \frac{1}{\psi_i(1-\psi_i)} \frac{d\psi_i}{dt_i}$$

$$m_{ij} = \frac{d}{dt_i} \text{Log } F_j = \frac{d}{dt_i} \text{Log } \frac{\psi_j(t_j)}{1-\psi_j(t_j)} +$$

$$+ \frac{d}{dt_i} \text{Log } F_k(t) = -F_i \frac{1}{\psi_i(1-\psi_i)} \frac{d\psi_i}{dt_i} \quad , \; i \neq j$$

Luego $M=M_1 Diag_1$, siendo $Diag_1$ una matriz diagonal $(k-1)\times(k-1)$ con elemento genérico $(Diag_1)_{ii}=1/(\psi_i(1-\psi_i))\ \frac{d\psi_i}{dt_i}$ mientras que M_1 es la matriz $(k-1)\times(k-1)$:

$$M_1 = I - Diag_2\ 1\ 1'$$

donde I es la matriz identidad, 1 es un vector con todas sus componentes iguales a uno y $Diag_2$ es una matriz diagonal con elemento genérico $(Diag_2)_{ii}=F_i$.

Ahora bien, si A es una matriz $r\times r$ y $a\in R^r$ entonces $|I-Aaa'| = 1-a'Aa$, ya que $aa'=PDP'$ siendo P una matriz ortogonal mientras que D es una matriz tal que $(D)_{11}=\|a\|^2$ y 0 en otro caso. Entonces $|I-Aaa'|=|I-APDP'|=|I-P'APD|$, ya que al ser P ortogonal, $PP'=I$, luego $|P||P'|=1$. Ahora bien $P'APD$, por la forma de la matriz D, será una matriz con todas las columnas, excepto la primera, iguales a 0, por tanto la matriz $I-P'APD$ es diagonal y su determinante es $|I-P'APD|=1-(P'APD)_{11}=1-Tr(P'APD)=1-Tr(APDP')=1-Tr(Axx')=1-Tr(x'Ax)=1-x'Ax$. Donde Tr representa la traza, y utilizando la propiedad $Tr(AB)=Tr(BA)$. (Graybill, 1961; Teorema 1.45, pag. 7).

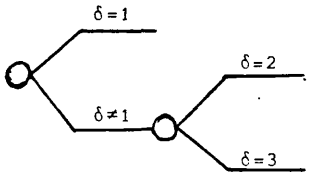
Por tanto,

$$|M| = |M_1 Diag_1| = |M_1||Diag_1| = |I-Diag_2 11'||Diag_1|.$$
$$= |Diag_1|\ (1-1'Diag_2 1) = F_k \prod_{i=1}^{k-1} \frac{1}{\psi_i(1-\psi_i)} \frac{d\psi_i}{dt_i} > 0$$

ya que, al ser ψ_i estrictamente creciente $\forall i$, $F_k>0$, $\psi_i>0$, $\frac{d\psi_i}{dt_i} > 0$.

■

Los modelos de Tipo 1 poseen un gran atractivo al relacionar tan claramente a cada combinación lineal, t_i, con su clase correspondiente. Sin embargo, en muchas situaciones, puede resultar mas interesante un planteamiento secuencial del problema. Así por ejemplo, en un problema de diagnóstico médico en el que el paciente pueda tener un síndrome, S, que puede ser producido por dos causas diferentes, A y B, las clases a considerar son: $(\delta=1) \equiv$ No padece el síndrome S; $(\delta=2) \equiv$ Padece el síndrome y la causa es A; $(\delta=3) \equiv$ Padece el síndrome y la causa es B.



En tales casos parece razonable modelizar $p(\delta=1)$, $p(\delta=2|\delta\neq1)$ y $p(\delta=3|\delta\neq1)$. Esta idea origina la siguiente

DEFINICION 2. *Modelos de Tipo 2.*

Un modelo de clasificación, $p(\delta|x,\theta)$, *se dice que es de tipo 2 si su estructura es*

$$p(\delta=i|x,\theta) = F_i(t(x,\theta)), \qquad i=1,\ldots,k$$

siendo $t(x,\theta)=(t_1,\ldots,t_{k-1})'=\theta'x$, *mientras que*

$$F_i(t) = \psi_1(t_1)$$

$$F_i(t) = \prod_{j=1}^{i-1} (1-\psi_j(t_j))\ \psi_i(t_i), \quad i=2,\ldots,k-1$$

$$F_k(t) = \prod_{j=1}^{k-1} (1-\psi_j(t_j))$$

Donde $\{\psi_i;\ i=1,\ldots,k-1\}$ *son* k-1 *funciones de distribución univariantes, posiblemente distintas, estrictamente crecientes y dos veces continuamente diferenciables.*

EJEMPLO 5. *Modelo Logístico Multiplicativo*

$$p(\delta=i\,|\,x,\theta) = F_i(t(x,\theta)), \qquad i=1,\ldots,k$$

siendo $t(x,\theta) = (t_1,\ldots,t_{k-1})' = \theta' x$, y donde:

$$F_1(t) = \exp(t_1)/(1+\exp(t_1))$$

$$F_i(t) = \exp(t_i)/\prod_{j=1}^{i}(1+\exp(t_j)) \qquad i=1,\ldots,k-1$$

$$F_k(t) = \left[\prod_{j=1}^{k-1}(1+\exp(t_j))\right]^{-1}$$

■■

Este es un modelo de Tipo 2 en el que

$$\psi_i(y) = \exp(y)/(1+\exp(y)), \quad \forall\, i=1,\ldots,k-1$$

EJEMPLO 6. *Modelo Normal Acumulado para* k *clases.*

$$p(\delta=i\,|\,x,\theta) = F_i(t(x,\theta)), \qquad i=1,\ldots,k$$

siendo $t(x,\theta) = (t_1,\ldots,t_{k-1})' = \theta' x$, y donde:

$$F_1(t) = \Phi(t_1) = \int_{-\infty}^{t_1}(2\pi)^{-\frac{1}{2}}\exp(-y^2/2)\,dy$$

$$F_i(t) = \Phi(t_i)\prod_{j=1}^{i-1}\Phi(-t_j), \qquad i=2,\ldots,k-1$$

$$F_k(t) = \prod_{j=1}^{k-1}\Phi(-t_j)$$

■■

Al igual que los modelos de Tipo 1, los modelos de Tipo 2 también son modelos de clasificación regulares. En efecto,

$$F_1(t)+F_2(t) = \psi_2(1-\psi_1)+\psi_1 = \psi_2(1-\psi_1)-(1-\psi_1)+1 = 1-(1-\psi_1)(1-\psi_2)$$

Por inducción, $F_{k-1}(t)+\ldots+F_1(t)=1-(1-\psi_1)\ldots(1-\psi_{k-1})=$ $=1-F_k(t)$, luego $F_k(t)+\ldots+F_1(t)=1$, lo que demuestra P1. La propiedad P3 es consecuencia de la definición, mientras que la demostración de P2 es el objetivo de la siguiente

PROPOSICION 3.

Los modelos de Tipo 2 cumplen la propiedad P2 de la definición 2.1.1.

DEMOSTRACION

Los mismos argumentos esgrimidos en la demostración de la proposición 2 son válidos para concluir que la matriz

$$G(t) = \sum_{\delta=1}^{k} F_\delta(t)\, D_t^1(\text{Log } F_\delta(t)) \left[D_t^1(\text{Log } F_\delta(t))\right]'$$

es definida positiva si y solo si la matriz M con elemento genérico m_{ij},

$$m_{ij} = \frac{d}{dt_j} \text{Log } F_i(t)$$

es no singular.

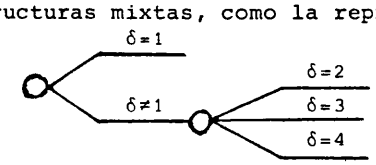
Como $F_i(t)$ no depende de $t_{i+1},\ldots,t_k$, la matriz M es triangular inferior. Por tanto $|M| = \prod_{i=1}^{k-1} m_{ii}$.

$$m_{ii} = \frac{d}{dt_i} \text{Log } F_i(t) = \frac{d}{dt_i}\left[\text{Log } \psi_i(t_i) + \sum_{j=1}^{i-1} \text{Log}(1-\psi_j(t_j))\right] = \frac{d}{dt_i} \text{Log } \psi_i(t_i) =$$

$$= \frac{1}{\psi_i} \frac{d\psi_i}{dt_i}$$

luego, al ser ψ_i estrictamente creciente, $m_{ii}>0$. M es no singular. ∎

Las dos estructuras básicas definidas mediante los modelos de Tipo 1 y Tipo 2 pueden combinarse para obtener estructuras mixtas, como la representada por el árbol



que corresponde a un modelo cuya estructura es:

$$p(\delta=1|t) = \psi_1(t_1)$$

$$p(\delta=2|t) = (1-\psi_1(t_1))\ \frac{\psi_2(t_2)}{1-\psi_2(t_2)}\ S^{-1}(t)$$

$$p(\delta=3|t) = (1-\psi_1(t_1))\ \frac{\psi_3(t_3)}{1-\psi_3(t_3)}\ S^{-1}(t)$$

$$p(\delta=4|t) = (1-\psi_1(t_1))\ S^{-1}(t)$$

siendo $S(t) = 1+ \dfrac{\psi_2(t_2)}{1-\psi_2(t_2)} + \dfrac{\psi_3(t_3)}{1-\psi_3(t_3)}$.

$\psi_1(.),\psi_2(.)$ y $\psi_3(.)$ son tres funciones de distribución univariantes, posiblemente distintas, estrictamente crecientes y dos veces continuamente diferenciables.

2.2 COMPORTAMIENTO ASINTOTICO DE LA FUNCION DE VEROSIMILITUD

Como ya se discutió anteriormente, apartado 1.2, la función de verosimilitud desde el enfoque clasificatorio es

$$V(\theta|D) = \prod_{i=1}^{n} p(\delta_i|x_i,\theta)$$

Si el modelo $p(\delta|x,\theta)$ es un modelo de clasificación regular, utilizando la notación introducida en el apartado 2.1,

$$V(\theta|D) = \prod_{i=1}^{n} F_{\delta i}(\theta' x_i) \qquad (1)$$

El comportamiento de esta función de verosimilitud no es en absoluto usual. En particular, la forma geométrica de la verosimilitud (1) no tiene porqué ser *acampanada*, pudiendo ser estrictamente creciente en determinadas direcciones del espacio paramétrico, como muestra el siguiente

EJEMPLO 1.

Considérese un problema de clasificación entre dos clases alternativas, $\delta\in\{1,2\}$, para el que el vector representante x posee dos componentes: un término constante, igual a uno, y un indicador continuo, y. Esto es, $x=(1,y)'$.

Si se asume el modelo logístico, y el banco de datos está formado por un solo dato, (δ_1,x_1), con $\delta_1=1$ y $x_1=(1,2)$, la función de verosimilitud es:

$$V(\theta|(\delta_1,x_1)) = F_{\delta_1}(\theta'x_1) = F_1((1,2)\ \theta) =$$

$$= \exp(\theta_1+2\theta_2)/(1+\exp(\theta_1+2\theta_2))$$

Considérese el subespacio Ψ y su ortogonal $\Psi^{\perp}$, de tal forma que Ψ es el subespacio generado por el vector $x_1=(1,2)'$. Por tanto, por ortogonalidad, $\omega'x=\omega_1+2\omega_2=0$, $\forall\ \omega\in\Psi^{\perp}$.

$\forall\ \theta\in\Theta\equiv R^2$, sea $\theta=\psi+\omega$ con $\psi\in\Psi$ y $\omega\in\Psi^{\perp}$. Entonces $\theta'x_1=$ $=\psi'x_1+\omega'x_1=\psi'x_1$, por tanto

$$V(\theta|(\delta_1,x_1)) = V(\psi|(\delta_1,x_1)) = \frac{\exp(s\,||\psi||\,||x_1||)}{1+\exp(s\,||\psi||\,||x_1||)}$$

donde $||.||$ representa la norma euclídea, y $s=\text{signo}(\psi'x_1)$.

La función de verosimilitud es, por tanto, constante en las rectas paralelas a $\Psi^{\perp}$, mientras que es creciente en las rectas paralelas a Ψ. De hecho, en estas rectas la función de verosimilitud coincide con la función de distribución logística, figura 2.1.1. En consecuencia, $V(\theta|(\delta_1,x_1))$ no posee máximo, convergiendo a 1 cuando $\theta_1+2\theta_2\uparrow+\infty$.

Algo parecido ocurre si se observan dos nuevos datos, (δ_2,x_2) y (δ_3,x_3) con $\delta_2=\delta_3=2$, $x_2'=(1,0)$, $x_3'=(1,-1)$. Entonces

$$V(\theta|D) = \frac{\exp(\theta_1+2\theta_2)}{1+\exp(\theta_1+2\theta_2)}\ \frac{1}{1+\exp(\theta_1)}\ \frac{1}{1+\exp(\theta_1-\theta_2)}$$

si $\theta_1=-\theta_2\downarrow-\infty$, entonces $V(\theta|D)\uparrow 1$. ∎∎

Una propiedad importante de la función de verosimilitud dada en la expresión (1) es su carácter de función acotada. En efecto, $V(\theta|D) \in (0,1)$ ya que es producto de funciones a valores en el intervalo (0,1).

Esta propiedad premite demostrar la siguiente

PROPOSICION 1

Sea $V(\theta|D(n))$ *la función de verosimilitud asociada a un modelo de clasificación regular, calculada a partir de un banco de datos,* $D(n)$*, de tamaño* n. *Una condición suficiente para que, con probabilidad 1, las colas de la función de verosimilitud bajen aproximándose a cero a medida que el tamaño muestral,* n, *crece,*

$$\lim_{n\to\infty} P\left\{ \lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n)) = 0 \right\} = 1 \qquad (2)$$

es que, con probabilidad 1, exista un subconjunto de datos, D_0*, incluído en* $D(n)$*, para el que* $\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D_0)=0$

Esto es,

$$\lim_{n\to\infty} P\left\{ \exists D_0 \subset D(n): \lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D_0)=0 \right\} = 1$$

implica (3)

$$\lim_{n\to\infty} P\left\{ \lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n))=0 \right\} = 1$$

DEMOSTRACION

La parte izquierda de (3) implica que, con probabilidad 1, la función de verosimilitud se puede descomponer como:

$$V(\theta|D) = V(\theta|D_0)\ V(\theta|D-D_0)$$

donde $\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D_0) = 0$ y donde $V(\theta|D-D_0)<1$,

por tanto, $0 \le \lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D) \le \lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D_0)=0$.

∎

La proposición 1 resulta ser una herramienta muy útil en la comprobación de las condiciones que proporcionan un comportamiento acampanado, al menos asintóticamente, de la función de verosimilitud. Entre otras, las condiciones bajo las cuales existe estimador máximo verosímil.

La siguiente proposición muestra una condición necesaria,

PROPOSICION 2

Si $V(\theta|D)$ *es la función de verosimilitud asociada a un modelo de clasificación regular, una condición necesaria para que* $\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D) = 0$ *es que los vectores representantes incluídos en* D, *formen un sistema de generadores de* R^m. *Siendo* m *el número de componentes de los vectores representantes.*

DEMOSTRACION

Sea X la matriz $m\times n$ cuya columna i-ésima es el vector representante i-ésimo de D, esto es x_i.

Considérese el espacio generado por X, S_X, y su ortogonal $S_X^{\perp}$:

$$S_X \equiv \{t \in R^m : t = X\lambda \text{ con } \lambda \in R^n\} \subset R^m$$

$$S_X^{\perp} \equiv \{t \in R^m : X't = 0\} \subset R^m$$

Si los vectores representantes no forman un sistema de generadores entonces $\text{Dim}(S_X) < m$, luego $\exists\, t^* \in S_X^{\perp}$, $t^* \neq 0$.

Sea $\theta_r \in \Theta$ la matriz $m \times (k-1)$ con elemento (i,j) rt_i^*, y considérese la sucesión $\{\theta_r : r = 1, \ldots\}$.

$$\|\theta_r\|^2 = \sum_{i=1}^{m} \sum_{j=1}^{k-1} r^2 t_i^{*2} = r^2 (k-1) \|t^*\|^2 \Rightarrow$$

$$\Rightarrow \|\theta_r\| = Cr \Rightarrow \lim_{r \to \infty} \|\theta_r\|^2 = +\infty$$

siendo $C = (k-1)^{\frac{1}{2}} \|t^*\|$ una constante positiva.

Además, $\forall x_i$, $i = 1, \ldots, n$, $\theta_r' x_i = 0$ ya que $t^* \in S_X^{\perp}$, luego

$$V(\theta_r | D) = \prod_{i=1}^{n} F_i(0) = \text{Cte. positiva}$$

por tanto:

$$\lim_{\Delta \to \infty} \sup_{\|\theta\| > \Delta} V(\theta | D) \geq \lim_{r \to \infty} V(\theta_r | D) = \prod_{i=1}^{n} F_i(0) > 0$$

∎

Sin embargo esta condición no es suficiente. Como contraejemplo considérese la situación en la cual el banco de datos está constituído solamente por m datos, todos ellos pertenecientes a la clase k, y cuyos vectores representantes son linealmente independientes:

$V(\theta | D) = \prod_{i=1}^{m} F_k(\theta' x_i)$, siendo $X = (x_1 \vdots \ldots \vdots x_m)_{m \times m}$ no singular.

Sea t^* la primera fila de la matriz X^{-1}, y sea $(\theta_r)_{ij} = -rt_j^*$, $r \in \mathbb{N}$. Al igual que en la demostración anterior, $\lim_{r\to\infty} \|\theta_r\| = +\infty$ mientras que, por definición de t^*,

$$\theta_r' x_i = -r(1,\ldots,1)' \quad \text{si } i=1$$
$$= 0, \text{ en otro caso}$$

luego, $V(\theta_r|D) = F_k(-r,\ldots,-r)\ (F_k(0))^{m-1}$, pero por la proposición 2.1.1, $\lim_{r\to\infty} F_k(-r,\ldots,-r) = 1$, luego

$$\lim_{r\to\infty} V(\theta_r|D) = (F_k(0))^{m-1} > 0$$

de donde se deduce,

$$\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D) \geq \lim_{r\to\infty} V(\theta_r|D) > 0$$

por tanto no es suficiente que la matriz X tenga rango m.

El siguiente teorema presenta las condiciones bajo las cuales se cumple (2) cuando la distribución de los vectores representantes es discreta.

TEOREMA 1

Si la distribución de probabilidades sobre el espacio χ de vectores representantes es discreta, entonces una condición necesaria y suficiente para que

$$\lim_{n\to\infty} P\left\{ \lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n))=0 \right\} = 1$$

es que existan m *vectores* $\{x_i, i=1,\ldots,m\} \subset \chi$, *linealmente independientes y tales que* $p(x_i)>0$, $i=1,\ldots,m$.

La demostración de este teorema necesita de los siguientes resultados previos:

PROPOSICION 3

Sea F una función vectorial de R^{k-1} *en* R^k, *que cumple las propiedades P1 y P3 de la definición de modelos de clasificación regulares, def. 2.1.1. Entonces*

$$\lim_{\|t\|\to\infty} \prod_{i=1}^{k} F_i(t) = 0$$

DEMOSTRACION

Por la propiedad P3, dado $\varepsilon>0$ $\exists\, \rho_{1i}(\varepsilon)>0$ y $\rho_{2i}(\varepsilon)>0$, $i=1,\ldots,k-1$, tales que

$$\forall\, t\in R^{k-1} \text{ con } t_i<-\rho_{1i},\ F_i(t)<\varepsilon$$

$$\forall\, t\in R^{k-1} \text{ con } t_i>\rho_{2i},\ t_j<-\rho_{2i} \quad j\neq i,\ F_i(t)>1-\varepsilon$$

Sea $\Delta(\varepsilon)=(k-1)^{\frac{1}{2}}\ \text{Max}\{\rho_{ji}\}, j=1,2,\ i=1,\ldots,k-1$. Si $\|t\|>\Delta(\varepsilon)$ entonces

$$\|t\|^2=\sum_{i=1}^{k-1} t_i^2>\Delta^2(\varepsilon) \Rightarrow t_{i*}^2=\text{Max}_i\ t_i^2 \geq \frac{1}{k-1} \sum_{i=1}^{k-1} t_i^2 >$$

$$> \frac{1}{k-1}\,\Delta^2(\varepsilon) \Rightarrow |t_{i*}| \quad \frac{\Delta(\varepsilon)}{(k-1)^{\frac{1}{2}}}$$

$$\text{Si } t_{i*}<0 \Rightarrow t_{i*}<-(k-1)^{-\frac{1}{2}}\ \Delta(\varepsilon) < -\rho_{1i*} \Rightarrow F_{i*}(t)<\varepsilon \Rightarrow$$

$$\Rightarrow \prod_{i=1}^{k} F_i(t)<F_{i*}(t)<\varepsilon\ .$$

Si $t_{i*}>0 \Rightarrow t_{i*}>(k-1)^{-\frac{1}{2}}\ \Delta(\varepsilon)>\rho_{2i*}$ además, como $|t_i|\leq|t_{i*}|\ \forall i$, $t_i\geq-|t_i|\geq-|t_{i*}|=-t_{i*}$, por tanto, utilizando la monotonicidad de F_k, $F_k(t)\leq F_k(\hat{t})$, siendo $\hat{t}_i=t_{i*}$ si $i=i*$, $\hat{t}_i=-t_{i*}$ si $i\neq i*$. Luego $\hat{t}_{i*}=t_{i*}>\rho_{2i}$ y

$\hat{t}_j = -t_{i*} < -\rho_{2i*}$ $\forall\, j \neq i*$, luego $F_{i*}(\hat{t}) > 1-\varepsilon$. Pero

$$F_k(t) \leq F_k(\hat{t}) = 1 - \sum_{i=1}^{k-1} F_i(\hat{t}) \leq 1 - F_{i*}(\hat{t}) < \varepsilon \Rightarrow$$

$$\Rightarrow \prod_{i=1}^{k} F_i(t) < F_k(t) < \varepsilon$$

Por tanto, $\forall \varepsilon > 0$ $\exists \Delta(\varepsilon)$ tal que $\|t\| > \Delta(\varepsilon)$ implica $\prod_{i=1}^{k} F_i(t) < \varepsilon$. Lo que demuestra la proposición. ∎

PROPOSICION 4

Sea $V(\theta|D)$ *la función de verosimilitud correspondiente a un banco de datos,* D, *formado por* m *vectores representantes distintos,* x_i i=1,...,m, *de manera que cada uno de ellos aparezca exactamente* k *veces, cada una de ellas perteneciendo a una clase distinta. Además,* $\{x_i;\ i=1,...,m\}$ *es una base de* R^m. *Entonces*

$$\lim_{\Delta \to \infty} \sup_{\|\theta\| > \Delta} V(\theta|D) = 0$$

DEMOSTRACION

Sea X la matriz cuadrada cuya columna i-ésima es el vector x_i. Como $\{x_i\}$ es una base de R^m, X es no singular. Considérese la matriz definida positiva y simétrica XX'. Los valores propios de XX' son todos estrictamente positivos ya que se trata de una matriz definida positiva. Sean $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m > 0$ esos valores propios.

Sea $\psi \in R^m$, considérese $\sum_{i=1}^{m} (\psi' x_i)^2 = \sum_{i=1}^{m} \psi' x_i x_i' \psi =$

$= \psi' \left(\sum_{i=1}^{m} x_i x_i' \right) \psi = \psi' XX' \psi = \|\psi\|^2\, u_\psi' X' X u_\psi$, donde u_ψ es el vector unitario en la dirección de ψ. Ahora bien, un cono-

cido teorema en formas cuadráticas,(ver por ejemplo: Gantmacher, 1977; vol. 1, teorema 10, pag. 319), demuestra que Min $\{u'Au;\ u\in R^m,\ \|u\|=1\} = \lambda_m(A)$, vector propio mas pequeño asociado a la matriz A. En consecuencia:

$$\sum_{i=1}^{m} (\psi' x_i)^2 = \|\psi\|^2\, u_\psi' X' X u_\psi \geq \|\psi\|^2\, \lambda_m$$

por tanto $\exists\, i^*$ tal que $(\psi' x_{i^*})^2 \geq \frac{1}{m} \|\psi\|^2\, \lambda_m$.

Dado $\varepsilon>0$ sea $\Delta_1(\varepsilon) \geq (m(k-1)/\lambda_m)^{\frac{1}{2}}\, \Delta(\varepsilon)$ donde $\Delta(\varepsilon)$ se define como en la proposición anterior. Si $\|\theta\|>\Delta_1(\varepsilon)$ entonces:

$$\|\theta\|^2 = \sum_{i=1}^{k-1} \sum_{j=1}^{m} \theta_{ji}^2 = \sum_{i=1}^{k-1} \|\theta_{.i}\|^2 > (\Delta_1(\varepsilon))^2,$$

donde $\theta_{.i}$ es la columna i-ésima de la matriz θ. Sea $\psi=\theta_{.i^{**}}$ con $\|\psi\| = \max_i \|\theta_{.i}\|$, entonces $\|\psi\|^2 > (\Delta(\varepsilon))^2/(k-1)$, por tanto:

$$\|\theta' x_{i^*}\|^2 = \sum_{i=1}^{k-1} (\theta_{.i}'\, x_{i^*})^2 \geq (\psi' x_{i^*})^2 \geq \|\psi\|^2 \lambda_m \ /m>$$

$$> \lambda_m (\Delta_1(\varepsilon))^2/(m(k-1)) \geq (\Delta(\varepsilon))^2 \Rightarrow$$

$$\Rightarrow \|\theta' x_{i^*}\| > \Delta(\varepsilon)$$

por la acotación de las verosimilitudes, y por la proposición 3,

$$V(\theta|D) \leq \prod_{j=1}^{k} F_j(\theta' x_{i^*}) < \varepsilon$$

∎

DEMOSTRACION DEL TEOREMA 1

Parte 1, suficiencia.

Sea D^* el banco de datos en la proposición 4. Como $p(x_i)>0$ y como $F_\delta(\theta' x)>0 \; \forall \theta \in R^{m(k-1)}$, la probabilidad de formar D^* con los $m \times k$ primeros datos es estrictamente positiva. Por tanto, con probabilidad 1 cuando el número de datos, n, tienda a infinito, el banco D^* será un subconjunto del banco $D(n)$. La suficiencia es, por consiguiente, una consecuencia inmediata de las proposiciones 1 y 4.

Parte 2, necesidad:

$$\text{Si } \lim_{n\to\infty} P\left\{\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n))=0\right\} = 1, \text{ entonces,}$$

dado un n suficientemente grande,

$$P\left\{\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n))=0\right\} > 0$$

Ahora bien, por la proposición 2, una condición necesaria para que $\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n))=0$ es que existan m vectores representantes linealmente independientes, pertenecientes al banco de datos, por tanto:

$$P\left\{\exists \{x_i;\ i=1,\ldots,m\}, \text{linealmente indep.}\right\} > 0$$

∎

EJEMPLO 2

Considérese el problema de clasificación comentado en el ejemplo 1. Esto es: dos clases, $\delta \in \{1,2\}$; vector representante bidimensional, $x'=(1,y)$; y modelo logístico.

Considérese el banco de datos de tamaño cuatro, $n=4$,

$D \equiv \{(x_i, \delta_i);\ i=1,2,3,4\}$, con $\delta_1=\delta_3=1$, $\delta_2=\delta_4=2$, $x_1=x_2=(1,2)'$ y $x_3=x_4=(1,1)'$. Banco de datos que cumple las condiciones de la proposición 4.

La función de verosimilitud es:

$$V(\theta|D) = \prod_{i=1}^{4} p(\delta_i|x_i,\theta) = V_1(\theta)\ V_2(\theta)$$

con

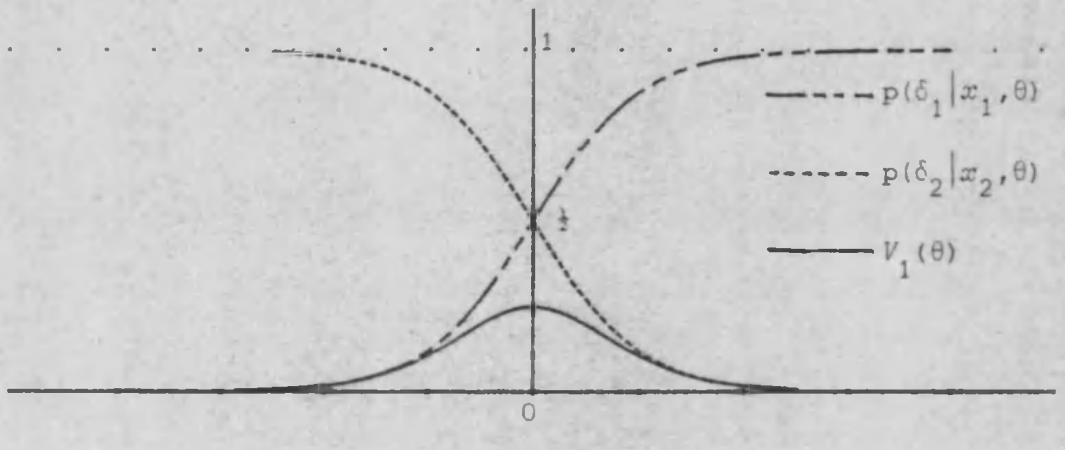
$$V_1(\theta) = p(\delta_1|x_1,\theta)\ p(\delta_2|x_2,\theta) = \frac{\exp(\theta_1+2\theta_2)}{1+\exp(\theta_1+2\theta_2)} \times \frac{1}{1+\exp(\theta_1+2\theta_2)}$$

$$V_2(\theta) = p(\delta_3|x_3,\theta)\ p(\delta_4|x_4,\theta) = \frac{\exp(\theta_1+\theta_2)}{1+\exp(\theta_1+\theta_2)} \times \frac{1}{1+\exp(\theta_1+\theta_2)}$$

Como ya se comprobó en el ejemplo 1, tanto $p(\delta_1|x_1,\theta)$ como $p(\delta_2|x_2,\theta)=1-p(\delta_1|x_1,\theta)$, permanecen constantes en todos los subespacios afines paralelos al subespacio $\Psi^{\perp}$, definido mediante la ecuación $\theta_1+2\theta_2=0$.

La figura 1 muestra el comportamiento de las funciones $p(\delta_1|x_1,\theta)$, $p(\delta_2|x_2,\theta)$ y $V_1(\theta)$ a lo largo del subespacio Ψ, definido por la ecuación $\theta_1=\frac{1}{2}\theta_2$.

El máximo de $V_1(\theta)$ es 1/4 y se alcanza cuando $\theta\in\Psi^{\perp}$, esto es , en todo $\theta\in\Theta$ tal que $\theta_1+2\theta_2=0$. Además, como la

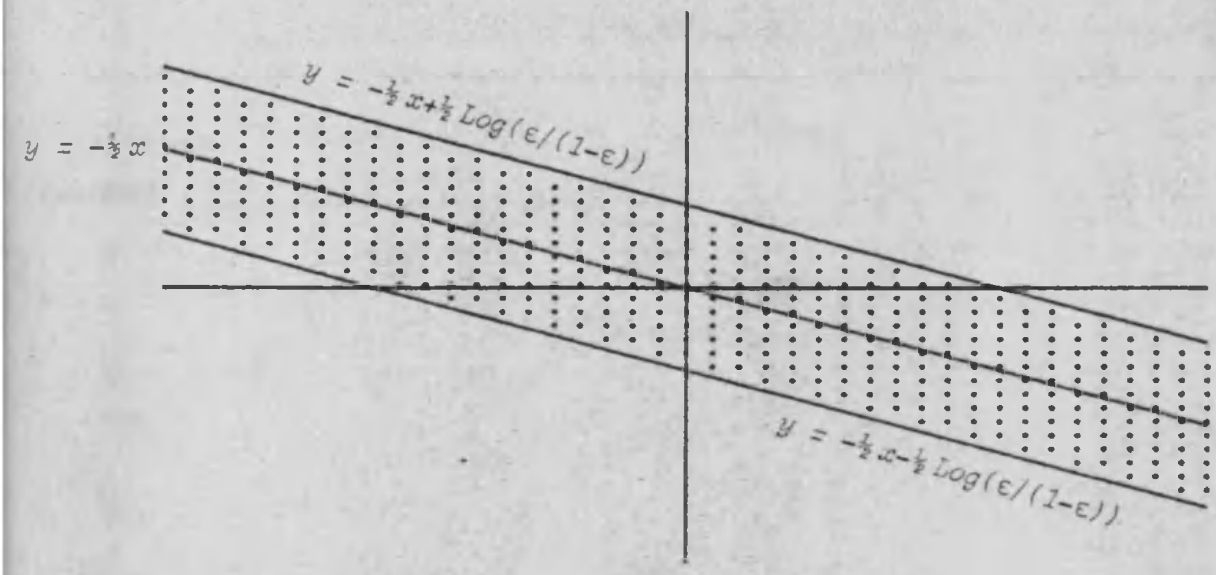


- fig. 1 -
Función $V_1(\theta)$

función $\exp(z)/(1+\exp(z))$ es monótona creciente, simétrica respecto al punto $(0,\frac{1}{2})$ y con inversa $\mathrm{Log}(r/(1-r))$, dado $\varepsilon>0$, si $|\theta_1+2\theta_2|>\mathrm{Log}(\varepsilon/(1-\varepsilon))$ entonces: o bien $p(\delta_1|x_1,\theta)<\varepsilon$; o bien $p(\delta_2|x_2,\theta)=1-p(\delta_1|x_1,\theta)<\varepsilon$. Luego $V_1(\theta)<\varepsilon$, esto es, dado $\varepsilon>0$, $V_1(\theta)<\varepsilon\ \forall\ \theta\in R_1(\varepsilon)$, siendo

$$R_1(\varepsilon) \equiv \{\theta\in\Theta\equiv\mathbb{R}^2;\ |\theta_1+2\theta_2|<\mathrm{Log}(\varepsilon/(1-\varepsilon))\}$$

Esta región se ha dibujado en la figura 2.

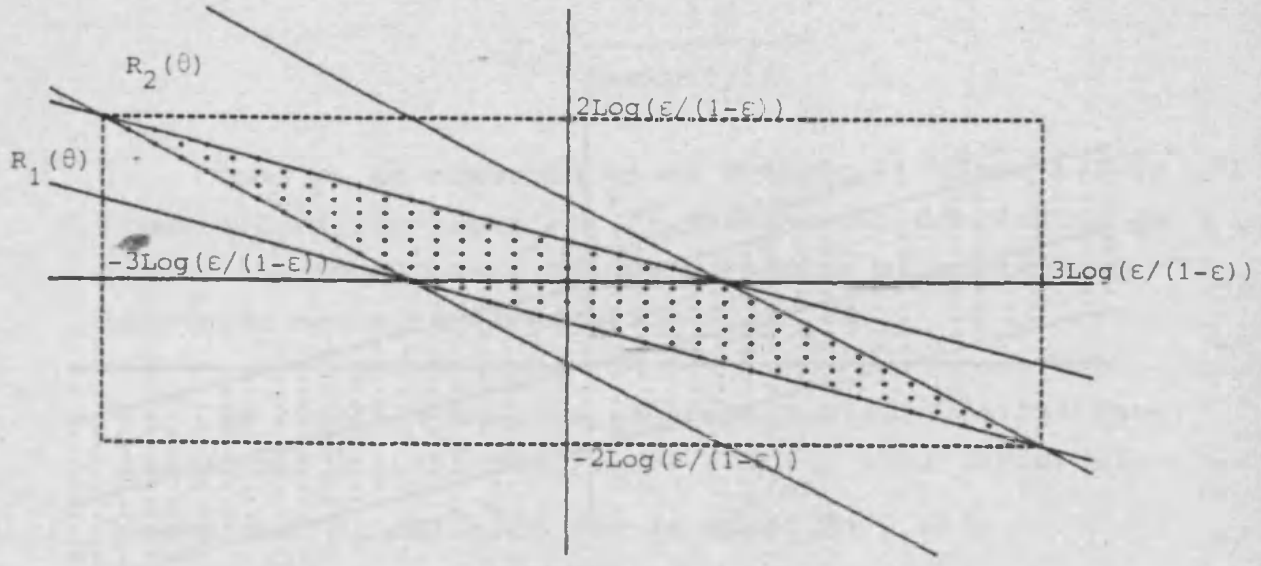


- fig. 2 -
Región $R_1(\varepsilon)$

Similarmente, el máximo para $\theta\in\Theta$ de $V_2(\theta)$ es 1/4, y ese máximo se alcanza para todo θ perteneciente a la recta $\theta_1+\theta_2=0$. Además, dado $\varepsilon>0$, $V_2(\theta)<\varepsilon \quad \forall\, \theta\in R_2(\varepsilon)$, siendo $R_2(\varepsilon)\equiv\{\theta\in R^2;\ |\theta_1+\theta_2|<\mathrm{Log}(\varepsilon/(1-\varepsilon))\}$.

En consecuencia, el máximo de $V(\theta|D)=V_1(\theta)V_2(\theta)$ se alcanzará en el punto de intersención de las rectas $\theta_1+2\theta_2=0$ y $\theta_1+\theta_2=0$, esto es, en $\theta_1=\theta_2=0$. Es mas, si dado $\varepsilon>0$ se define $R(\varepsilon)=R_1(\varepsilon)\cap R_2(\varepsilon)$, (ver figura 3), esa región está incluída en el conjunto $\{\theta\in R^2;\ \|\theta\|<\sqrt{13}\ \mathrm{Log}(\varepsilon/(1-\varepsilon))\}$.

Por tanto, para todo $\theta\in\Theta$ tal que su norma euclídea sea mayor que $\sqrt{13}\ \mathrm{Log}(\varepsilon/(1-\varepsilon))$, la función $V(\theta|D)$ será menor que ε, ya que $\theta\notin R(\varepsilon)$ y por tanto $V_1(\theta)<\varepsilon$ o $V_2(\theta)<\varepsilon$, siendo ambas funciones menores que uno.



- *fig. 3* -
Región R(ε)

■■

El teorema 1 proporciona una condición necesaria y suficiente para que se cumpla la expresión (1) en el caso discreto, no necesariamente finito. La extensión de este resultado al caso general presenta ciertas complicaciones adicionales puesto que, si algunas componentes del vector representante, x, son continuas, la probabilidad de que en el banco de datos existan dos vectores representantes iguales es cero.

Esta dificultad puede resolverse considerando, en lugar de vectores representantes iguales pertenecientes a distintas clases, vectores representantes *suficientemente parecidos*. Así, puede considerarse que en el espacio de vectores representantes existen m bolas, $\{B_i; i=1,\ldots,m\}$ con radios suficientemente pequeños para que, elegido un conjunto cualquiera formado por m vectores representantes, uno de cada una de las m bolas, dicho conjunto sea linealmente independiente. De esta forma, las ideas implícitas en la demostración del teorema 1 pueden utilizarse para demostrar el caso general.

Sea $\{x_i;\ i=1,\ldots,m\}$, una base de R^m y sea X la matriz no singular cuadrada cuya columna i-ésima es el vector x_i. Sean $\{r_i;\ i=1,\ldots,m\}$ los números reales positivos definidos como $r_i=\frac{1}{2}\|x_i\|/(1+m\lambda_m^{-\frac{1}{2}}\|x_i\|)$, donde λ_m es el valor propio mas pequeño de la matriz definida positiva XX'. Con esta notación, la generalización del teorema 1 al caso continuo es.

TEOREMA 2

Dado un modelo de-clasificación regular, $p(\hat{c}|x,\theta)$, *y un banco de datos de tamaño* n, $D(n)$, *sea* $V(\theta|D(n))$ *la*

función de verosimilitud correspondiente. Si la distribución de probabilidades sobre el espacio de vectores representantes, χ, es tal que $P(y \in B_{r_i}(x_i)) > 0$ $i=1,\ldots,m$, *siendo* $B_{r_i}(x_i)$ *la bola abierta de centro* x_i *y radio* r_i, *entonces*

$$\lim_{n\to\infty} P\left[\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n)) = 0\right] = 1$$

La definición utilizada por el teorema para los radios r_i, introducida anteriormente, es un tanto arbitraria. Cualquier valor estrictamente positivo, mas pequeño, sería válido; así como posiblemente valores mayores. Han sido elegidos concretamente esos valores para simplificar los cálculos.

Los radios r_i propuestos por el teorema están bien definidos. En efecto, $r_i > 0$ ya que $m \geq 1$, $\|x_i\| > 0$ y $\lambda_m > 0$, pues es un valor propio de la matriz definida positiva XX'. Además, $r_i < \|x_i\|$ puesto que $1 + m\lambda_m^{-\frac{1}{2}}\|x_i\| > 1$, por tanto $0 \notin B_{r_i}(x_i)$ para todo $i=1,\ldots,m$.

Un resultado previo a la demostración del teorema 2 es la siguiente

PROPOSICION 5

Sean $\{B_{r_i}(x_i);\ i=1,\ldots,m\}$ *las* m *bolas abiertas definidas en el teorema 1. Sea* $\psi \in R^m$ *y sea* $x=x_{i*}$ *tal que* $|\psi'x| = \max_i |\psi'x_i|$. *Entonces, para todo* $y \in B_r(x)$ *con* $x=x_{i*}$ $r=r_{i*}$, *se cumple que* $|\psi'y| > \frac{1}{2}|\psi'x|$. *Es mas, si* $\psi'x > 0$ *entonces* $\psi'y > \frac{1}{2}\psi'x > 0$, *mientras que si* $\psi'x < 0$ *entonces* $\psi'y < \frac{1}{2}\psi'x < 0$, *y si* $\psi'x = 0$ *entonces* $\psi'y = 0$

DEMOSTRACION

Si $\psi' x=0$ entonces, para todo $i=1,\ldots,m$, $\psi' x_i=0$. Pero como $\{x_i;\ i=1,\ldots,m\}$ es una base de R^m, ψ debe ser cero, por tanto $\psi' y=0$ para todo $y\in R^m$. Supóngase pues, que $|\psi' x|>0$.

Sea $y\in B_r(x)$ y sean y_1, y_2 tales que $y=y_1+y_2$ con $y_1=ax$, $y_2 \perp x$, esto es, y_1 es la proyección de y en el espacio generado por el vector x, mientras que y_2 es la proyección de y en el espacio ortogonal al generado por x. Como $y\in B_r(x)$, $r^2>\|x-y\|^2=\|x\|^2+\|y\|^2-2x'y$, y como $r<\|x\|$, $r^2>\|x\|^2+\|y\|^2-2x'y>r^2+\|y\|^2-2x'y \Rightarrow x'y>\frac{1}{2}\|y\|^2\geq 0$. Además, $x'y=x'(y_1+y_2)=x'y_1=ax'x=a\|x\|^2>0 \Rightarrow a>0 \Rightarrow$
$\Rightarrow \|y_1\|=|a|\,\|x\|=a\|x\| \Rightarrow a=\|y_1\|/\|x\| \Rightarrow y_1=\|y_1\|x/\|x\|$

Por otra parte, como $y_2 \perp x \Rightarrow y_2 \perp y_1-x$, entonces:

$$r^2>\|y-x\|^2=\|y_1-x+y_2\|^2=\|y_1-x\|^2+\|y_2\|^2 \Rightarrow$$

$$\Rightarrow \begin{cases} \|y_2\|<r \\ \|y_1-x\|<r \Rightarrow y_1\in B_r(x) \end{cases}$$

Como $y_1\in B_r(x) \Rightarrow \|x\|=\|x-y_1+y_1\|\leq\|x-y_1\|+\|y_1\|<$
$<r+\|y_1\| \Rightarrow \|y_1\|>\|x\|-r$, luego $\|y_1\|/\|x\|>1-r/\|x\|$.

Por otra parte, $y_2=\|y_2\|u$ con $\|y_2\|<r$ y $u\in R^m$, vector unitario en la dirección y_2. Como $\{x_i\}$ es una base de R^m, existe $t\in R^m$ tal que $u=Xt \Rightarrow t=X^{-1}u$, donde X

es la matriz cuadrada cuya columna i-ésima es el vector x_i. Además,

$$\|t\|^2 = t't = u'(X^{-1})'X^{-1}u = u'(XX')^{-1}u$$

ahora bien, Max $\{u'Mu;\ u\in R^m,\ \|u\|=1\} = \lambda_1(M)$, donde $\lambda_1(M)$ es el valor propio mas grande de la matriz M, (Gantmacher, 1977; vol. 1, teorema 13, pag. 322), por tanto:

$$\|t\|^2 = u'(XX')^{-1}u \leq \lambda_1((XX')^{-1}) = 1/\lambda_m$$

ya que los valores propios de M^{-1} son los inversos de los valores propios de M.

Por tanto, $\sum_{i=1}^{m} t_i^2 = \|t\|^2 \leq 1/\lambda_m \Rightarrow t_i^2 \leq 1/\lambda_m \quad \forall i=1,\ldots,m \Rightarrow |t_i| \leq \lambda_m^{-\frac{1}{2}}$

Considérese ahora:

$$|\psi'y_2| = \|y_2\|\,|\psi'u| = \|y_2\| \left|\sum_{i=1}^{m} t_i\psi'x_i\right| \leq$$

$$\leq \|y_2\| \sum_{i=1}^{m} |t_i|\,|\psi'x_i| \leq \|y_2\|\,|\psi'x| \sum_{i=1}^{m} |t_i| \leq$$

$$\leq \|y_2\|\,|\psi'x|\,m\lambda_m^{-\frac{1}{2}} < rm\lambda_m^{-\frac{1}{2}}|\psi'x|$$

Si $\psi'x>0$:

$$\psi'y = \psi'y_1 + \psi'y_2 = \psi'x\,\|y_1\|/\|x\| + \psi'y_2 > (1-r/\|x\|)\psi'x + \psi'y_2 >$$

$$> (1-r/\|x\|)\psi'x - |\psi'y_2| > (1-r/\|x\|)\psi'x - rm\lambda_m^{-\frac{1}{2}}\psi'x =$$

$$= (1-r(1/\|x\|+m\lambda_m^{-\frac{1}{2}}))\psi'x = \tfrac{1}{2}\psi'x > 0.$$

Si $\psi'x<0$:

$$\psi'y = \psi'y_1+\psi'y_2 = \psi'x\|y_1\|/\|x\|+\psi'y_2 < (1-r/\|x\|)\psi'x+\psi'y_2$$

$$< (1-r/\|x\|)\psi'x+|\psi'y_2| < (1-r/\|x\|)\psi'x+rm\lambda_m^{-\frac{1}{2}}|\psi'x|$$

$$= (1-r/\|x\|)\psi'x-rm\lambda_m^{-\frac{1}{2}}\psi'x = (1-r(1/\|x\|+m\lambda_m^{-\frac{1}{2}}))\psi'x =$$

$$= \tfrac{1}{2}\psi'x < 0$$

∎

PROPOSICION 6

Sea $V(\theta|D)$ *la función de verosimilitud correspondiente a un banco de datos,* D, *de tamaño* m×k, *de forma que exactamente* k *datos posean un vector representante incluído en cada una de las* m *bolas* $\{B_{r_i}(x_i);\ i=1,\dots,m\}$, *definidas en el teorema 2. Además, los* k *datos con vector representante perteneciente a cada una de las* m *bolas, pertenecen a clases distintas. Entonces:*

$$\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D) = 0$$

DEMOSTRACION

Dado $\varepsilon>0$, sea $\Delta_2(\varepsilon)=2\Delta_1(\varepsilon)$, donde $\Delta_1(\varepsilon)$ se define como en la demostración de la proposición 4. Siguiendo el mismo razonamiento que allí:

$\forall\theta$, $\|\theta\|>\Delta_2(\varepsilon)$ $\exists i^*$ y $\exists i^{**}$ tales que, si $\psi=\theta_{i^*}$ y $x=x_{i^{**}}$, entonces $|\psi'x|>\lambda_m(m(k-1))^{-\frac{1}{2}}\Delta_2(\varepsilon)=2\Delta(\varepsilon)$.

Si $\psi'x<0$:

$V(\theta|D) \leq F_{i*}(\theta' y_{i*})$ con $y_{i*} \in B_{r_{i**}}(x_{i**})$, vector representante del dato perteneciente a la clase $i*$. La componente $i*$ de $\theta' y_{i*}$ será $\psi' y_{i*}$ que, por la proposición anterior, $\psi' y_{i*} < \frac{1}{2}\psi' x < -\Delta(\varepsilon)$. Por tanto, propiedad P3, $F_{i*}(\theta' y_{i*}) < \varepsilon \Rightarrow V(\theta|D) < \varepsilon$.

Si $\psi'x>0$:

$V(\theta|D) \leq F_k(\theta' y_{i*})$ con $\psi' y_{i*} > \frac{1}{2}\psi' x > \Delta(\varepsilon)$, por tanto, $F_k(\theta' y_{i*}) < \varepsilon$, luego $V(\theta|D) < \varepsilon$

En resumen, dado $\varepsilon>0$ existe $\Delta_2(\varepsilon)$ tal que si $\|\theta\| > \Delta_2(\varepsilon)$ entonces $V(\theta|D) < \varepsilon$, luego $\sup_{\|\theta\| > \Delta_2(\varepsilon)} V(\theta|D) < \varepsilon$. ∎

DEMOSTRACION DEL TEOREMA 2

Sea D^* el banco de datos utilizado en la proposición 6. Como $p(y \in B_{r_i}(x_i)) > 0$ para todo $i=1,\ldots,m$ y $p(\delta|y,\theta) > 0$ para todo $\theta \in R^{m(k-1)}$, la probabilidad de formar un banco de datos similar a D^* con los $m \times k$ primeros datos es estrictamente positiva. Por tanto, con probabilidad uno cuando el número de datos, n, tiende a infinito, el banco de datos D^* será un subconjunto del banco $D(n)$. Este hecho, conjuntamente con las proposiciones 1 y 6, demuestra el teorema. ∎

El teorema 2 demuestra que, si bien las propiedades

asintóticas de la función de verosimilitud dependen en alguna medida del modelo probabilístico que genera los vectores representantes $x \in \mathcal{X}$, esta dependencia no es importante. Así por ejemplo, si la distribución sobre $\mathcal{X}$ es absolutamente continua y existe un abierto $E \subset \{x \in \mathcal{X} ; p(x)>0\}$, entonces las condiciones del teorema 2 se cumplen automáticamente; lo mismo ocurre si alguna de las componentes de x es discreta, pero solamente una de ellas, como máximo, es constante y no existen *demasiadas* combinaciones imposibles.

Otro resultado interesante, íntimamente relacionado con ese teorema, es la siguiente

PROPOSICION 7

Si se cumplen las hipótesis del teorema 2, la probabilidad de que exista estimador máximo verosímil converge a uno conforme el número de datos, n, *tiende a infinito.*

DEMOSTRACION

Supongase que $\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D) = 0$. Sea θ_0 punto interior de Θ. Como $p(\delta|x,\theta)>0$ para todo $\theta\in\Theta$ y para todo (δ,x), entonces $V(\theta|D(n))>0$.

$$\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D)=0 \Rightarrow \text{Dado } \varepsilon>0 \ \exists \Delta(\varepsilon) \text{ tal que}$$

$$\sup_{\|\theta\|>\Delta} V(\theta|D)<\varepsilon \Rightarrow V(\theta|D)<\varepsilon \ \forall \theta \text{ tal que } \|\theta\|>\Delta(\varepsilon).$$

Sea $\Delta^* = \Delta(V(\theta_0|D)/2)$ y considérese la bola compacta de centro el origen y radio Δ^*, $E_{\Delta^*}(0)$. Como la

función $V(\theta|\mathcal{D})$ es continua en $B_{\Delta^*}(0)$, alcanzará un máximo en algún punto $\hat{\theta}$ de esa bola, luego:

$$\forall \theta \in B_{\Delta^*}(0),\ V(\theta|D) \leq V(\hat{\theta}|D),$$

en particular $V(\hat{\theta}|D) \geq V(\theta_0|D)$. Además,

$$\forall \theta \notin B_{\Delta^*}(0),\ \|\theta\| > \Delta^*, \text{ luego } V(\theta|D) < \tfrac{1}{2} V(\theta_0|D) \leq V(\hat{\theta}|D).$$

Por tanto $\hat{\theta}$ proporciona un máximo absoluto de la función $V(\theta|\mathcal{D})$.

Aplicando ahora el teorema 2,

$$\lim_{n\to\infty} P\left\{\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} V(\theta|D(n)) = 0\right\} = 1$$

lo que demuestra la proposición. ∎

CAPITULO 3

RESULTADOS ASINTOTICOS

En este capítulo se propone un conjunto de condiciones de regularidad de la distribución de vectores representantes. Se demuestra que esas condiciones son suficientes para la convergencia en distribución de la distribución final del parámetro de un modelo de clasificación regular a la distribución Normal asintótica usual.

Los modelos de clasificación regulares no cumplen las axiomáticas sobre normalidad asintótica propuestas en la literatura especializada. Por ello, este capítulo necesita los resultados obtenidos en el apéndice 1, en el que se debilita la axiomática propuesta por Walker (1969).

Por último, se estudia la velocidad de convergencia en dos ejemplos en los que se utilizan bancos de datos simulados. Se comprueba que, aunque la distribución final converge muy lentamente a la aproximación Normal, las distribuciones de clasificación proporcionadas por la aproximación asintótica son suficientemente buenas para tamaños muestrales no demasiado grandes.

3.1 COMPORTAMIENTO ASINTOTICO DE LA DISTRIBUCION FINAL

En todo análisis bayesiano de un modelo probabilístico, o de una familia de modelos, un apartado importante lo debe constituír el estudio del comportamiento asintótico de la distribución final, no solamente por su valor intrínseco, proporcionando por lo general una aproximación sencilla a la distribución final, tanto mejor cuanto mayor sea el número de datos; sino también como elemento importante en la búsqueda de *distribuciones de referencia* (Bernardo, 1979).

Sin duda, el resultado mas importante en teoría asintótica bayesiana, es el que establece que, bajo *condiciones de regularidad* suficientemente generales, la distribución final es aproximadamente Normal con media el estimador máximo verosímil y matriz de precisión, la matriz de información de Fisher. Sin embargo, los modelos de clasificación regulares no cumplen ninguno de los conjuntos de condiciones de regularidad hasta de ahora propuestos para la normalidad asintótica de la distribución final, ni tan siquiera para la consistencia del estimador máximo verosímil. Por ejemplo, como se demuestra en el apéndice 1, no cumplen el axioma A5 en Walker (1969), ni el axioma 6 en Johnson (1970), ni el axioma 5 en Wald (1945), ni la hipótesis 2 en Kiefer and Wolfowitz (1956), etc...

Los conjuntos de condiciones de regularidad para la normalidad asintótica de la distribución final que aparecen en la literatura especializada, son todos ellos condiciones suficientes pero no necesarias. En particular, la axiomática propuesta por Walker (1969) ya ha

sido debilitada por Dawid (1970), aunque manteniendo el axioma A5. En el apéndice 1 se propone una condición alternativa, CA, estrictamente mas débil que A5. En efecto, se comprueba que A5 implica CA pero que, por ejemplo, los modelos de clasificación regulares cumplen CA pero no cumplen A5. Además se demuestra, teorema A1.3.2, que la axiomática que se obtiene al sustituír CA por A5 sigue siendo suficiente, aunque todavía no necesaria, para la normalidad asintótica de la distribución final.

En los problemas de clasificación desde el enfoque clasificatorio, la generación de los datos, $\{(\delta_i, x_i); i=1,\ldots,n\}$, está determinada por el modelo de clasificación, $p(\delta|x,\theta)$, y por la distribución de los vectores representantes, $p(x)$. Para asegurar la normalidad asintótica de la distribución final es necesario imponer condiciones tanto a $p(\delta|x,\theta)$ como a $p(x)$. El siguiente teorema recoge un conjunto de condiciones *suficientes* para la normalidad asintótica de la distribución final.

TEOREMA 1

Sea $p(\delta|\mathbf{x},\theta)$ *un modelo de clasificación regular, y sean* $\mathbf{x}_1$ *y* $\mathbf{x}_2$ *los subvectores correspondientes a las componentes continua y discreta, respectivamente, del vector representante* $\mathbf{x}\in\chi\subset R^m$. *Sean* m_1 *y* m_2 *el número de componentes de* $\mathbf{x}_1$ *y* $\mathbf{x}_2$, *por tanto* $m_1+m_2=m$, *y* $\mathbf{x}_1\in\chi_1\subset R^{m_1}$, $\mathbf{x}_2\in\chi_2\subset R^{m_2}$.

Si la distribución de los vectores representantes cumple las condiciones:

C1.

Para todo $\mathbf{x}_2\in\chi_2$, *la distribución de* $\mathbf{x}_1$ *dado* $\mathbf{x}_2$ *es*

absolutamente continua, y la función de densidad $p(.|x_2)$ *es continua.*

C2.

Existen m *vectores linealmente independientes,* $\{x^i;\ i=1,\ldots,m\}$, *tales que* $p(x^i)=p(x_1^i|x_2^i)p(x_2^i)>0$.

C3.

Dado θ_0 *y* θ_1, *puntos interiores de* Θ, *existe* $\varepsilon_0>0$ *tal que la esperanza respecto a* x *de la función*

$$\sum_{\delta=1}^{k} G_{\varepsilon_0}(x,\delta,\theta_1)\ p(\delta|x,\theta_1)$$

existe y es finita. Siendo,

$$G_{\varepsilon_0}(x,\delta,\theta_1)=\sup_{\|\theta-\theta_1\|<\varepsilon} |Log\ p(\delta|x,\theta)-Log\ p(\delta|x,\theta_1)|.$$

C4.

Dado θ_0 *punto interior de* Θ, *existe* $\varepsilon_1>0$ *tal que la esperanza con respecto a* x *de las funciones*

$$\sum_{\delta=1}^{k} m_{ij}(\varepsilon_1,x,\delta,\theta_0)\ p(\delta|x,\theta_0), \quad i,j=1,\ldots,m\times(k-1)$$

existen y son finitas. Siendo

$$m_{ij}(\varepsilon_1,x,\delta,\theta_0)=\sup_{\|\theta-\theta_0\|<\varepsilon} \left| \frac{\partial^2 Log\ p(\delta|x,\theta)}{\partial\theta_i\partial\theta_j} - \frac{\partial^2 Log\ p(\delta|x,\theta_0)}{\partial\theta_i\partial\theta_j} \right|$$

Y si la distribución inicial, $p(\theta)$, *cumple la condición*

C5.

La distribución $p(\theta)$ *es continua y no se anula en ningún punto de* Θ.

Entonces la distribución final converge en distribución a una Normal con media el estimador máximo verosímil, $\hat{\theta}$, *y matriz de precisión la matriz*

$$D^2_\theta(-Log\ V(\hat{\theta}|D))$$

Así como la condición C3, como se verá mas adelante, parece ser necesaria, C4 es una condición relacionada con el comportamiento de la derivada segunda del logaritmo de la función de verosimilitud. Según diversos autores, (ver por ejemplo Le Cam, 1970), las condiciones sobre las derivadas de orden superior del logaritmo de la función de verosimilitud pueden no ser necesarias para la normalidad asintótica del estimador máximo verosímil. En el caso de probarse dicha conjetura, la condición C4 podría no ser necesaria en la demostración del teorema 1. Sin embargo, la demostración aquí presentada necesita de dicha condición.

Las condiciones C1 y C2 implican el siguiente resultado.

PROPOSICION 1

Si se cumplen C1 y C2 entonces, para todo $\varepsilon>0$ *las* m *bolas abiertas* $B_\varepsilon(x^i)$, $i=1,\ldots,m$, *de centro* x^i *y radio* ε, *tienen probabilidad no nula.*

DEMOSTRACION

Sea $p(x_1^i|x_2^i)=\rho>0$. Como la función $p(.|x_2^i)$ es continua en x_1^i, existe un entorno de x_1^i, N_i, tal que $\forall\ y_1 \in N_i$, $p(y_1|x_2^i)>\rho/2>0$.

Dado $\varepsilon>0$ sea τ, $\varepsilon>\tau>0$, tal que la bola $B^*(x_1^i)\subset R^{m_1}$ esté incluída en N_i.

Como la medida de borel de toda bola no vacía es estrictamente positiva, $\mu(B^*(x_1^i))>0$, y como $p(x_2^i)>0$, entonces

$$P(y\in R^m;\ y\in B_\varepsilon(x^i)) \geq P(y\in R^m;\ y\in B_\tau(x^i)) \geq$$

$$\geq P(y;\ y_1\in B_\tau^*(x_1^i)\subset R^{m_1},\ y_2=x_2^i) =$$

$$= P(y_1\in B_\tau^*(x_1^i)\,|\,y_2=x_2^i)\ p(x_2^i) \geq \rho/2\ \mu(B_\tau^*(x_1^i))\ p(x_2^i) >$$

∎

Este resultado, conjuntamente con la independencia lineal de los vectores $\{x^i\}$, implica las condiciones en las que se basa el teorema 2.3.2, de donde se deduce entre otras propiedades, la existencia con probabilidad uno del estimador máximo verosímil (proposición 2.3.7) y la forma asintótica acampanada de la función de verosimilitud (teorema 2.3.2).

Obviamente, para conseguir las condiciones del teorema 2.3.2 no es estrictamente necesario exigir la existencia de funciones de densidad continuas, siendo posible por tanto la búsqueda de condiciones mas débiles que C1 y C2. Sin embargo, se han presentado aquí dichas condiciones en parte para conseguir una mayor claridad en la exposición, pero principalmente por que C1 y C2 no suponen ninguna restricción en aplicaciones. En la práctica siempre se trabaja con modelos probabilísticos que poseen funciones de densidad continuas.

La comprobación de las condiciones de integrabilidad exigidas a la distribución $p(x)$ en el enunciado del teorema 1, condiciones C3 y C4, puede presentar serias dificultades. Sin embargo, en situaciones menos generales que la postulada por el teorema 1, pueden encontrarse condiciones alternativas, quizás mas restrictivas que C3 y C4 pero suficientemente generales, como muestran los siguientes resultados.

COROLARIO 1

Sea $p(\delta|x,\theta)$ un modelo de clasificación regular. Si la distribución inicial, $p(\theta)$, es continua y no se anula en ningún punto, y si la distribución de los vectores representantes, $p(x)$, es discreta finita y tal que existen m vectores linealmente independientes, $\{x_i; i=1,\ldots,m\}$, con $p(x_i)>0$. Entonces, la distribución final converge en distribución a una Normal de media el estimador máximo verosímil, $\hat{\theta}$, y matriz de precisión la matriz $D^2_\theta(-Log\ V(\theta|D))$ calculada en el punto $\hat{\theta}$.

El corolario 1 es evidente a partir del teorema 1. En efecto, $p(\delta|x,\theta)$ toma valores en el intervalo abierto (0,1), luego los valores de $G_\varepsilon(x,\delta,\theta_1)$ y $m_{ij}(\varepsilon,x,\delta,\theta_0)$ son finitos para todo $x\in\chi$. Como χ es un conjunto finito,

$$\max_x G_\varepsilon(x,\delta,\theta_1) = G$$

existe y es finito, por tanto $\mathop{E}_x(G_\varepsilon(x,\delta,\theta_1)) \leq G \leq +\infty$. Un razonamiento similar demuestra que la esperanza de $m_{ij}(\varepsilon,x,\delta,\theta_0)$ también existe y es finita, luego se cumplen las condiciones C3 y C4. Por el enunciado del corolario también se cumplen C1, C2 y C5.

PROPOSICION 2

Sea $p(\delta|x,\theta)$ un modelo de clasificación regular. Si $p(x)$ es tal que existe la esperanza de Log $p(\delta|x,\theta)$ y es finita, entonces se cumple la condición C3 del teorema 1.

DEMOSTRACION

Dado $\theta\in\Theta$, si existe θ^* tal que $p(\delta|x,\theta)\geq p(\delta|x,\theta^*)$ entonces

$$|\text{Log } p(\delta|x,\theta)| = -\text{Log } p(\delta|x,\theta) \leq -\text{Log } p(\delta|x,\theta^*)$$

Como $p(\delta|x,\theta)$ es un modelo de clasificación regular, cumple la condición de monotonicidad P3 (Definición 2.1.1). Por tanto, dado $x\in\chi$ y $\varepsilon>0$, para todo θ se puede encontrar un $\theta^*(x,\varepsilon)$ tal que $p(\delta|x,\theta_1)\geq p(\delta|x,\theta^*)$ para todo θ_1 con $\|\theta-\theta_1\|<\varepsilon$. Por ejemplo:

$$\theta^*_{\delta j} = \theta_{\delta j} - \text{Signo}(x_j) \quad , \quad \theta^*_{ij} = \theta_{ij} + \text{Signo}(x_j) \quad \text{para } i\neq\delta$$

Esta definición de $\theta^*(x,\varepsilon)$ solo depende del vector $x\in R^m$ a través de los signos de sus m componentes. Sea por tanto $A^*(\varepsilon)\equiv\{\theta^*_r(\varepsilon);\ r=1,\ldots,2^m\}$, los 2^m vectores que se pueden formar con las posibles combinaciones de los m signos, entonces $\theta^*(x,\varepsilon)\in A^*(\varepsilon)$ para todo x. En consecuencia, dado $\varepsilon>0$ y dado $\theta\in\Theta$:

$$\sup_{\|\theta-\theta_1\|<\varepsilon} |\text{Log } p(\delta|x,\theta)| = -\inf_{\|\theta-\theta_1\|<\varepsilon} \text{Log } p(\delta|x,\theta) \leq$$

$$\leq -\min_r \text{Log } p(\delta|x,\theta^*_r(\varepsilon)) = \max_r |\text{Log } p(\delta|x,\theta^*_r(\varepsilon))| \leq$$

$$\leq \sum_r |\text{Log } p(\delta|x,\theta^*_r)|$$

Luego:

$$\underset{x}{E}\{G_\varepsilon(x,\delta,\theta_1)\ p(\delta|x,\theta_0)\} \leq \underset{x}{E}\ G_\varepsilon(x,\delta,\theta_1) \leq$$

$$\leq \underset{x}{E}|\text{Log } p(\delta|x,\theta_1)| + \underset{x}{E} \sup_{\|\hat{\theta}-\theta_1\|<\varepsilon} |\text{Log } p(\delta|x,\theta)| \leq$$

$$\leq \underset{x}{E}|\text{Log } p(\delta|x,\theta_1)| + \sum_{r=1}^{2^m} \underset{x}{E}|\text{Log } p(\delta|x,\theta_r^*)| < +\infty$$

ya que, por hipótesis la esperanza de $p(\delta|x,\theta)$ existe y es finita para todo $\theta\in\Theta$. ∎

PROPOSICION 3

Sea $p(\delta|x,\theta)$ el modelo logístico aditivo (ejemplo 2.2.4). Si la distribución sobre los vectores representantes, $p(x)$, es tal que su varianza VAR(x), existe y es finita, entonces $p(x)$ cumple las condiciones C3 y C4 del teorema 1.

DEMOSTRACION

Si $p(\delta|x,\theta)$ sigue un modelo logístico:

$$p(\delta|x,\theta) = \exp(\theta'_{.\delta}x)/\left(1+\sum_{i=1}^{k-1}\exp(\theta'_{.\delta}x)\right)$$

donde $\theta_{.\delta}$ es la columna δ de la matriz θ. Por tanto

$$|p(\delta|x,\theta)| = -\text{Log } p(\delta|x,\theta) = \text{Log}\left(1+\sum_{i=1}^{k-1}\exp(\theta'_{.i}x)\right) - \theta'_{.\delta}x$$

Utilizando que $\text{Log}\sum_{i=1}^{r} a_i \leq (r-1)\text{Log } 2 + \sum_{i=1}^{r}|\text{Log } a_i|$,

desigualdad facilmente demostrable por inducción, partiendo de $\text{Log}(a_1+a_2) \leq \text{Max}(\text{Log } 2a_1, \text{Log } 2a_2) = \text{Log } 2 + \text{Max}(\text{Log } a_1, \text{Log } a_2) \leq \text{Log } 2 + |\text{Log } a_1| + |\text{Log } a_2|$, se obtiene:

$$0 \leq |\text{Log } p(\delta|x,\theta)| \leq k\text{Log } 2 + \sum_{i=1}^{k-1} |\theta'_{.i}x| - \theta'_{.\delta}x$$

para todo $\theta \in \Theta$.

Como $\text{VAR}(x)$ existe y es finita, $E(x)$ y $E(|t'x|)$ también existen y son finitas para todo $t \in R^m$. Por tanto $E|\text{Log } p(\delta|x,\theta)|$ existe y es finita para todo $\theta \in \Theta$. Esto, conjuntamente con la proposición 2, demuestra C3.

Por otra parte,

$$\text{Log } p(\delta|x,\theta) = \text{Log } \left\{ \exp(\theta'_{.\delta}x)\, p(\delta=k|x,\theta) \right\} =$$

$$= \theta'_{.\delta}x + \text{Log } p(\delta=k|x,\theta)$$

Luego:

$$\frac{\partial^2}{\partial\theta_{j_1 i_1}\partial\theta_{j_2 i_2}} \text{Log } p(\delta|x,\theta) = \frac{\partial^2}{\partial\theta_{j_1 i_1}\partial\theta_{j_2 i_2}} \text{Log } p(\delta=k|x$$

$$= -\frac{\partial^2}{\partial\theta_{j_1 i_1}\partial\theta_{j_2 i_2}} \text{Log} \left\{ 1 + \sum_{i=1}^{k-1} \exp(\theta'_{.i}x) \right\} =$$

$$= \begin{cases} -x_{j_1} x_{j_2}\, p(\delta=i_1|x,\theta)\, p(\delta=i_2|x,\theta) & \text{si } i_1 \neq i_2 \\ -x_{j_1} x_{j_2}\, p(\delta=i_1|x,\theta)(1-p(\delta=i_1|x,\theta)) & \text{si } i_1 = i_2 \end{cases}$$

por tanto, debido a que las probabilidades son menores que uno, para todo $\theta \in \Theta$:

$$\left| \frac{\partial^2}{\partial\theta_{j_1 i_1}\partial\theta_{j_2 i_2}} \text{Log } p(\delta|x,\theta) \right| < |x_{j_1} x_{j_2}|$$

Luego,

$$m_{i_1 j_1 i_2 j_2}(\varepsilon, x, \delta, \theta_0) \leq \left| \frac{\partial^2}{\partial\theta_{j_1 i_1}\,\partial\theta_{j_2 i_2}} \operatorname{Log} p(\delta|x,\theta_0) \right| +$$

$$+ \sup_{\|\theta-\theta_0\|<\varepsilon} \left| \frac{\partial^2}{\partial\theta_{j_1 i_1}\,\partial\theta_{j_2 i_2}} \operatorname{Log} p(\delta|x,\theta) \right| < 2|x_{j_1} x_{j_2}| \qquad (2)$$

Como VAR(x) existe y es finita, también existe y es finita $E(|x_i x_j|)$ para todo $i,j=1,\ldots,m$. Esto, conjuntamente con (2), demuestra C4.

∎

Para la demostración del teorema 1 puede utilizarse el teorema 3.2 del apéndice 1. Con tal motivo es necesario comprobar que las condiciones asumidas por el teorema 1 implican las condiciones asumidas por el teorema A1.3.2. (Ver apéndice 1 para una descripción detallada de las mismas).

Algunas de esas condiciones son inmediatas, (CR.1, CR.2, CR.6, CR.10), mientras que CA.3 es precisamente el resultado del teorema 2.3.2. Las demás condiciones se demuestran en las proposiciones siguientes.

PROPOSICION 4 (Condición CR.3)

Si $p(\delta|x,\theta)$ es un modelo de clasificación regular y si $p(x)$ cumple las condiciones C1 y C2 entonces: Dados $\theta_1 \neq \theta_2$, dos puntos cualesquiera de Θ, el conjunto de puntos (δ,x) tales que $p(\delta,x|\theta_1) \neq p(\delta,x|\theta_2)$ tiene probabilidad no nula.

DEMOSTRACION

$$P\{(x,\delta);\ p(x,\delta|\theta_1)\neq p(x,\delta|\theta_2)\} =$$

$$= P\{(x,\delta);\ F_\delta(\theta_1'x)\ p(x) \neq F_\delta(\theta_2'x)\ p(x)\} =$$

$$= P\{(x,\delta);\ F_\delta(\theta_1'x)\neq F_\delta(\theta_2'x)\}$$

Sea $\{x_i;\ i=1,\ldots,m\}$ la base de R^m tal que $P(y\in B_\varepsilon(x_i))>0$, cuya existencia está garantizada por la proposición 1.

Si $\theta_1\neq\theta_2$ entonces debe existir i_0 tal que $\theta_1'x_{i_0}\neq\theta_2'x_{i_0}$, (en caso contrario $\theta_1'X=\theta_2'X$ siendo X la matriz $m\times m$ cuya columna i-ésima es x_i. X es no singular y por tanto $\theta_1=\theta_2$, absurdo).

Por la propiedad P1 de los modelos de clasificación regulares, la función vectorial $F(t)$ es uno a uno, por tanto, para todo $t_1\neq t_2\in R^{k-1}$, existe δ_0 tal que $F_{\delta_0}(t_1)\neq F_{\delta_0}(t_2)$. Luego existe δ_0 tal que $F_{\delta_0}(\theta_1'x_{i_0})-F_{\delta_0}(\theta_2'x_{i_0})\neq 0$. Sin pérdida de generalidad se puede considerar que esa diferencia es estrictamente positiva.

Como las funciones involucradas son continuas, existe $\varepsilon>0$ tal que para todo $y\in B_\varepsilon(x_{i_0})$, $F_{\delta_0}(\theta_1'y)-F_{\delta_0}(\theta_2'y)>0$. Utilizando esto y la proposición 1:

$$P\{(x,\delta);\ F_\delta(\theta_1'x)\neq F_\delta(\theta_2'x)\} > P\{(x,\delta_0);\ F_{\delta_0}(\theta_1'x)\neq F_{\delta_0}(\theta_2'x)\} >$$

$$> P\{(x,\delta_0);\ x\in B_\varepsilon(x_{i_0})\} > 0$$

■

PROPOSICION 5 (Condición CR.4)

Si $p(\delta|x,\theta)$ *es un modelo de clasificación regular entonces, dado* $\varepsilon>0$ *suficientemente pequeño y dado* $\theta_1\in\Theta$, $|Log\ p(\delta,x|\theta)-Log\ p(\delta,x|\theta_1)| \leq G_\varepsilon(\delta,x,\theta_1)$ *para todo* $\theta\in\Theta$ *tal que* $\|\theta-\theta_1\|<\varepsilon$, *con* $\lim_{\varepsilon\to 0} G_\varepsilon(\delta,x,\theta_1)$.

Además, si se cumple C3,

$$\lim_{\varepsilon\to 0} \int \sum_{\delta=1}^{k} G_\varepsilon(\delta,x,\theta_1)\ p(\delta,x|\theta_0)\ d\mu = 0 \qquad \forall\ \theta_0\in\Theta$$

DEMOSTRACION

$$|Log\ p(\delta,x|\theta)-Log\ p(\delta,x|\theta_1)| = |Log\ p(\delta|x,\theta)-Log\ p(\delta|x,\theta_1)| =$$

$$= |Log\ F_\delta(\theta'x)-Log\ F_\delta(\theta_1'x)| \leq$$

$$\leq \sup_{\|\theta-\theta_1\|<\varepsilon} |Log\ F_\delta(\theta'x)-Log\ F_\delta(\theta_1'x)| = G_\varepsilon(x,\delta,\theta_1)$$

Como las funciones $\theta'x$, $F_\delta(.)$ y $Log(.)$ son continuas en todo el rango de definición, entonces $Log\ F_\delta(\theta'x)$ es una función continua y por tanto $\lim_{\varepsilon\to 0} G_\varepsilon(x,\delta,\theta_1)=0$.

Además, si existe la esperanza de $G_{\varepsilon_0}(x,\delta,\theta_1)$, entonces la función $\sum_{\delta=1}^{k} G_{\varepsilon_0}(x,\delta,\theta_1)\ p(\delta,x|\theta_0)$ será integrable. Mas aún, la función positiva $G_\varepsilon(x,\delta,\theta_1)$ es monótona decreciente en relación a ε, luego para todo $\varepsilon<\varepsilon_0$,

$$\sum_{\delta=1}^{k} G_\varepsilon(x,\delta,\theta_1)\ p(\delta,x|\theta_0) \leq \sum_{\delta=1}^{k} G_{\varepsilon_0}(x,\delta,\theta_1)\ p(\delta,x|\theta_0).$$

Por tanto se podrá aplicar el teorema de la Convergencia

Dominada de Lebesgue,

$$\lim_{\varepsilon\to 0} \int \sum_{\delta=1}^{k} G_{\varepsilon}(x,\delta,\theta_1)\ p(\delta|x,\theta_0)\ p(x)\ d\mu =$$

$$= \int \sum_{\delta=1}^{k} \lim_{\varepsilon\to 0} G_{\varepsilon}(x,\delta,\theta_1)\ p(\delta|x,\theta_0)\ p(x)\ d\mu = 0$$

∎

PROPOSICION 6 (Condiciones CA.1 y CA.2)

Si $p(\delta|x,\theta)$ es un modelo de clasificación regular, entonces para todo θ_0, punto interior de Θ, y para todo $\Delta>0$, existe $M_{\Delta}(x,\delta,\theta_0)$ tal que

$$Log\ p(\delta,x|\theta)-Log\ p(\delta,x|\theta_0) < M_{\Delta}(x,\delta,\theta_0)$$

para todo θ tal que $\|\theta\|>\Delta$, y donde

$$\lim_{\Delta\to\infty} \int \sum_{\delta=1}^{k} M_{\Delta}(x,\delta,\theta_0)\ p(\delta,x|\theta_0)\ d\mu \le M(\theta_0) < +\infty$$

DEMOSTRACION

$$\text{Log } p(\delta,x|\theta)-\text{Log } p(\delta,x|\theta_0) = \text{Log } p(\delta|x,\theta)-$$

$$- \text{Log } p(\delta|x,\theta_0) < - \text{Log } p(\delta|x,\theta_0)$$

ya que $p(\delta|x,\theta)<1$. Sea $M_{\Delta}(x,\delta,\theta_0)=-\text{Log } p(\delta|x,\theta_0)$, que es independiente de Δ.

$$\int \sum_{\delta=1}^{k} M_{\Delta}(x,\delta,\theta_0)\ p(\delta,x|\theta_0)\ d\mu =$$

$$= \int \sum_{\delta=1}^{k} \{-\text{Log } p(\delta|x,\theta_0)\}\ p(\delta,x|\theta_0)\ d\mu =$$

$$= - \int p(x) \sum_{\delta=1}^{k} p(\delta|x,\theta_0)\ \text{Log } p(\delta|x,\theta_0)\ d\mu =$$

$$= \int p(x)\ H\{p(\delta|x,\theta_0)\}\ d\mu$$

donde

$$H\{p(\delta|x,\theta_0)\} = -\sum_{\delta=1}^{k} p(\delta|x,\theta_0)\ \text{Log}\ p(\delta|x,\theta_0)$$

es la entropía de la distribución discreta $p(\delta|x,\theta_0)$.

Ahora bien, la entropía de una distribución discreta tiene una cota superior en la entropía de la distribución con probabilidades constantes, (Renyi, 1976, pag. 540), por tanto la constante $M(\theta_0)$ buscada puede definirse

$$M(\theta_0) = \text{Log}\ k < +\infty$$

■

PROPOSICION 7 (Condición CR.7)

Si $p(\delta|x,\theta)$ es un modelo de clasificación regular entonces la matriz $J(\theta_0)$, θ_0 punto interior de Θ, cuyo elemento genérico viene dado por

$$J_{ij}(\theta_0) = \int \sum_{\delta=1}^{k} \left\{\frac{\partial Log\ p(\delta,x|\theta_0)}{\partial\theta_{0i}}\right\}\left\{\frac{\partial Log\ p(\delta,x|\theta_0)}{\partial\theta_{0j}}\right\} p(\delta,x|\theta_0)\ d\mu$$

es definida positiva

DEMOSTRACION

Dado $i=1,\ldots,m(k-1)$, sean $i_1 \leq k-1$ y $i_2 \leq m$, dos enteros tales que $i=(i_1-1)m+i_2$, entonces:

$$\frac{\partial \text{Log}\ p(\delta,x|\theta_0)}{\partial\theta_{0i}} = \frac{\partial}{\partial\theta_{0i}}\{\text{Log}\ F_\delta(\theta_0'x)+\text{Log}\ p(x)\} =$$

$$= \frac{\partial}{\partial \theta_{0i}} \text{Log } F_\delta(\theta_0' x) = x_{i_2} \frac{\partial \text{Log } F_\delta(t)}{\partial t_{i_1}}$$

por tanto:

$$\left\{\frac{\partial \text{Log } p(\delta, x|\theta_0)}{\partial \theta_{0i}}\right\}\left\{\frac{\partial \text{Log } p(\delta, x|\theta_0)}{\partial \theta_{0j}}\right\} = x_{i_2} x_{j_2} \frac{\partial \text{Log } F_\delta(t)}{\partial t_{i_1}} \frac{\partial \text{Log } F_\delta(t)}{\partial t_{j_1}}$$

calculada en el punto $t=\theta_0' x$. Así,

$$J_{ij}(\theta_0) = \int \sum_{\delta=1}^{k} x_{i_2} x_{j_2} \frac{\partial \text{Log } F_\delta(t)}{\partial t_{i_1}} \frac{\partial \text{Log } F_\delta(t)}{\partial t_{j_1}}\Bigg|_{t=\theta_0' x} F_\delta(\theta_0' x)\ p(x)\ d\mu =$$

$$= \int x_{i_2} x_{j_2} G_{i_1 j_1}(\theta_0' x)\ p(x)\ d\mu$$

siendo $G_{i_1 j_1}(\theta_0' x)$, el elemento genérico de la matriz $G(\theta_0' x)$ que es definida positiva, por la propiedad P2 de los modelos de clasificación regulares. Por tanto,

$$J(\theta_0) = \int \{x'x \otimes G(\theta_0' x)\}\ p(x)\ d\mu$$

donde $\otimes$ representa el producto de Kronecker de matrices, esto es:

$$xx' \otimes G = \begin{bmatrix} xx'G_{1\,1} & xx'G_{1\,2} \cdots & xx'G_{1\,k-1} \\ \vdots & \vdots & \vdots \\ xx'G_{k-1\,1} & xx'G_{k-1\,1} \cdots & xx'G_{k-1\,k-1} \end{bmatrix}$$

Sea $a \in \mathbb{R}^{m(k-1)}$ con $a_{i_1 i_2}$, $i_1=1,\ldots,k-1$, $i_2=1,\ldots,m$, el elemento $(i_1-1)m+i_2$ del vector a, y sea a_{i_1} el subvector de a formado por las m componentes $\{a_{i_1 r};\ r=1,\ldots,m\}$. Entonces:

$$a'J(\theta_0)a = \int a'\{xx'\otimes G\}a \quad p(x)\ d\mu$$

siendo

$$a'\{xx'\otimes G\}a = \sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(a_i'x)(a_j'x)\ G_{ij}$$

Como G es definida positiva, si $A(x)\in R^{k-1}$ es el vector con componente i-ésima $A(x)_i=a_i'x$,

$$a'\{xx'\otimes G\}a = 0 \quad \text{si y solo si } A(x) = 0$$

por tanto

$$a'J(\theta_0)a = 0 \iff \int a'\{xx'\otimes F\}a\ p(x)\ d\mu = 0 \iff$$

$\iff A(x) = 0$ para casi todo x $\iff$

$\iff a_1,\ldots,a_{k-1}\in L^{\perp}(x)$ para casi todo x

siendo $L^{\perp}(x)$ el espacio ortogonal al espacio generado por el vector x. Pero esa última expresión solo es posible si $a_i=0$ para todo $i=1,\ldots,k-1$, esto es, si $a=0$, ya que en χ existen m entornos linealmente independientes con probabilidad no nula. En consecuencia, $J(\theta_0)$ es definida positiva. ∎

PROPOSICION 8 (Condición CR.8)

Si $p(\delta|x,\theta)$ es un modelo de clasificación regular y θ_0 es un punto interior de Θ, entonces

$$\int \sum_{\delta=1}^{k} \frac{\partial p(\delta,x|\theta_0)}{\partial\theta_{0i}}\,d\mu = \int \sum_{\delta=1}^{k} \frac{\partial^2 p(\delta,x|\theta_0)}{\partial\theta_{0i}\,\partial\theta_{0j}}\,d\mu = 0, \qquad \forall\, i,j$$

DEMOSTRACION

$$\frac{\partial\, p(\delta,x|\theta_0)}{\partial\theta_{0i}} = p(x)\,\frac{\partial}{\partial\theta_{0i}}\,F_\delta(\theta_0'x) \quad \Rightarrow$$

$$\Rightarrow \sum_{\delta=1}^{k}\frac{\partial\, p(\delta,x|\theta_0)}{\partial\theta_{0i}} = p(x)\,\frac{\partial}{\partial\theta_{0i}}\left\{\sum_{\delta=1}^{k}F_\delta(\theta_0'x)\right\} = 0$$

ya que $\sum F_\delta(t)=1$ para todo t, por tanto:

$$\int \sum_{\delta=1}^{k}\frac{\partial\, p(\delta,x|\theta_0)}{\partial\theta_{0i}}\,d\mu = 0$$

El mismo razonamiento demuestra que la integral de las derivadas segundas también es cero. ∎

PROPOSICION 9 (Condición CR.9)

Sea $p(\delta|x,\theta)$ un modelo de clasificación regular y sea $p(x)$ tal que cumple la condición C4. Si θ_0 es un punto interior de Θ, entonces:

Para todo $\theta\in\Theta$, $\|\theta-\theta_0\|<\varepsilon$, siendo $\varepsilon>0$ suficientemente pequeño,

$$\left|\frac{\partial^2 Log\ p(\delta,x|\theta)}{\partial\theta_i\partial\theta_j} - \frac{\partial^2 Log\ p(\delta,x|\theta_0)}{\partial\theta_{0i}\partial\theta_{0j}}\right| < m_{ij}(\varepsilon,x,\delta,\theta_0)$$

de manera que:

$$\lim_{\delta\to 0}\int \sum_{\delta=1}^{k} m_{ij}(\varepsilon,x,\delta,\theta_0)\ p(\delta,x|\theta_0)\ d\mu = 0$$

DEMOSTRACION

$$\left| \frac{\partial^2 \text{Log } p(\delta,x|\theta)}{\partial\theta_i \partial\theta_j} - \frac{\partial^2 \text{Log } p(\delta,x|\theta_0)}{\partial\theta_{0i} \partial\theta_{0j}} \right| =$$

$$= \left| \frac{\partial^2 \text{Log } F_\delta(\theta'x)}{\partial\theta_i \ \partial\theta_j} - \frac{\partial^2 \text{Log } F_\delta(\theta_0'x)}{\partial\theta_{0i} \ \partial\theta_{0j}} \right| \leq$$

$$\leq \sup_{\|\theta-\theta_0\|<\varepsilon} \left| \frac{\partial^2 \text{Log } F_\delta(\theta'x)}{\partial\theta_i \ \partial\theta_j} - \frac{\partial^2 \text{Log } F_\delta(\theta_0'x)}{\partial\theta_{0i} \ \partial\theta_{0j}} \right| = m_{ij}(\varepsilon,x,\delta,\theta_0)$$

Como las funciones $F_\delta(\theta'x)$ son dos veces continuamente diferenciables, para todo $\tau>0$, existe $\varepsilon_\tau>0$ tal que si $\|\theta-\theta_0\|<\varepsilon_\tau$ entonces:

$$\left| \frac{\partial^2 \text{Log } F_\delta(\theta'x)}{\partial\theta_i \ \partial\theta_j} - \frac{\partial^2 \text{Log } F_\delta(\theta_0'x)}{\partial\theta_{0i} \ \partial\theta_{0j}} \right| < \tau$$

luego $m_{ij}(\varepsilon_\tau,x,\delta,\theta_0)\leq\tau$. Esto es

$$\lim_{\varepsilon\to 0} m_{ij}(\varepsilon,x,\delta,\theta_0) = 0$$

Además, si la distribución $p(x)$ cumple C4, entonces existe ε_1 tal que

$$E_x\left\{ \sum_{\delta=1}^{k} m_{ij}(\varepsilon_1,x,\delta,\theta_0) \ F_\delta(\theta_0'x) \right\} < +\infty$$

por lo que se puede aplicar el teorema de la Convergencia Dominada de Lebesgue, obteniéndose:

$$\lim_{\varepsilon\to 0} \int \sum_{\delta=1}^{k} m_{ij}(\varepsilon,x,\delta,\theta_0) \ p(\delta,x|\theta_0) \ d\mu =$$

$$= \int \sum_{\delta=1}^{k} \lim_{\varepsilon\to 0} m_{ij}(\varepsilon,x,\delta,\theta_0) \ p(\delta,x|\theta_0) \ d\mu = 0$$

∎

3.2 CAMBIOS DE LOCALIZACION Y ESCALA EN EL VECTOR REPRESENTANTE

La aplicación de los resultados obtenidos en el apartado anterior requiere el cálculo tanto del estimador máximo verosímil como de la matriz de derivadas segundas parciales del logaritmo de la función de verosimilitud. Sin embargo, el estimador máximo verosímil para un modelo de clasificación regular no tiene forma cerrada, siendo necesario el uso de métodos iterativos para encontrar una solución de la ecuación de verosimilitud, $D^1_\theta(\text{Log } V(\theta|D))=0$. En particular, se suelen obtener buenos resultados con el método de Newton-Raphson, que no solamente proporciona el máximo buscado sino también, como un paso intermedio, la matriz $D^2_\theta(-\text{Log } V(\hat{\theta}|D))$, matriz de precisión de la aproximación Normal asintótica juntificada por el teorema 3.1.1.

EJEMPLO 1

Considérese el modelo logístico aditivo. Esto es,

$$\text{Log } p(\delta|x,\theta) = \theta'_{.\delta}x + \text{Log } p(k|x,\theta) \quad \text{si } \delta=1,\ldots,k-1$$

$$\text{Log } p(k|x,\theta) = -\text{Log}\left\{1+\sum_{j=1}^{k-1}\exp(\theta'_{.j}x)\right\}$$

El vector de derivadas parciales es $D^1_\theta(\text{Log } p(\delta|x,\theta))$, con elemento genérico $(i-1)\times(k-1)+j$,

$$\frac{\partial}{\partial\theta_{ij}}\text{Log } p(\delta|x,\theta) = x_i\,(1-p(\delta|x,\theta)) \quad \text{si } j=\delta$$

$$= x_i\,p(j|x,\theta) \quad \text{si } j\neq\delta$$

La matriz de derivadas segundas, $D^2_\theta(\text{Log } p(\delta|x,\theta))$, tiene por elemento genérico $((i_1-1)(k-1)+j_1)\times((i_2-1)(k-1)+j_2)$ a:

$$\frac{\partial^2}{\partial\theta_{i_1 j_1}\partial\theta_{i_2 j_2}} \text{Log } p(\delta|x,\theta) = x_{i_1} x_{i_2} \; p(j|x,\theta)(1-p(j|x,\theta)) \quad \text{si } j_1=j_2=j$$

$$= -x_{i_1} x_{i_2} \; p(j_1|x,\theta)\; p(j_2|x,\theta) \quad \text{si } j_1 \neq j_2$$

Por tanto, el vector $D^1_\theta(\text{Log } V(\theta|D))$ y la matriz $D^2_\theta(\text{Log } V(\theta|D))$ son fácilmente programables.

El método de Newton-Raphson, (ver por ejemplo, Ralston 1970, apartado 8.8), consiste en calcular en la i-ésima iteración,

$$\theta^{(i)} = \theta^{(i-1)} - D^2_\theta(\text{Log } V(\theta^{(i-1)}|D)) \times \times D^1_\theta(\text{Log } V(\theta^{(i-1)}|D))$$

El cálculo del estimador maximo verosímil en los paquetes de programas usuales, BMDP, SPSS, etc..., se realiza mediante este algoritmo. Para el cálculo de los ejemplos numéricos empleados en esta memoria, apartado 3.3 y capítulo 5, se ha utilizado el algoritmo de Newton-Raphson partiendo de un valor inicial $\theta^{(0)}=0$. La regla de parada empleada consiste en una comparación entre los vectores $\theta^{(i)}$ y $\theta^{(i-1)}$ mediante la distancia euclídea. Así, si $||\theta^{(i)}-\theta^{(i-1)}||<0.01$ el algoritmo se detiene, mientras que si la distancia euclídea es mayor, el algoritmo comienza una nueva iteración. El valor

crítico 0.01 fue elegido después de comparar la relación número de iteraciones / bondad del resultado para distintos valores críticos.

Reglas de parada alternativas pueden definirse por comparación de la función de verosimilitud en los puntos $\theta^{(i)}$ y $\theta^{(i-1)}$, esto es $V(\theta^{(i)}|D)$ y $V(\theta^{(i-1)}|D)$; sin embargo, al variar el número de datos, n, la función de verosimilitud cambia de escala considerablemente; el valor máximo de $V(\theta|D(n))$ disminuye al aumentar n. En consecuencia, la regla de parada debe definirse a través del porcentaje de aumento, o lo que es equivalente, a traves del cociente $V(\theta^{(i)}|D)\ /\ V(\theta^{(i-1)}|D)$. La interpretación de este cociente es bastante menos intuitiva que la proporcionada por la distancia euclídea $\|\theta^{(i)}-\theta^{(i-1)}\|$, por ello se eligió la distancia euclídea como regla de parada.

■■

En la aplicación de métodos numéricos es conveniente que los datos sean lo mas homogéneo posible; con ello se puede conseguir una disminución de los errores de redondeo y, en ocasiones, un aumento de la velocidad de cálculo. Por estas razones, mimetizando un método de trabajo frecuente en numerosas áreas de cálculo numérico, resulta aconsejable utilizar el siguiente algoritmo para el cálculo del estimador maximo verosímil:

Paso 1: *Tipificación* de los vectores representantes incluídos en el banco de datos, utilizando para ello la media y varianza muestrales. Así, cada vector representante, $x=(x^1,\ldots,x^m)'$, se transforma en un nuevo vector, $y=(y^1,\ldots,y^m)'$

mediante el cambio

$$y^j = (x^j - \bar{x}_j)/s_j .$$

Siendo $\bar{x}_j$ y s_j la media y varianza muestrales de las componentes j-ésimas de los vectores representantes.

Paso 2: Reparametrización de la función de verosimilitud obteniendo

$$V(\psi|D) = \prod_{i=1}^{n} p(\delta_i|\psi, y_i) .$$

Cálculo de la distribución asintótica final de ψ.

Paso 3: Utilizando las fórmulas de cambio de variables, obtener la distribución asintótica final del parámetro original, θ, a partir del resultado del Paso 2.

El paso 2 en el algoritmo anterior tiene sentido puesto que todo cambio de localización y escala de los vectores representantes induce una transformación lineal en el espacio paramétrico. En efecto, todo cambio de localización y escala,

$$y^j = a_j(x^j + b_j) \qquad j=1,\ldots,m$$

corresponde a una transformación lineal; es mas, si en el vector representante se ha considerado un término constante, $x^1=1$ para todo $x \in \chi$, entonces, el cambio de localización y escala puede expresarse como $y=Mx$, siendo

M la matriz:

$$M \equiv \begin{pmatrix} a_1(1+b_1) & 0 & 0 & \dots & 0 \\ a_2 b_2 & a_2 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ a_m b_m & 0 & 0 & \dots & a_m \end{pmatrix}$$

M es no singular si $b_1 \neq -1$ y $a_j \neq 0$ para todo $j=1,\dots,m$. En particular para $a_1=1$, $b_1=0$, $a_j=1/s_j$ y $b_j=\bar{x}_j$, transformación efectuada en el Paso 1, la matriz M es no singular. En consecuencia existe M^{-1} con lo que $\theta' x = \theta' M M^{-1} x = \theta' M y$. Sea $\psi = M'\theta$, entonces:

$$V(\theta|D) = \prod_{i=1}^{n} p(\delta_i|\theta, x_i) = \prod_{i=1}^{n} F_{\delta i}(\theta' x_i) =$$

$$= \prod_{i=1}^{n} F_{\delta i}(\psi' y_i) = V(\psi|D).$$

Esa es la reparametrización exigida en el Paso 2.

Aplicando el teorema 3.1.1, la distribución final de ψ, si el tamaño muestral n, es suficientemente grande, es aproximadamente Normal con media $\hat{\psi}$ y matriz de precisión

$$D^2_{\psi}(-\text{Log } V(\hat{\psi}|D)).$$

Esto es,

$$p(\psi|D) \cong N(\psi|\hat{\psi}, D^2_{\psi}(-\text{Log } V(\hat{\psi}|D))).$$

Por otra parte, la expresión que relaciona las ma-

trices de parámetros ψ y θ es $\theta=(M')^{-1}\psi$. Expresados los parámetros en forma vectorial, esa igualdad se convierte en $\theta=(I\otimes(M')^{-1})\psi$. Sea H el producto de Kronecker $I\otimes(M')^{-1}$, como toda transformación lineal de una cantidad aleatoria Normal también es Normal, la distribución final de θ es:

$$p(\theta|D) \cong N(\theta|H\hat{\psi},(M')^{-1}D^2_{\psi}(-\text{Log } V(\hat{\psi}|D))H^{-1}).$$

La siguiente proposición demuestra que el mismo resultado se obtendría al aplicar directamente el teorema 3.1.1 a la verosimilitud $V(\theta|D)$.

PROPOSICION 1

Con la notación anterior, $\hat{\theta} = H\hat{\psi}$, *y*

$$D^2_{\theta}(Log\ V(\theta|D)) = (H')^{-1}D^2_{\psi}(Log\ V(\psi|D))H^{-1}$$

DEMOSTRACION

Por la regla de la cadena, como $\psi=H^{-1}\theta$,

$$D^1_{\theta}(\text{Log } V(\theta|D)) = D^1_{\psi}(\text{Log } V(\psi|D))\ D^1_{\theta}(\psi) =$$

$$= D^1_{\psi}(\text{Log } V(\psi|D))\ H^{-1}$$

además, al ser H no singular, las soluciones a las ecuaciones $D^1_{\theta}(\text{Log } V(\theta|D))=0$ y $D^1_{\psi}(\text{Log } V(\psi|D))=0$ coinciden. Por tanto, los estimadores maximo verosímiles $\hat{\theta}$ y $\hat{\psi}$ también están ligados por la ecuación $\hat{\theta}=H\hat{\psi}$.

Por otra parte, una nueva aplicación de la regla de la cadena demuestra que,

$$D^2_\theta(\text{Log } V(\theta|D)) = D^1_\theta\{D^1_\psi(\text{Log } V(\psi|D))H^{-1}\} =$$

$$= D^1_{D^1_\psi(\text{Log } V(\psi|D))}\{D^1_\psi(\text{Log } V(\psi|D))H^{-1}\}D^2_\psi(\text{Log } V(\psi|D))D^1_\theta(\psi) =$$

$$= (H')^{-1}D^2_\psi(\text{Log } V(\psi|D))H^{-1}$$

■

Esta proposición comprueba la validez teórica del algoritmo anterior cuando en el vector representante se incluye un término constante.

Por el contrario, no considerar un término constante como parte del vector representante es equivalente a suponer que ese término si ha sido incluído, pero que los parámetros correspondientes, θ_{1i}, $i=1,\ldots,k-1$, son conocidos e iguales a cero. Estas restricciones sobre los parámetros θ se traducen, tras el cambio de localización y escala de los vectores representantes, en las restricciones sobre los parámetros ψ:

$$a_1(1+b_1)\psi_{1i} + \sum_{j=2}^{m} a_j b_j \psi_{ji} = 0, \quad i=1,\ldots,k-1$$

Sigue existiendo una transformación uno a uno entre los parámetros θ y ψ, por tanto los resultados comentados anteriormente siguen siendo válidos. Sin embargo, todas las constantes $\{a_i\}$, $\{b_i\}$ aparecen ahora en la función $V(\psi|D)$ ya que

$$\psi_{1i} = -(a_1(1+b_1))^{-1} \sum_{j=2}^{m} a_j b_j \psi_{ji}$$

con lo que las ventajas computacionales de maximizar $V(\psi|D)$ en lugar de $V(\theta|D)$ desaparecen. En consecuencia, si no se introduce un término independiente, es mas aconsejable el cálculo de la distribución asintótica final de θ directamente a partir de $V(\theta|D)$.

3.3 EJEMPLOS NUMERICOS

Considérese el modelo logístico para dos clases, $k=2$, con vector representante, x, formado por un término constante e igual a uno y un indicador continuo, y:

$$p(\delta=1|y,\theta) = 1-p(\delta=2|y,\theta) = \exp(\theta_0+\theta_1 y)/(1+\exp(\theta_0+\theta_1 y)) \quad (1)$$

La sencillez de este modelo, involucrando únicamente dos parámetros, posibilita el uso de técnicas de análisis numérico para el cálculo de sus características. Por ese motivo ha sido escogido para el estudio, a través de bancos de datos simulados, de la rapidez de convergencia de la distribución final a su aproximación asintótica.

Una vez obtenido un banco de datos simulado, conociendo por tanto el verdadero valor del parámetro θ, se calcula la distribución normal asintótica, $N(\theta|\hat{\theta},\hat{H})$, siendo $\hat{\theta}$ el estimador maximo verosímil, y $\hat{H}$ la matriz de precisión, inversa de la matriz de varianzas, estimador usual de la precisión asintótica. Existen diferentes medidas de la bondad del resultado, en este apartado se utilizan las siguientes:

En primer lugar, la *distancia euclídea*,

$$DIS_E(\theta,\hat{\theta}) = \|\theta-\hat{\theta}\| = \left((\theta_0-\hat{\theta}_0)^2+(\theta_1-\hat{\theta}_1)^2\right)^{\frac{1}{2}}$$

calculada para diversos tamaños muestrales, debe converger a cero ya que $\hat{\theta}$ es un estimador consistente. Por tanto, DIS_E puede utilizarse para medir la bondad del

estimador media asintótica.

La matriz $\hat{H}$ puede contrastarse utilizando la *distancia de Mahalanobis.*

$$DIS_M(\theta,\hat{\theta}) = \left((\theta-\hat{\theta})' \hat{H}(\theta-\hat{\theta})\right)^{\frac{1}{2}}$$

En efecto, la distancia de Mahalanobis es la distancia euclídea al origen de la transformación tipificadora usual aplicada al punto θ; esto es, si $\theta \sim N(\hat{\theta},\hat{H})$ entonces $\hat{H}^{\frac{1}{2}}(\theta-\hat{\theta}) \sim N(0,I)$, con lo que

$$DIS_M(\theta,\hat{\theta}) = DIS_E(\hat{H}^{\frac{1}{2}}(\theta-\hat{\theta}),0)$$

Si la distribución $N(\hat{\theta},\hat{H})$ es una buena aproximación de la distribución final entonces, el cuadrado de DIS_M debe seguir una distribución Chi-cuadrado con tantos grados de libertad como número de elementos tenga el vector paramétrico θ. (ver, por ejemplo, John, 1971, teorema 2, pag. 30).

Sin embargo, en un problema de clasificación, el énfasis debe situarse en la distribución de clasificación en lugar de situarlo en la distribución final sobre los parámetros del modelo. Ninguna de las distancias anteriores mide la bondad de la distribución predictiva obtenida a partir de la distribución asintótica.

Una vez obtenida la distribución asintótica, la distribución predictiva para el modelo (1), $p(\delta|y,D)$, es fácilmente calculable mediante integración numérica, ya que solo depende de la combinación lineal $\theta_0+\theta_1 y$

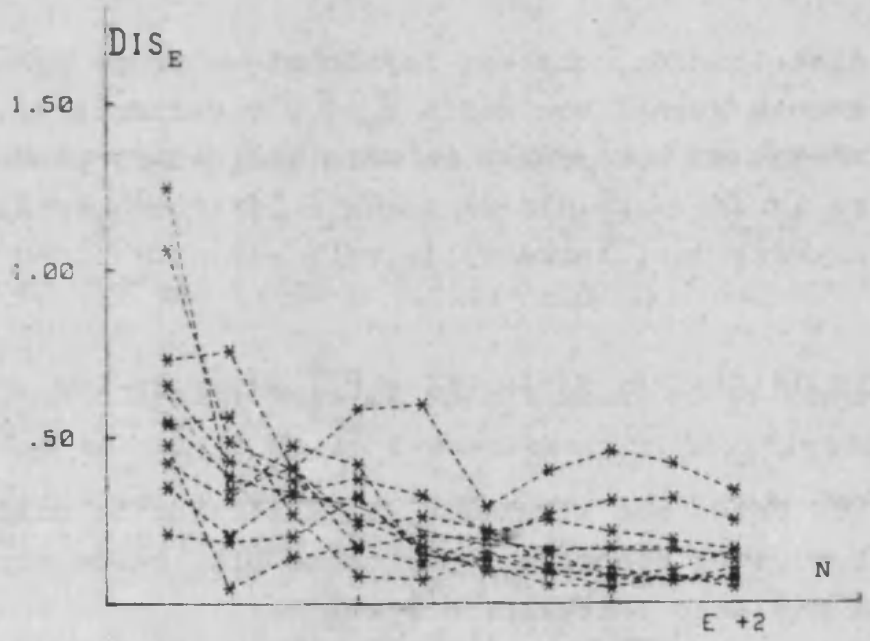
cuya distribución, una vez fijado el valor de y, es asintóticamente Normal con media $\hat{\theta}_0+\hat{\theta}_1 y$ y varianza $(1,y)\ \hat{H}^{-1}(1, y)'$. Además al considerar un solo indicador, puede calcularse la *discrepancia esperada* o *distancia esperada de Kullback-Liebler*, entre $p(\delta|y,\theta)$ y $p(\delta|y,D)$,

$$DIS_K(p(\delta|y,\theta),\ p(\delta|y,D)) = \underset{y}{E} \sum_{\delta} p(\delta|y,\theta)\ \text{Log}\ \frac{p(\delta|y,\theta)}{p(\delta|y,D)}$$

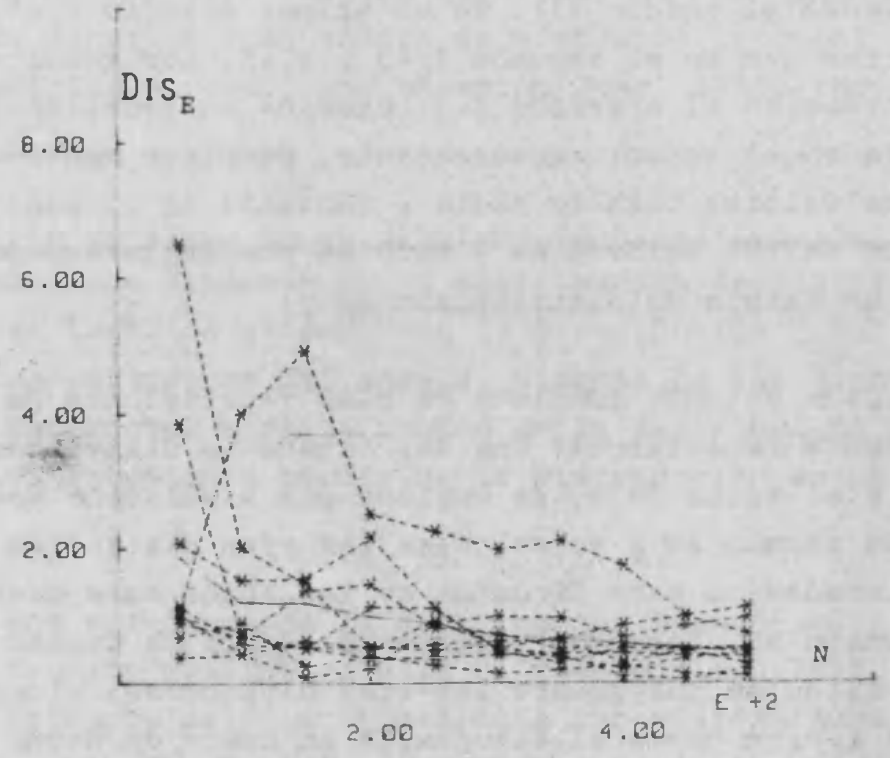
Como los datos son simulados, la distribución unidimensional $p(y)$ es conocida y por tanto DIS_K puede ser calculada mediante integración numérica.

Estas tres distancias han sido calculadas en dos estudios paralelos, ambos estudios se realizan con muestras simuladas obtenidas a partir de $p(y)\sim N(y|0,1)$ y $p(\delta|y,\theta)$ siguiendo el modelo (1). En el primer estudio $\theta_0=0$, $\theta_1=1$; mientras que en el segundo $\theta_0=3$ y $\theta_1=5$. Los resultados obtenidos en el apartado 3.2, cambios de localización y escala en el vector representante, permiten mantener los mismos valores para la media y varianza de y, pues un cambio en los valores de θ también pueden interpretarse como un cambio de distribución de y.

Cada estudio consiste en diez repeticiones de la siguiente experiencia: Una vez fijada la distribución $p(y)$ y el valor de θ, se obtiene por simulación una muestra de tamaño 50 y se calculan las tres distancias antes mencionadas; a esos 50 datos se les añade otra muestra de tamaño 50, formando un banco de datos de tamaño 100, y se calculan nuevamente las tres distancias; el proceso se repite hasta el estudio de un banco de datos de tamaño 500.



- Fig. 1 -
-Distancia euclídea con $\theta_0=0$, $\theta_1=1$-

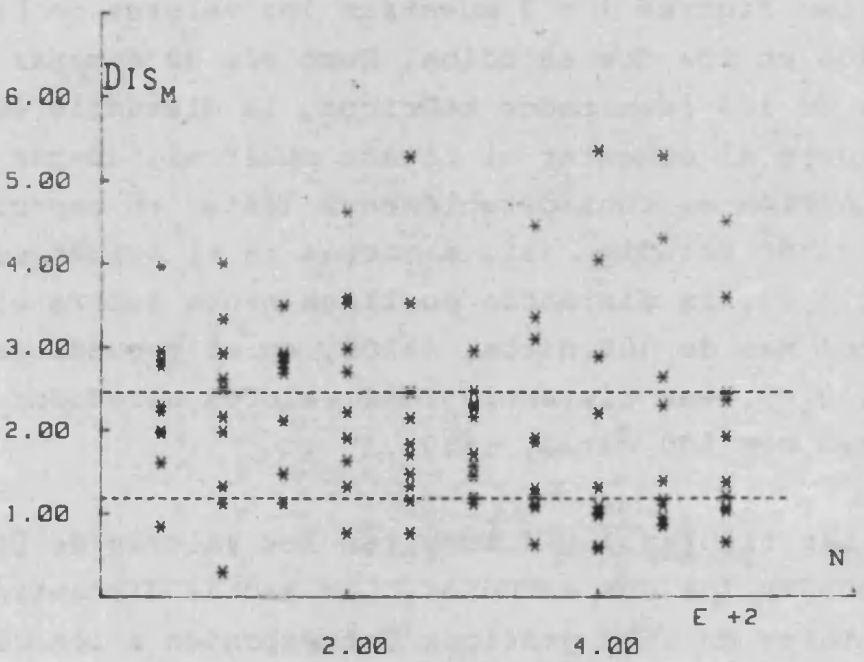


- Fig. 2 -
-Distancia euclídea con $\theta_0=3$, $\theta_1=5$-

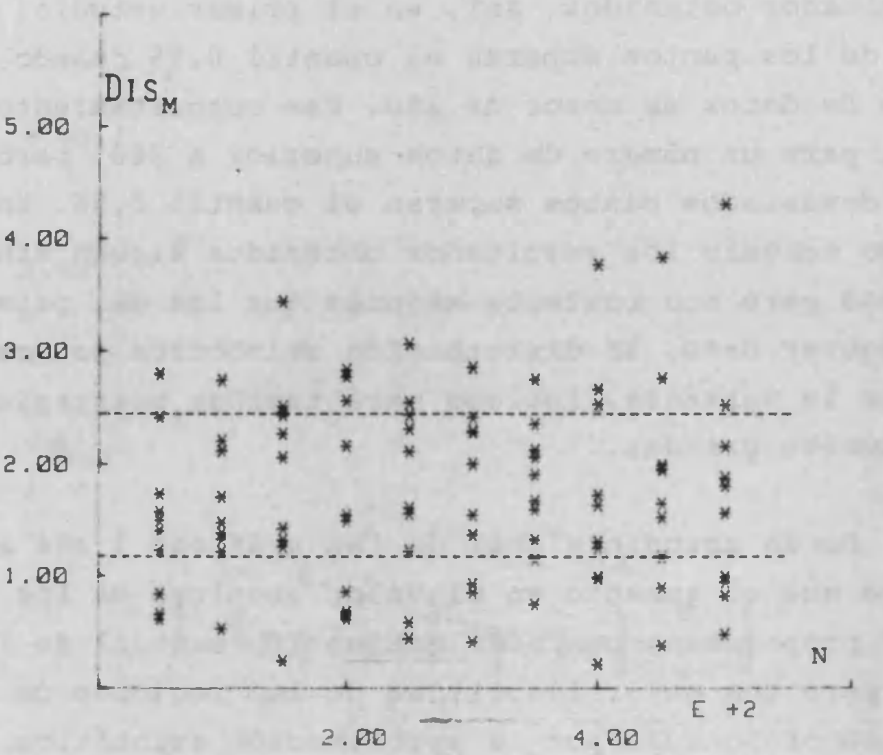
Las figuras 1 y 2 muestran los valores de DIS_E obtenidos en los dos estudios. Como era de esperar a la vista de los resultados teóricos, la distancia euclídea disminuye al aumentar el tamaño muestral, aunque esa disminución es considerablemente lenta, en especial en el segundo estudio. Así, mientras en el primer caso, $\theta_0=0$, $\theta_1=1$, la distancia euclídea nunca supera el valor 0.5 con mas de 300 datos, $n\geq 300$, en el segundo caso, $\theta_0=3$, $\theta_1=5$, esa distancia toma valores alrededor de 1 incluso con 500 datos, $n=500$.

Las figuras 3 y 4 muestran los valores de DIS_M obtenidos en los dos estudios. Las rectas discontinuas horizontales en esas gráficas corresponden a los cuantiles de orden 0.5 y 0.95 de DIS_M obtenidas a partir de la distribución Chi-cuadrado con dos grados de libertad. Estos cuantiles teóricos no se corresponden con los resultados obtenidos. Así, en el primer estudio, la mitad de los puntos superan el cuantil 0.95 cuando el número de datos es menor de 250. Ese comportamiento se mejora para un número de datos superior a 300, pero todavía demasiados puntos superan el cuantil 0.95. En el segundo estudio los resultados obtenidos siguen sin ser buenos pero son bastante mejores que los del primero. En cualquier caso, la distribución asintótica parece subestimar la varianza, incluso para tamaños muestrales relativamente grandes.

De un estudio global de las gráficas 1 a 4 se infiere que un aumento en el valor absoluto de los parámetros proporciona una peor estimación puntual de los mismos pero una mayor fiabilidad de las regiones de confianza proporcionadas por la aproximación asintótica.



- Fig. 3 -
-Distancia de Mahalanobis con $\theta_0=0$, $\theta_1=1$-



- Fig. 4 -
-Distancia de Mahalanobis con $\theta_0=3$, $\theta_1=5$-

Los valores de discrepancia son bastante menos intuitivos que los proporcionados por las distancias euclídea y de Mahalanobis. Por ello puede resultar interesante la realización de un pequeño estudio sobre las implicaciones de los diversos valores de discrepancia antes de comentar los resultados obtenidos con esta distancia.

Dado un valor para el indicador y, la distribución de clasificación teórica en los modelos aquí estudiados es

$$p(\delta=1|y,\theta) = \exp(\theta_0+\theta_1 y)/(1+\exp(\theta_0+\theta_1 y)) = p$$

$$p(\delta=2|y,\theta) = 1-p$$

mientras que la proporcionada por los resultados asintóticos es

$$p(\delta=1|y,D) = \underset{\theta|D}{E}\{p(\delta=1|y,\theta)\} = p+\varepsilon$$

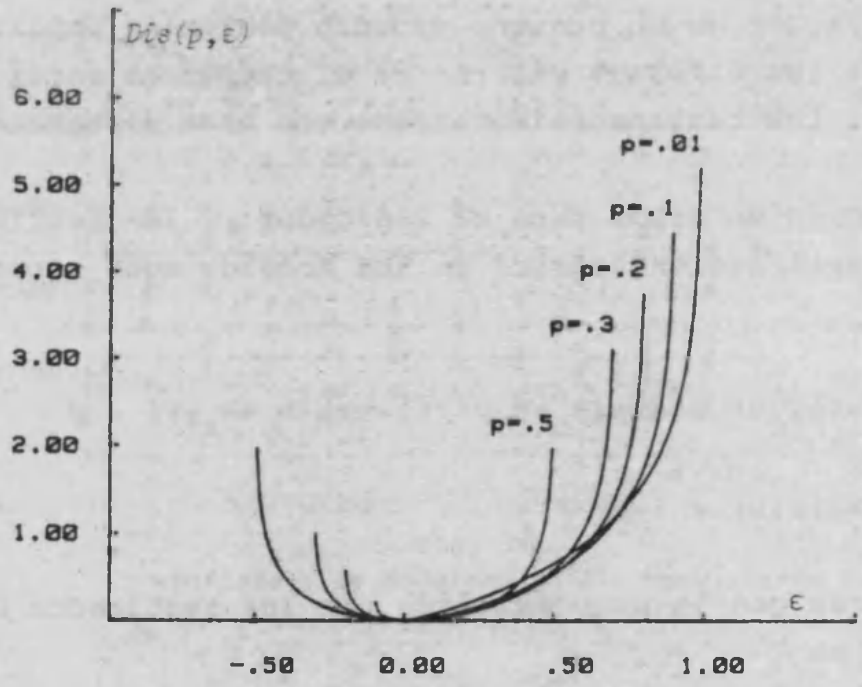
$$p(\delta=2|y,D) = 1-(p+\varepsilon)$$

siendo ε un número real perteneciente al intervalo $(-p, 1-p)$. La discrepancia entre estas dos funciones es

$$Dis(p,\varepsilon) = p \text{ Log } \frac{p}{p+\varepsilon} + (1-p) \text{ Log } \frac{1-p}{1-(p+\varepsilon)}$$

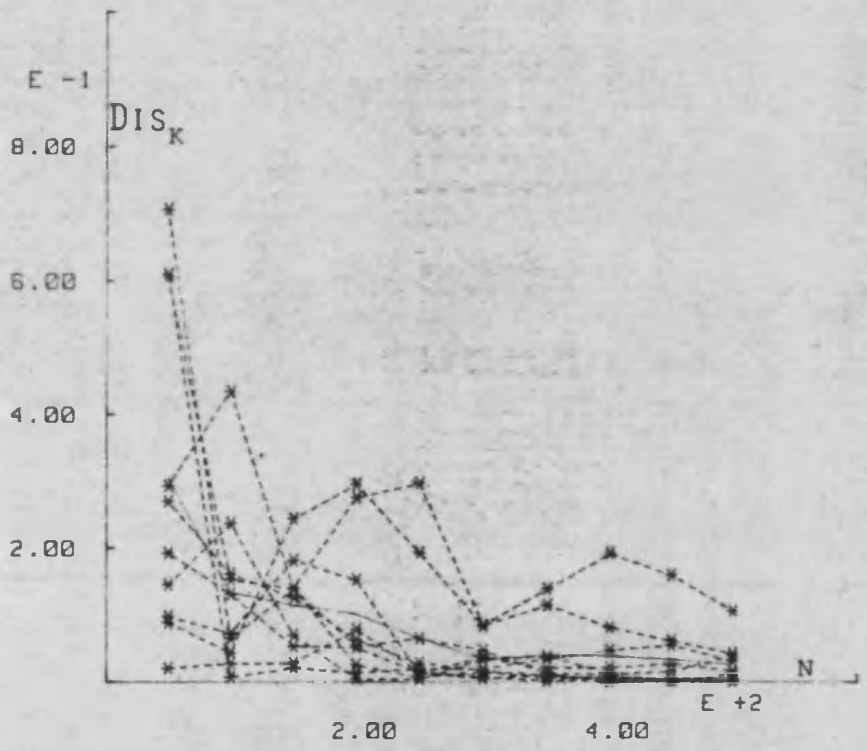
En la figura 5 se muestran los valores de $Dis(p,\varepsilon)$, como función de ε, para cinco valores distintos de p, 0.01, 0,1, 0.2, 0.3 y 0.5. Obviamente, por la simetría de la función $Dis(p,\varepsilon)$, su gráfica es la imagen especular de la gráfica correspondiente a $Dis(1-p,\varepsilon)$, por ello

solo se muestran valores de p no mayores a 0.5.

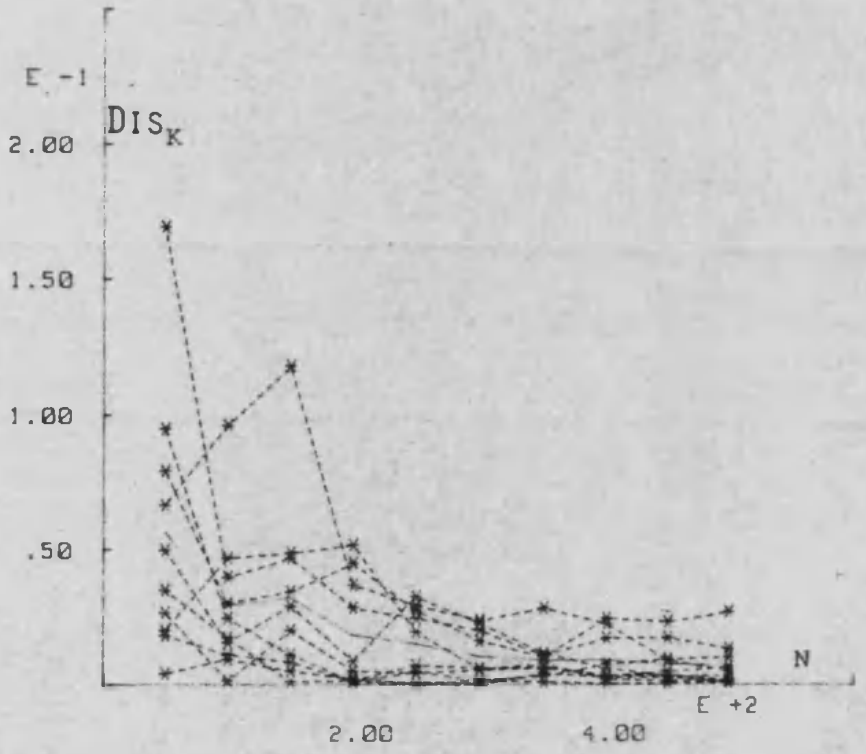


-fig. 5-
-Estudio de la discrepancia-

Las figuras 6 y 7 muestran los valores de DIS_K obtenidos en los dos estudios. En el primero, la discrepancia solo supera el valor 0.5 en tres casos, todos ellos con tamaño muestral 50, siendo menor a 0.1 en casi todos los puntos con 300 o mas datos. En el segundo estudio la discrepancia se comporta todavía mejor, no superando nunca el valor 0.2 y siendo inferior a 0.05 en todos los puntos obtenidos con un tamaño muestral superior a 200, $n \geq 250$.



- Fig. 6 -
-Discrepancia con $\theta_0=0$, $\theta_1=1$-



- Fig. 7 -
-Discrepancia con $\theta_0=3$, $\theta_1=5$-

CAPITULO 4

PROCESO INFERENCIAL PARA MODELOS DE CLASIFICACION REGULARES

En este capítulo se pormenorizan las etapas que constituyen la estructura bayesiana en el enfoque clasificatorio. En particular, dicha estructura se ejemplifica mediante el estudio detallado del modelo Normal Acumulado para dos clases.

El objetivo de la segunda parte del capítulo es la búsqueda de distribuciones de referencia. Se obtiene el resultado general y se proponen aproximaciones para que este resultado teórico, de cálculo difícil, pueda ser utilizado en aplicaciones. Se estudia con detalle las distribuciones de referencia correspondientes a los modelos de tipo 1 y tipo 2 obteniendo, como ejemplos, las distribuciones de referencia de los modelos logístico aditivo y logístico multiplicativo.

4.1 INFERENCIA Y PREDICCION

Desde un enfoque clasificatorio, la metodología bayesiana conlleva un proceso bietápico para la obtención de la distribución de clasificación. En efecto, como ya se discutió en el apartado 1.2, la distribución de clasificación es el resultado del proceso:

$$p(\delta|x,D) = \int p(\delta|x,\theta)\ p(\theta|D)\ d\theta$$

$$p(\theta|D) \propto V(\theta|D)\ p(\theta)$$

donde $p(\theta)$ es la distribución inicial sobre el parámetro desconocido $\theta\in\Theta$; $V(\theta|D)$ es la función de verosimilitud correspondiente al modelo $p(\delta|x,\theta)$; $p(\theta|D)$ es la distribución final sobre $\theta\in\Theta$, y por último, $p(\delta|x,D)$ es la distribución de clasificación buscada.

El cálculo de la distribución final $p(\theta|D)$ no es el objetivo del problema de clasificación, sino solamente un paso intermedio necesario. El énfasis de todo el proceso debe situarse, por tanto, en el cálculo de la distribución predictiva $p(\delta|x,D)$.

En el capítulo anterior se ha demostrado que si el número de datos, n, es suficientemente grande entonces la distribución final, $p(\theta|D)$, puede aproximarse razonablemente bien mediante la distribución Normal $N(\theta|\hat{\theta},H(\hat{\theta}))$ En este caso,

$$p(\delta|x,D) = \int p(\delta|x,\theta)\ \ N(\theta|\hat{\theta},\ H(\hat{\theta}))\ d\theta$$

Por el contrario, si el número de datos no es sufi-

ciente y/o se desea introducir información inicial sobre el parámetro desconocido $\theta\in\Theta$, esa aproximación asintótica no será válida. En este caso será necesario especificar la distribución inicial $p(\theta)$ antes de realizar las dos etapas que conducen al cálculo de la distribución de clasificación.

En el siguiente ejemplo se estudia el modelo Normal Acumulado. La distribución de clasificación para ese modelo se obtiene: primero, utilizando la aproximación normal asintótica a la distribución final; después, utilizando una distribución normal para representar las opiniones iniciales del experto sobre el parámetro $\theta\in\Theta$, $p(\theta) = N(\theta|\mu,H)$.

EJEMPLO 1. *Modelo Normal Acumulado para 2 clases.*

En el apartado 2.2 se definió el modelo de clasificación Normal-Acumulado para dos clases, $k=2$, como:

$$p(\delta=1|\theta,x) = \Phi(\theta'x) = \int_{-\infty}^{\theta'x} (2\pi)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}z^2\}dz =$$

$$= \int_{-\infty}^{0} (2\pi)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(y+\theta'x)^2\}dy$$

donde θ es un vector de dimensión igual al número de componentes del vector representante, esto es m. De hecho $\theta\in\Theta\equiv R^m$.

Si la cantidad aleatoria δ que representa a las clases se define de forma que tome los valores +1 y -1, correspondientes a las clases 1 y 2 respectivamente, entonces:

$$p(\delta|\theta,x) = \int_{-\infty}^{0} (2\pi)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(y+\delta\theta'x)^2\}\,dy, \quad \text{con } \delta=\pm 1 \qquad (1)$$

En esta notación, la función de verosimilitud es:

$$V(\theta|D) \propto \prod_{i=1}^{n} p(\delta_i|\theta,x_i) =$$

$$= \prod_{i=1}^{n} \int_{-\infty}^{\delta_i \theta' x_i} (2\pi)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}z_i^2\} dz_i$$

$$\propto \Phi_n \ (Diag\ X'\theta) \qquad (2)$$

donde $\Phi_n(Diag\ X'\theta)$ representa la función de distribución Normal n-dimensional de media 0 y matriz de precisión la identidad, calculada en el punto $Diag\ X'\theta$; $Diag$ es una matriz diagonal $n\times n$ tal que $(Diag)_{ii} = \delta_i$, y X es una matriz $m\times n$ cuya columna i-ésima es el vector representante x_i.

Como resultado del teorema 3.1.1, la distribución final puede ser aproximada, si el número de datos, n, es suficientemente grande, mediante la distribución Normal $N(\theta|\hat{\theta}, H(\hat{\theta}))$. Donde $\hat{\theta}$ es el estimador maximo verosímil y $H(\hat{\theta})$ es el estimador asintótico usual de la matriz de precisión, inversa de la matriz de covarianzas.

El algoritmo para obtener $\hat{\theta}$ y $H(\hat{\theta})$ por el método de Newton-Raphson puede expresarse como (Aitchison and Lauder, 1979):

$$\theta^{(r+1)} = \left(H^{(r+1)}\right)^{-1} \left(\sum_{i=1}^{n} w_i^{(r)} y_i^{(r)} x_i\right)$$

$$w_i^{(r)} = \phi^2(x_i'\ \theta^{(r)}) / \{\Phi(x_i'\ \theta^{(r)})\ (1-\Phi(x_i'\ \theta^{(r)}))\}$$

$$y_i^{(r)} = x_i' \theta^{(r)} + \{(\delta_i + 1)/2 - \Phi(x_i' \theta^{(r)})\} / \phi(x_i' \theta^{(r)})$$

$$H^{(r+1)} = \sum_{i=1}^{n} w_i^{(r)} x_i x_i'$$

donde $\phi(.)$ y $\Phi(.)$ son, respectivamente, las funciones de densidad y de distribución de una Normal unidimensional, N(0,1).

Utilizando la aproximación asintótica, la distribución predictiva puede ser calculada de forma analítica, obteniéndose (Aitchison and Lauder, 1979):

$$p(\delta | x, D) = \int p(\delta | x, \theta)\ p(\theta | D)\ d\theta =$$

$$= \Phi(\delta x' \hat{\theta} / \sqrt{1 + x' H^{-1}(\hat{\theta}) x}) \qquad (3)$$

Por el contrario si se desea introducir información inicial sobre los parámetros del modelo, será necesario especificar una distribución inicial, $p(\theta)$, sobre el espacio paramétrico Θ. Si las opiniones iniciales se representan mediante un miembro de la familia Normal, $p(\theta) = N(\theta | \mu, H)$, elección razonable habida cuenta de que el espacio paramétrico coincide con R^m, entonces:

$$p(\theta | D) \propto p(\theta)\ V(\theta | D) \propto N(\theta | \mu, H)\ \Phi_n(Diag\ X' \theta) \propto$$

$$\propto \exp\{-\tfrac{1}{2}(\theta - \mu)' H (\theta - \mu)\} \prod_{i=1}^{n} \left(\int_{-\infty}^{\delta_i x_i' \theta} \exp\{-\tfrac{1}{2} z_i^2\}\ dz_i \right)$$

Tras los cambios de variables apropiados:

$$p(\theta|D) \propto \int_{R^n_{(-)}} \exp\left\{-\tfrac{1}{2}\{(\theta-\mu)'H(\theta-\mu)+ +(y+Diag\ X'\theta)'(y+Diag\ X'\theta)\}\right\} dy \qquad (4)$$

donde $R^n_{(-)} \equiv \{z \in R^n;\ t_i<0,\ i=1,\ldots,n\}$. ■■

Antes de continuar con este ejemplo es necesario introducir la siguiente

PROPOSICION 1

Sean H *y* G *dos matrices definidas positivas y simétricas, entonces:*

$$(\theta-\mu)'H(\theta-\mu) + (B\theta-m)'G(B\theta-m) =$$

$$= \{\theta-(H+B'GB)^{-1}(H\mu+B'Gm)\}'(H+B'GB)\{\theta-(H+B'GB)^{-1} (H\mu+B'Gm)\}+(m-B\mu)'(G^{-1}+BH^{-1}B')^{-1}(m-B\mu) \qquad (5)$$

DEMOSTRACION

$$(\theta-\mu)'H(\theta-\mu) + (B\theta-m)'G(B\theta-m) =$$

$$= \theta'H\theta+\mu'H\mu-2\theta'H\mu+\theta'B'GB\theta+m'Gm-2\theta'B'Gm =$$

$$= \theta'(H+B'GB)\theta-2\theta'(H+B'GB)(H+B'GB)^{-1}(H\mu+B'Gm)+\mu'H\mu+m'Gm$$

Completando la forma cuadrática en θ, la proposición quedará demostrada si se comprueba que

$$\mu'H\mu+m'Gm-(H\mu+B'Gm)'(H+B'GB)^{-1}(H\mu+B'Gm) \qquad (6)$$

coincide con el segundo sumando de la parte derecha de la expresión (5).

Ahora bien, (6) es:

$$\mu'(H-H(H+B'GB)^{-1}H)\mu + m'(G-G'B(H+B'GB)^{-1}B'G)m - \\ - 2m'G(B(H+B'GB)^{-1}H)\mu \qquad (7)$$

Esta expresión puede simplificarse mediante la igualdad de matrices (Rao, 1973, ejercicio 2.9, pag. 33),

$$C-CA(B+A'CA)^{-1}A'C = (C^{-1}+AB^{-1}A')^{-1} \qquad (8)$$

de la que se deducen, como corolarios,

$$A'(C+ABA')^{-1}C = B^{-1}(B^{-1}+A'C^{-1}A)^{-1}A' \qquad (9)$$

$$C-C(C+ABA')^{-1}C = A(B^{-1}+A'C^{-1}A)^{-1}A' \qquad (10)$$

Utilizando (10), (8) y (9) respectivamente en los tres sumandos de (7), se obtiene:

$$\mu'B'(G^{-1}+BH^{-1}B')^{-1}B\mu + m'(G^{-1}+BH^{-1}B')^{-1}m - \\ - 2m'(G^{-1}+BH^{-1}B')^{-1}B\mu$$

■

EJEMPLO 1 (Continuación)

Aplicando la proposición 1 a la expresión (4) y multiplicando por las constantes apropiadas:

$$p(\theta|D) = C\int_{R^n_{(-)}} p(\theta|y)\ f(y)\ dy$$

donde:

$$p(\theta|y) = N(\theta|(H+XDiag^2X')^{-1}(H\mu-XDiag\ y),(H+XDiag^2X')$$

$$f(y) = N(y|-Diag\ X'\mu,(I+Diag\ X'H^{-1}XDiag)^{-1})$$

y siendo C la constante de proporcionalidad, esto es,

$$C^{-1} = \int\int_{R^n_{(-)}} p(\theta|y)\ f(y)\ dy\ d\theta = \int_{R^n_{(-)}} f(y)\ dy =$$

$$= \Phi_n(0|-Diag\ X'\mu,(I+Diag\ X'H^{-1}XDiag)^{-1}) \qquad (11)$$

Por tanto, $p(\theta|D) = \int p(\theta|y)\ p(y)\ dy$. Siendo $p(y)$ la función de densidad de una cantidad aleatoria Normal truncada. Los momentos de la distribución final se pueden poner en función de los momentos de la cantidad aleatoria y, de forma que:

$$E(\theta|D) = \underset{y}{E}\{E(\theta|y)\} = \underset{y}{E}\{(H+XX')^{-1}(H\mu-XDiag\ y)\} =$$

$$= (H+XX')^{-1}(H\mu-XDiag\,E(y)) \qquad (12)$$

puesto que $Diag^2=I$.

$$E(\theta^2|D) = \underset{y}{E}\{E(\theta^2|y)\}$$

$$= \underset{y}{E}\{(H+XX')^{-1}+(H+XX')^{-1}(H\mu-XDiag\ y)$$

$$(H\mu-XDiag\ y)'(H+XX')^{-1}\} =$$

$$= (H+XX')^{-1}+(H+XX')^{-1}\underset{y}{E}\{(H\mu-XDiag\ y)$$

$$(H\mu-XDiag\ y)'\}(H+XX')^{-1} =$$

$$= (H+XX')^{-1}+E(\theta|D)E'(\theta|D)+(H+XX')^{-1}XDiag\,VAR(y)$$

$$Diag\ X'(H+XX')^{-1}$$

en consecuencia,

$$VAR(\theta|D) = (H+XX')^{-1}+(H+XX')^{-1}XDiag\,VAR(y) \qquad (13)$$

$$Diag\ X'(H+XX')^{-1}$$

Los momentos de la distribución Normal truncada $p(y)$, $E(y)$ y $VAR(y)$, pueden expresarse (Tallis, 1961) como combinaciones lineales de funciones de distribución Normales n, $n-1$ y $n-2$ dimensionales.

De forma similar, la distribución predictiva buscada es,

$$p(\delta|x,D) \propto \int p(\delta|x,\theta)\ V(\theta|D)\ p(\theta)\ d\theta \propto$$

$$\propto \int V^*(\theta|D,(\delta,x))\ p(\theta)\ d\theta$$

donde $V^*(\theta|D,(\delta,x))$ representa a la función de verosimilitud del banco de datos ampliado por el elemento (δ,x). Por tanto, si X^* y $Diag^*(\delta)$ son las matrices correspondientes al banco de datos ampliado,

$$p(\delta|x,D) \propto \iint_{R^n_{(-)}} \exp\Big\{-\tfrac{1}{2}\{(\theta-\mu)'H(\theta-\mu)+$$

$$+(y^*+Diag^*(\delta)X^{*\prime}\theta)'(y^*+Diag^*(\delta)X^{*\prime}\theta)\}\Big\}\ dy^*\ d\theta$$

Teniendo en cuenta que $|H+X^*X^{*\prime}|$ no depende de δ y que

$$|I+Diag^*(\delta)X^{*\prime}H^{-1}X^*Diag^*(\delta)| =$$

$$= |Diag^*(\delta)(I+X^{*\prime}H^{-1}X^*)Diag^*(\delta)| =$$

$$= |I+X^{*\prime}H^{-1}X^*|,$$

tampoco depende de δ. La distribución predictiva, tras aplicar la proposición 1 y multiplicar por las constantes apropiadas, es:

$$p(\delta|x,D) \propto \int\int_{R_{(-)}^{n+1}} N(\theta|(H+X^*X^{*\prime})^{-1}(H\mu-X^*Diag^*(\delta)y^*),$$

$$(H+X^*X^{*\prime}))\ N(y^*|-Diag^*(\delta)X^{*\prime}\mu,(I+Diag^*(\delta)X^{*\prime}$$

$$H^{-1}X^*Diag^*(\delta))^{-1})\ dy^*\ d\theta \propto$$

$$\propto \int_{R_{(-)}^{n+1}} N(y^*|-Diag^*(\delta)X^{*\prime}\mu,(I+Diag^*(\delta)X^{*\prime}H^{-1}X^*$$

$$Diag^*(\delta))^{-1})\ dy^*$$

$$\propto \{C^*(\delta)\}^{-1}$$

con $\{C^*(\delta)\}^{-1}=\Phi_{n+1}(0|-Diag^*(\delta)X^{*\prime}\mu,$

$$(I+Diag^*(\delta)X^{*\prime}H^{-1}X^*Diag^*(\delta))^{-1})$$

Como se trata de una distribución discreta, la constante de proporcionalidad es fácilmente calculable con lo que,

$$p(\delta=1|x,D) = C^*(-1)/(C^*(1)+C^*(-1)) \qquad (14)$$

Las fórmulas obtenidas tanto para la esperanza y

varianza de la distribución final como para la distribución predictiva, expresiones (12), (13) y (14), tienen una forma analítica sencilla, sin embargo su cálculo no es posible en la práctica. Esto es debido a que involucran el cálculo de funciones de distribución normales de dimensión el número de datos, n. Si n no es muy pequeño, no mayor que 3 ó 4, el cálculo de $\Phi_n(\cdot)$ es prohibitivo.

El hecho de que las expresiones (12), (13) y (14) sean fácilmente calculables si n es muy pequeño, en particular para un tamaño muestral 1, apunta hacia la aproximación de los resultados anteriores mediante el siguiente método iterativo. En cada iteración se introduce un nuevo dato, calculando la distribución final correspondiente; dicha distribución final se aproxima mediante su mejor aproximación Normal en el sentido de la discrepancia de Kullbach-Leibel, esto es, por la distribución Normal con la misma media y varianza; esta aproximación Normal se utiliza como distribución inicial en la siguiente iteración.

Así, en la $(r+1)$ iteración:

$$p(\theta|(\delta_1,x_1),\ldots,(\delta_r,x_r)) \doteq N(\theta|\mu_r,H_r)$$

$$\mu_{r+1} = E(\theta|(\delta_1,x_1),\ldots,(\delta_{r+1},x_{r+1})) =$$

$$= (H_r+x_rx_r')^{-1}\ (H_r\mu_r+\delta_r\ E(y_r)x_r)$$

$$(H_{r+1})^{-1} = VAR(\theta|(\delta_1,x_1),\ldots,(\delta_{r+1},x_{r+1})) =$$

$$= (H_r+x_rx_r')^{-1} + VAR\ (y_r)\ (H_r+x_rx_r')^{-1}\ x_rx_r'(H_r+x_rx_r')^{-1}$$

donde $E(y_r)$ y $VAR(y_r)$ son, respectivamente, la media y varianza de una distribución Normal de parámetros $-\delta_r x_r' \mu_r$ y $(1+x_r' H_r^{-1} x_r)^{-1}$, truncada en 0, por tanto Johnson and Kotz (1970; pag. 81-83):

$$E(y_r) = -\delta_r x_r' \mu_r - (1+x_r' H_r^{-1} x_r)^{\frac{1}{2}} \frac{\phi(a_r)}{\Phi(a_r)}$$

$$VAR(y_r) = (1+x_r' H_r^{-1} x_r) \left\{ 1 - a_r \frac{\phi(a_r)}{\Phi(a_r)} - \left[\frac{\phi(a_r)}{\Phi(a_r)} \right]^2 \right\}$$

con $a_r = (1+x_r' H_r^{-1} x_r)^{-\frac{1}{2}} \delta_r x_r' \mu_r$.

Este método iterativo no es invariante con respecto al orden en el que se introducen las obsevaciones, por lo que en su aplicación será aconsejable el estudio de su robustez frente a cambios en el orden de introducción de los datos.

Una vez obtenida la aproximación Normal a la distribución final, $p(\theta|D) \cong N(\theta|\mu_n, H_n)$, las propiedades de la distribución multinormal aseguran que,

$$p(\theta' x|D) \cong N(\theta' x \mid \mu_n' x, (x' H_n x)^{-1})$$

sea cual sea el vector representante x. Por tanto, aplicando la proposición 1,

$$p(\delta=1|x,D) = \int p(\delta=1|x,\theta)\, p(\theta|D)\, d\theta \cong$$

$$\cong \int_{-\infty}^{0} (2\pi)^{-\frac{1}{2}} (x' H_n x)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x' H_n^{-1} x)^{-1} (y-\mu_n' x)^2\}\, dy =$$

$$= \Phi(\delta \mu_n' x\ (1+x' H_n\ x)^{-\frac{1}{2}})$$

Esta aproximación, a parte de su dependencia del orden de introducción de los datos, presenta el incoveniente de necesitar invertir una matriz, la matriz H_r, en cada iteración. Por tanto, si el número de datos, n, es suficientemente grande este método puede resultar demasiado lento y por ello poco aconsejable. Sin embargo, es en ese caso, n grande, cuando puede esperarse un buen funcionamiento de la aproximación asintótica. ■■

Introducir información inicial sobre los parámetros de un modelo de clasificación regular presenta graves inconvenientes, no solo desde el punto de vista matemático, búsqueda de modelos multivariantes tratables, sino también desde un punto de vista práctico. En efecto, el experto puede tener cierta información a priori sobre la distribución de clasificación, pero traducir esa información en una distribución inicial, $p(\theta)$, sobre el parámetro desconocido $\theta \in \Theta$, puede ser difícil.

En consecuencia, en el enfoque clasificatorio resulta particularmente importante la obtención de distribuciones de referencia. Son esas distribuciones las correspondientes a aquellas situaciones en las que o no se desea introducir información inicial o esta es demasiado débil para tenerla en cuenta. Además, las distribuciones de referencia constituyen una forma de medir la influencia de la información inicial en el resultado final.

El resto de este capítulo trata sobre la obtención teórica y el cálculo práctico de las distribuciones de referencia para los modelos de clasificación regulares.

4.2 DISTRIBUCIONES DE REFERENCIA

La *Regla de Jeffreys* (Jeffreys, 1946, 1967) es el primer resultado importante conseguido en la búsqueda de distribuciones iniciales no informativas sobre parámetros continuos. Jeffreys, basándose en argumentos de invarianza, propone el uso como distribución inicial no informativa de la raíz cuadrada del determinante de la matriz de información de Fisher del modelo muestral, i.e.,

$$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}}, \quad \text{siendo } I(\theta) = \underset{x|\theta}{E}\{D_\theta^2 - \text{Log } p(x|\theta)\}$$

Este resultado parece funcionar perfectamente en el caso unidimensional continuo, sin embargo ha sido bastante criticado, incluso por el mismo Jeffreys, en el caso multidimensional.

Basándose en argumentos de Teoría de la Información Bernardo (1975,1979) propone un método para el cálculo de las distribuciones de referencia que, en cierta forma, generaliza la Regla de Jeffreys caracterizando condiciones suficientes para su validez. Así, cuando el parámetro es continuo y el modelo *regular*, i.e. la función de verosimilitud cumple las condiciones de regularidad para la normalidad asintótica, los dos métodos coinciden si el parámetro de interes es el vector paramétrico θ, completo; esto es, si el objetivo es hacer inferencias sobre todo el vector paramétrico y no solamente sobre parte del mismo.

En particular, si el objetivo principal es la predicción de resultados futuros, el parámetro de interés

es la nueva observación, mientras que el vector paramétrico juega un papel marginal. Por ello puede considerarse como distribución predictiva de referencia la calculada a partir de la distribución de referencia del vector paramétrico completo, θ. (Ver Bernardo, 1979, en contestación a Bartholomew). Así pues, parece razonable utilizar

$$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}} \propto \left| \underset{x,\delta|\theta}{E} \{-D_\theta^2 \text{Log } p(\delta,x|\theta)\} \right|^{\frac{1}{2}} \qquad (1)$$

como distribución inicial de referencia para problemas de clasificación.

4.2.1 MODELOS DE CLASIFICACION REGULARES

La distribución de referencia dada por la expresión (1) para los modelos de clasificación regulares,

$$p(\delta,x|\theta) \propto p(\delta|x,\theta) = F_\delta(\theta' x)$$

es

$$\pi(\theta) \propto \left| \underset{x}{E} \{-\sum_{\delta=1}^{k} F_\delta(\theta' x)\ D_\theta^2(\text{Log } F_\delta(\theta' x))\} \right|^{\frac{1}{2}}$$

En este apartado se estudia con detalle esa distribución inicial proponiendo aproximaciones para el caso, muy frecuente, en el que la distribución de los vectores representantes es desconocida y, por tanto, la esperanza $\underset{x}{E}$ no puede ser calculada directamente.

Por la regla de la cadena, denotando mediante t a $t(x,\theta)=\theta'x$,

$$\frac{\partial^2}{\partial\theta_{i_1 j_1}\partial\theta_{i_2 j_2}} \operatorname{Log} F_\delta(\theta'x) =$$

$$= \frac{\partial}{\partial\theta_{i_2 j_2}} \left\{ \frac{\partial}{\partial t_{j_1}} \operatorname{Log} F_\delta(t) \frac{\partial}{\partial\theta_{i_1 j_1}} t_{j_1} \right\} =$$

$$= x_{i_1} \frac{\partial^2}{\partial t_{j_1}\partial t_{j_2}} \operatorname{Log} F_\delta(t) \frac{\partial}{\partial\theta_{i_2 j_2}} t_{j_2} =$$

$$= x_{i_1} x_{i_2} \frac{\partial^2}{\partial t_{j_1}\partial t_{j_2}} \operatorname{Log} F_\delta(t)$$

luego, si $D^2_\theta(\operatorname{Log} F_\delta(\theta'x))$ es la matriz de derivadas parciales segundas con respecto a θ de la función $\operatorname{Log} F_\delta(\theta'x)$, y $D^2_t(\operatorname{Log} F_\delta(t))$ se define de forma similar, entonces:

$$-D^2_\theta(\operatorname{Log} F_\delta(\theta'x)) = xx' \otimes \{-D^2_t(\operatorname{Log} F_\delta(t))\}$$

donde $\otimes$ representa el producto de Kronecker de matrices.

Además,

$$\frac{\partial^2}{\partial t_i \partial t_j} \operatorname{Log} F_\delta(t) = \frac{\partial}{\partial t_j}\left\{ \frac{1}{F_\delta(t)} \frac{\partial}{\partial t_i} F_\delta(t) \right\} =$$

$$= -(F_\delta(t))^{-2} \frac{\partial}{\partial t_j} F_\delta(t) \frac{\partial}{\partial t_i} F_\delta(t) +$$

$$+ \frac{1}{F_\delta(t)} \frac{\partial^2}{\partial t_i \partial t_j} F_\delta(t) =$$

$$= - \frac{\partial}{\partial t_i} \operatorname{Log} F_\delta(t) \frac{\partial}{\partial t_j} \operatorname{Log} F_\delta(t) +$$

$$+ \frac{1}{F_\delta(t)} \frac{\partial^2}{\partial t_i \partial t_j} F_\delta(t)$$

de donde se deduce:

$$\mathop{E}_{\delta|x} \left\{ - \frac{\partial^2}{\partial t_i \partial t_j} \text{Log } F_\delta(t) \right\} =$$

$$= \sum_{\delta=1}^{k} F_\delta(t) \frac{\partial}{\partial t_i} \text{Log } F_\delta(t) \frac{\partial}{\partial t_j} \text{Log } F_\delta(t) +$$

$$+ \sum_{\delta=1}^{k} \frac{\partial^2}{\partial t_i \partial t_j} F_\delta(t)$$

Ahora bien, $\sum F_\delta(t)=1$, constante, luego su derivada se anula. Por tanto,

$$\mathop{E}_{\delta|x} \{-D_t^2(\text{Log } F_\delta(t))\} = H$$

siendo H la matriz $(k-1)\times(k-1)$ con elemento genérico

$$H_{ij} = \sum_{\delta=1}^{k} F_\delta(t) \frac{\partial}{\partial t_i} \text{Log } F_\delta(t) \frac{\partial}{\partial t_j} \text{Log } F_\delta(t)$$

La matriz H es definida positiva, propiedad P2 de los modelos de clasificación regulares.

En resumen, la matriz de información de Fisher es:

$$I(\theta) = \mathop{E}_{x} \mathop{E}_{\delta|x} \left\{ xx' \otimes \{-D_t^2(\text{Log } F_\delta(t))\} \right\} =$$

$$= \mathop{E}_{x} \left\{ xx' \otimes \mathop{E}_{\delta|x} -D_t^2(\text{Log } F_\delta(t)) \right\} =$$

$$= \mathop{E}_{x} \{xx' \otimes H(\theta,x)\} \qquad (2)$$

Si la distribución que genera los vectores representantes fuese conocida entonces la matriz $I(\theta)$ podría ser calculada exactamente, siempre que las integrales involucradas tengan solución analítica. Por el contrario, en numerosas aplicaciones, la distribución de los vectores representantes es desconocida. En tal caso es preciso buscar aproximaciones a $I(\theta)$.

Utilizando el propio banco de datos para conseguir información sobre la distribución de los vectores representantes, la matriz $I(\theta)$ puede ser aproximada por Monte-Carlo, de manera que si el muestreo ha sido prospectivo,

$$I(\theta) \cong n^{-1} \sum_{i=1}^{n} \{x_i x_i' \otimes H(\theta, x_i)\}$$

donde n representa el tamaño muestral.

Por el contrario, si el muestreo ha sido retrospectivo obteniendose n_δ datos de la clase δ,

$$I(\theta) \cong \sum_{\delta=1}^{k} p(\delta)\ n^{-1} \sum_{i=1}^{n_\delta} \{x_{i\delta} x_{i\delta}' \otimes H(\theta, x_{i\delta})\}$$

La aproximación por Monte-Carlo a la distribución final es, por tanto,

$$\pi_{MC}(\theta) \propto \left| \sum_{i=1}^{n} \{x_i x_i' \otimes H(\theta, x_i)\} \right|$$

para el muestreo prospectivo.

Esa expresión tiene sentido si y solo si el determinante de la matriz involucrada es no nulo. Bajo condi-

ciones muy generales esa matriz es definida positiva. En efecto,

PROPOSICION 1

Sea x *un vector* m*-dimensional y* H=H(x) *una matriz* (k-1)×(k-1). *La matriz* $\underset{x}{E}\{xx' \otimes H\}$ *es definida positiva si la distribución que genera los vectores representantes,* p(x), *es discreta finita y existen* m *vectores* $\{x_i;\ i=1,\ldots,m\}$, *linealmente independientes, con* $p(x_i)>0$.

DEMOSTRACION

Se desea demostrar que el único vector a tal que $a'\underset{x}{E}\{xx' \otimes H\}a=0$ es $a=0$.

Dado $a \in R^{m(k-1)}$ sea $\{a_i \in R^{k-1};\ i=1,\ldots,m\}$ tal que $a=(a_1',\ldots,a_m')'$ y sea A la matriz $m\times(k-1)$ cuya columna i-ésima es el vector a_i.

$$a'(xx' \otimes H)a = \sum_{r=1}^{m}\sum_{s=1}^{m} a_r' x_r x_s H a_s = \sum_{r=1}^{m} x_r x_s a_r' H a_s =$$

$$= x'A'HAx$$

Como H es definida positiva, $a'(xx' \otimes H)a=x'A'HAx \geq 0$ con igualdad si y solo si $Ax=0$.

$$a'\underset{x}{E}\{xx' \otimes H\}a = \underset{x}{E}\{a'(xx' \otimes H)a\} = \underset{x}{E}\{x'A'HAx\} \geq 0$$

ya que se trata de la esperanza de una variable no negativa. Además, por la misma razón,

$$\underset{x}{E}\{x'A'HAx\} = 0 \iff P\{x'A'HAx>0\} = 0 \iff$$

$$\iff P\{Ax>0\} = 0 \iff P\{x \in A^{\perp}\} = 1$$

donde $A^{\perp}$ representa el espacio ortogonal al generado por A. Ahora bien, para que la última probabilidad en la expresión anterior valga uno, los m vectores linealmente independientes $\{x_i\}$, deben pertenecer al espacio $A^{\perp}$, luego $\text{Rango}(A^{\perp}) \geq m$. Por tanto,

$$P(x \in A^{\perp}) = 1 \Rightarrow \text{Rango}(A^{\perp}) \geq m \Rightarrow 0 \leq \text{Rango}(A) = m - \text{Rango}(A^{\perp}) \leq 0 \Rightarrow$$

$$\Rightarrow \text{Rango}(A) = 0 \Rightarrow A \equiv 0 \Rightarrow a \equiv 0$$

∎

COROLARIO 1

Si los vectores representantes incluídos en el banco de datos, $\{x_i;\ i=1,\ldots,n\}$, *forman un sistema de generadores de* R^m *entonces la matriz*

$$\sum_{i=1}^{n} \{x_i x_i' \otimes H(\theta, x_i)\}$$

es definida positiva.

DEMOSTRACION

$\sum_{i=1}^{n} \{x_i x_i' \otimes H(\theta, x_i)\}$ puede considerarse como n veces la esperanza de la variable aleatoria discreta con n puntos posibles, cada uno de ellos con probabilidad $1/n$. Como entre esos vectores existen m linealmente independientes, puede aplicarse la proposición anterior obteniendose directamente el resultado buscado.

∎

COROLARIO 2

Si los n_δ *vectores representantes incluídos en el banco de datos y correspondientes a la clase* δ *forman un sistema de generadores de* R^m *entonces la matriz* $\sum p(\delta) \sum \{x_{i\delta} x_{i\delta}' \otimes H(\theta, x_{i\delta})\}$ *es definida positiva.*

El corolario 2 es la réplica del corolario 1 para muestreos retrospectivos. Su demostración es totalmente similar a la del corolario 1.

EJEMPLO 1

Sean $x_1=(1,0)'$ y $x_2=(1,1)'$ los únicos valores posibles del vector representante; sean p_1 y $p_2=1-p_1$ las frecuencias relativas con las que x_1 y x_2 han aparecido en el banco de datos. La aproximación por Monte-Carlo a la distribución inicial de referencia es:

$$\pi_{MC}(\theta) \propto \left|p_1 x_1 x_1' H(\theta,x_1) + p_2 x_2 x_2' H(\theta,x_2)\right|^{\frac{1}{2}} =$$

$$= \begin{vmatrix} p_1H(\theta,x_1)+p_2H(\theta,x_2) & p_2H(\theta,x_2) \\ p_2H(\theta,x_2) & p_2H(\theta,x_2) \end{vmatrix}^{\frac{1}{2}}$$

Ahora bien, utilizando que si $A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ con A_{22} no singular, entonces $|A|=|A_{22}||A_{11}-A_{12}A_{22}^{-1}A_{21}|$, (Graybill, 1961, teorema 1.50, pag. 8):

$$\pi_{MC}(\theta) \propto \left|p_2H(\theta,x_2)\right|^{\frac{1}{2}} \left|p_1H(\theta,x_1)\right|^{\frac{1}{2}} \propto \left|H(\theta,x_1)\right|^{\frac{1}{2}} \left|H(\theta,x_2)\right|^{\frac{1}{2}}$$

Resultado que no depende en absoluto de las frecuencias p_1 y p_2, por lo que coincide con el resultado que se obtendría si se conociese la función $p(x)$, esto es, con el resultado exacto.

■■

Sin embargo, si el número de vectores representantes es mayor de dos, la aproximación por Monte-Carlo puede depender de las frecuencias relativas con las que

los vectores representantes aparecen en el banco de datos. Como consecuencia, la aproximación a la distribución inicial de referencia así obtenida no ha de coincidir necesariamente con la verdadera distribución inicial de referencia. Una prueba de esto es el siguiente

EJEMPLO 2

Sean $x_1=(1,0)'$, $x_2=(1,1)'$ y $x_3=(1,-1)'$ los únicos valores posibles del vector representante. Sean p_1, p_2 y p_3, respectivamente las frecuencias relativas con las que x_1, x_2 y x_3 han aparecido en el banco de datos.

La aproximación por Monte-Carlo a la distribución inicial de referencia es:

$$\pi_{MC}(\theta) \propto |p_1x_1x_1'H_1+p_2x_2x_2'H_2+p_3x_3x_3'H_3|^{\frac{1}{2}} =$$

$$= |p_2H_2+p_3H_3|^{\frac{1}{2}} \; |p_1H_1+p_2H_2+p_3H_3-$$

$$-(p_2H_2-p_3H_3)(p_2H_2+p_3H_3)^{-1}(p_2H_2-p_3H_3)|^{\frac{1}{2}}$$

donde $H_i=H(\theta,x_i)$, son matrices definidas positivas.

Si $p_2=p_3$ y se denota por q el cociente p_1/p_2, entonces la expresión anterior se convierte en

$$\pi_{MC}(\theta) \propto |H_2+H_3|^{\frac{1}{2}} \; |qH_1+H_2+H_3-(H_2-H_3)(H_2+H_3)^{-1}(H_2-H_3)|^{\frac{1}{2}}$$

que, obviamente, depende del cociente q ya que H_1 es una matriz no nula.

■■

En cualquier caso, las frecuencias relativas pueden ser una aproximación suficientemente buena a la distribución de los vectores representantes, $p(x)$, si esta distribución es discreta finita y existen pocos vectores representantes distintos en relación con el número de datos, n. Por tanto, bajo esas condiciones es de esperar que la aproximación por Monte-Carlo sea bastante precisa.

Sin embargo, el uso de $\pi_{MC}(\theta)$ como distribución inicial presenta graves inconvenientes desde el punto de vista computacional; esto es así ya que no parece existir una solución analítica sencilla al determinante que aparece en la definición $\pi_{MC}(\theta)$, siendo su desarrollo, en general, prohibitivo.

Por ejemplo, considérese la situación en la que el vector representante está formado por un término constante y tres indicadores dicotómicos, m=4, -con lo que el número de vectores representantes distintos es 8- y existen solo dos clases distintas, k=2, -con lo que $H(\theta,x)$ es un número real-. El determinante involucrado en el cálculo de $\pi_{MC}(\theta)$ corresponde a una matriz 4×4, por tanto su desarrollo conlleva una suma formada por 18 sumandos, cada uno resultado del producto de 4 factores. Como, además, el determinante a desarrollar corresponde a la suma de ocho matrices, el número total de operaciones involucradas es enorme.

Una alternativa a la aproximación por Monte-Carlo se obtiene si se supone que las matrices xx' y $H(\theta,x)$ son aproximadamente independientes, entonces:

$$\underset{x}{E}(xx' \otimes H(\theta,x)) \cong \underset{x}{E}(xx') \otimes \underset{x}{E}(H(\theta,x))$$

La matriz H depende de θ y x a través de la combinación lineal $\theta'x$, por tanto la covarianza de H y xx' será, en general, no nula. Sin embargo, si la varianza de x es suficientemente pequeña para no esperar grandes cambios en las probabilidades $p(\delta|x,\theta)=F_\delta(\theta'x)$ a través de las cuales se define la función H, la suposición de independencia puede proporcionar una aproximación suficientemente buena.

Aceptado esto, la distribución inicial de referencia puede aproximarse mediante:

$$\pi(\theta) \propto \left|\mathop{E}_{x}(xx')\otimes\mathop{E}_{x}(H(\theta,x))\right|^{\frac{1}{2}} =$$

$$= \left|\mathop{E}_{x}(xx')\right|^{(k-1)/2}\left|\mathop{E}_{x}(H(\theta,x))\right|^{m/2} \propto \left|\mathop{E}_{x}(H(\theta,x))\right|^{m/2}$$

ya que si A y B son dos matrices cuadradas de dimensión $a\times a$ y $b\times b$ entonces $|A\otimes B|=|A|^b|B|^a$. (Anderson, 1958, teorema 10, pag. 348).

Para el cálculo de la esperanza de H, si no se desea modelizar $p(x)$ o las integrales involucradas no hacen posible su cálculo exacto, puede proponerse desarrollos de Taylor. En particular, una primera aproximación es $E(H(\theta,x)) = H(\theta,E(x))$. Obviamente, añadiendo nuevos términos al desarrollo de Taylor puede mejorarse la bondad de la aproximación tanto como se desee. Sin embargo, considerar mas de un término implica el cálculo del determinante de una suma de matrices.

Antes de continuar, puede resultar muy interesante el estudio del siguiente

EJEMPLO 3

Considérese un problema de clasificación en el que el vector representante está formado por un término constante y un indicador continuo, esto es, $m=2$, y $x=(1,y)'$. Si el número de clases es dos, $k=2$, entonces $H(\theta,x)$ es un número real y por tanto:

$$\left|\underset{x}{E}\{xx'\otimes H(\theta,x)\}\right| = \underset{y}{E}(H)\ \underset{y}{E}(y^2H) - \underset{y}{E}^2(yH)$$

Si μ y σ^2 representan la media y varianza de y, el desarrollo en series de Taylor de las funciones H, yH y y^2H, proporciona las aproximaciones (Lindley, 1965, teorema 3.4.1):

$$\underset{y}{E}(H) \cong H(\mu) + \tfrac{1}{2}\sigma^2 \frac{d^2}{dy^2} H(\mu)$$

$$\underset{y}{E}(yH) \cong \mu H(\mu) + \tfrac{1}{2}\sigma^2\left(2\frac{d}{dy}H(\mu)+\mu\frac{d^2}{dy^2}H(\mu)\right)$$

$$\underset{y}{E}(y^2H) \cong \mu^2 H(\mu) + \tfrac{1}{2}\sigma^2\left(2H(\mu)+4\mu\frac{d}{dy}H(\mu)+\mu^2\frac{d^2}{dy^2}H(\mu)\right)$$

por tanto, tras un sencillo cálculo,

$$E(H)\ E(y^2H) - E^2(yH) \cong$$

$$\cong \sigma^2\{H(\mu)\}^2 + \tfrac{1}{2}\sigma^4\left(H(\mu)\frac{d^2}{dy^2}H(\mu) - 2\left\{\frac{d}{dy}H(\mu)\right\}^2\right)$$

si la desviación típica, σ, es suficientemente pequeña, el segundo sumando será despreciable frente al primero. Por tanto,

$$\pi(\theta) \propto \left|E(xx'\otimes H)\right|^{\frac{1}{2}} \doteq H(\mu) = \left|H(\theta,E(x))\right|^{m/2}$$

■■

Un compromiso entre bondad de aproximación y facilidad de cálculo puede aconsejar el uso de la función

$$|H(\theta, E(x))|^{m/2}$$

como distribución inicial de referencia.

Si no se conoce la esperanza de x, el propio banco de datos, o un banco alternativo, puede utilizarse para encontrar un estimador $\bar{x}$ de $E(x)$. En tal caso

$$\pi_E(\theta) \propto |H(\theta, \bar{x})|^{m/2}$$

proporciona una aproximación estimativa a la distribución inicial de referencia.

En el siguiente ejemplo se estudia la bondad de la aproximación $\pi_E(\theta)$ en un modelo en el que es posible no solo la obtención de la distribución inicial de referencia, $\pi(\theta)$, sino también, mediante integración numérica, el cálculo de los momentos de las respectivas distribuciones finales, $\pi_E(\theta|D)$ y $\pi(\theta|D)$.

EJEMPLO 4 *Datos simulados.*

Dos clases, $k=2$; un solo indicador dicotómico, $y \in \{-1,1\}$; y no se considera término constante en el vector representante, esto es, se supone θ_1 conocido e igual a 0, por tanto $m=1$. Si se asume el modelo logístico,

$$p(\delta=1|\theta,x) = 1-p(\delta=2|\theta,x) = \exp(y\theta)/(1+\exp(y\theta)),$$

$$y=-1,1$$

la distribución inicial de referencia para el parámetro θ, es

$$\pi(\theta) \propto \left| \underset{y}{E} \underset{\delta|y}{E} \left(-D^2_{\theta} \text{Log } p(\delta|y,\theta)\right) \right|^{\frac{1}{2}}$$

$$\propto \exp(\theta/2)/(1+\exp(\theta))$$

que coincide, al igual que en el ejemplo 1, con la aproximación $\pi_{MC}(\theta)$. Por el contrario,

$$\pi_E(\theta) \propto |H(\theta,\bar{y})|^{m/2}$$

$$\propto \exp(\bar{y}\theta)/(1+\exp(\bar{y}\theta))^2$$

siendo $\bar{y}$ la media muestral de los indicadores.

En la tabla 1 se recoge una muestra de tamaño 10 obtenida mediante simulación a partir de una población en la que $p(y=1)=0.75$ y siendo $p(\delta|y,\theta)$ un modelo logístico con parámetro $\theta=2$.

y	1	-1	1	1	1	-1	1	1	1	1
δ	1	2	1	2	1	1	1	1	1	1

- tabla 1-

- banco de datos simulado-

La función de verosimilitud es:

$$V(\theta|D) \propto \left(\exp(\theta)/(1+\exp(\theta))\right)^7 \left(1/(1+\exp(\theta))\right) \times$$

$$\times \left(\exp(-\theta)/(1+\exp(-\theta))\right) \left(1/(1+\exp(-\theta))\right)$$

$$\propto \exp(8\theta) \left(1+\exp(\theta)\right)^{-10}$$

por tanto, como $\bar{y}=0.6$,

$$\pi(\theta|D) = C\ \exp(8.5\theta)(1+\exp(\theta))^{-11}$$

$$\pi_E(\theta|D) = C_E\ \exp(8.6\theta)(1+\exp(\theta))^{-10}\ (1+\exp(0.6\theta))^{-2}$$

Tanto las constantes de integración, C y C_E, como la media y varianza de estas distribuciones pueden obtenerse fácilmente por integración numérica, ver tabla 2. De igual forma, tabla 3, pueden calcularse las distribuciones de clasificación predictivas.

	$\pi(\theta\|D)$	$\pi_E(\theta\|D)$
Cte. de integración	194.5059	367.5408
media	1.3769	1.4070
varianza	0.6152	0.6247

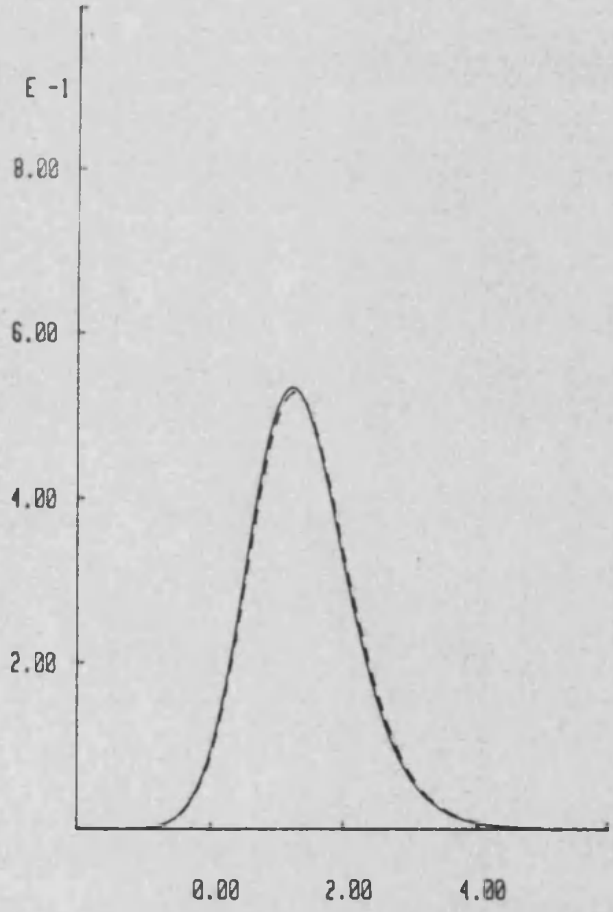
- tabla 2 -

	Dist. Final		
$p(\delta=1)$	$\pi(\theta\|D)$	$\pi_E(\theta\|D)$	Discrep.
$y=1$	0.772727	0.777160	0.000056
$y=-1$	0.227273	0.222840	0.000056

- tabla 3 -

Las gráficas de las distribuciones $\pi(\theta|D)$ y $\pi_E(\theta|D)$ se muestran en la figura 1.

A pesar de la sencillez de este ejemplo, sencillez inevitable para que el error producido en la integración numérica pueda considerarse despreciable, los buenos resultados obtenidos resultan, cuando menos, esperanzadores.



- fig. 1 -

■■

EJEMPLO 5. *Modelo Normal-Acumulado para dos clases.*

En dos poblaciones $H(\theta,x)$ no es una matriz sino un número real, exactamente

$$H(\theta,x) = F_1(t) \frac{d}{dt} \text{Log } F_1(t) + F_2(t) \frac{d}{dt} \text{Log } F_2(t)$$

Si se utiliza el modelo Normal-Acumulado,

$$F_1(t) = \Phi(t), \quad F_2(t) = 1-F_1(t) = \Phi(-t)$$

$$\frac{d}{dt} F_1(t) = \phi(t), \qquad \frac{d}{dt} F_2(t) = -\phi(t)$$

donde $\Phi(.)$ es la función de distribución y $\phi(.)$ la de densidad de una distribución Normal con media cero y varianza uno. Por tanto,

$$H(\theta,x) = \phi^2(t)/\Phi(t) + \phi^2(t)/\Phi(-t) =$$

$$= \phi^2(t)\ \Phi^{-1}(t)\ \Phi^{-1}(-t)$$

Por tanto, utilizando la segunda aproximación, la distribución inicial de referencia es

$$\pi_E(\theta) \propto |H(\theta,\bar{x})|^{m/2} \propto \phi^m(\theta'\bar{x})\ \Phi^{-m/2}(\theta'\bar{x})\ \Phi^{-m/2}(-\theta'\bar{x})$$

siendo

$$\phi^m(\theta'\bar{x}) \propto \exp(-m/2\ \theta'\bar{x}\bar{x}'\theta) \propto N(\theta|0,m\bar{x}\bar{x}')$$

■■

4.2.2. MODELOS DE TIPO 1

Los modelos de clasificación regulares de Tipo 1 se definieron en el apartado 2.2 como aquellos modelos de clasificación tales que:

$$p(\delta=i|x,\theta) = F_i(t), \qquad i=1,\ldots,k$$

con:

$$F_i(t) = \frac{\psi_i(t_i)}{1-\psi_i(t_i)} F_k(t), \qquad i=1,\ldots,k-1$$

$$F_k(t) = \left[1 + \sum_{i=1}^{k-1} \frac{\psi_i(t_i)}{1-\psi_i(t_i)}\right]^{-1} \tag{3}$$

siendo $t=t(\theta,x)=\theta' x \in R^{k-1}$ con $t_i=\theta'_i x$, componente i-ésima del vector t, y donde $\{\psi_i(.);\ i=1,\ldots,k-1\}$ son $k-1$ funciones, posiblemente distintas, de R en el intervalo abierto (0,1) que cumplen las condiciones exigidas para los modelos de clasificación regulares en dos clases.

PROPOSICION 2

Si $F_\delta(t)$ corresponde a un modelo de Tipo 1, expresión (3), entonces:

$$H(\theta,x) = Diag_1 (Diag_2 - 1_{k-1} 1'_{k-1}) Diag_1$$

siendo $Diag_1$ y $Diag_2$ dos matrices diagonales $(k-1)\times(k-1)$, tales que,

$$(Diag_1)_{ii} = (1-\psi_i)^{-2} F_k \frac{d\psi_i}{dt_i}, \quad (Diag_2)_{ii} = F_i^{-1}$$

y donde 1_{k-1} representa el vector $(k-1)$ con todas sus componentes iguales a uno.

DEMOSTRACION

Los elementos de la diagonal de la matriz de derivadas segundas parciales son:

$$\frac{\partial^2}{\partial t_i^2} \operatorname{Log} F_\delta(t) = \frac{\partial}{\partial t_i}\left[\frac{\partial}{\partial \psi_i} \operatorname{Log} F_\delta \frac{d\psi_i}{dt_i}\right] =$$

$$\frac{\partial^2}{\partial\psi_i}\,\mathrm{Log}\,F_\delta\left(\frac{d\psi_i}{dt_i}\right)^2+\frac{\partial}{\partial\psi_i}\,\mathrm{Log}\,F_\delta\,\frac{d^2\psi_i}{dt_i}$$

mientras que fuera de la diagonal,

$$\frac{\partial^2}{\partial t_i\partial t_j}\,\mathrm{Log}\,F_\delta(t)=\frac{\partial}{\partial t_j}\left[\frac{\partial}{\partial\psi_i}\,\mathrm{Log}\,F_\delta\,\frac{d\psi_i}{dt_i}\right]=$$

$$=\frac{\partial^2}{\partial\psi_i\partial\psi_j}\,\mathrm{Log}\,F_\delta\,\frac{d\psi_i}{dt_i}\,\frac{d\psi_j}{dt_j},\qquad \text{si}\quad i\neq j$$

con lo que $D_t^2(\mathrm{Log}\,F_\delta(t))=\mathrm{Diag}_3(\delta)+\mathrm{Diag}_4\mathrm{Mat}(\delta)\mathrm{Diag}_4$, donde $\mathrm{Diag}_3(\delta)$ y Diag_4 son dos matrices diagonales $(k-1)\times(k-1)$ tales que

$$(\mathrm{Diag}_3(\delta))_{ii}=\frac{\partial}{\partial\psi_i}\,\mathrm{Log}\,F_\delta\,\frac{d^2\psi_i}{dt_i},\qquad(\mathrm{Diag}_4)_{ii}=\frac{d\psi_i}{dt_i}$$

mientras que $\mathrm{Mat}(\delta)$ es la matriz $(k-1)\times(k-1)$ con elemento genérico

$$(\mathrm{Mat}(\delta))_{ij}=\frac{\partial^2}{\partial\psi_i\partial\psi_j}\,\mathrm{Log}\,F_\delta$$

Por tanto

$$H=\underset{\delta|x}{E}\{-D_t^2(\mathrm{Log}\,F_\delta(t))\}=$$

$$=-\underset{\delta|x}{E}\{\mathrm{Diag}_3(\delta)\}+\mathrm{Diag}_4\underset{\delta|x}{E}\{-\mathrm{Mat}(\delta)\}\mathrm{Diag}_4$$

ahora bien:

$$\underset{\delta|x}{E}\left[\frac{\partial}{\partial\psi_i}\ \text{Log}\ F_\delta\right] = \sum_{\delta=1}^{k} F_\delta\ \frac{\partial}{\partial\psi_i}\ \text{Log}\ F_\delta = \frac{\psi}{\partial\psi_i}\ \sum_{\delta=1}^{k} F_\delta = 0$$

ya que $\sum F_\delta=1$, constante. Luego $E\{\text{Diag}_3(\delta)\}=0$. Por otra parte:

$$\frac{\partial^2}{\partial\psi_i^2}\ -\text{Log}\ F_k = \frac{\partial^2}{\partial\psi_i^2}\ \text{Log}\left[1+\sum_{\delta=1}^{k-1}\frac{\psi_\delta}{1-\psi_\delta}\right] = \frac{\partial}{\partial\psi_i}\ (1-\psi_i)^{-2}F_k =$$

$$= \frac{2}{(1-\psi_i)^3}\ F_k - \frac{1}{(1-\psi_i)^4}\ F_k^2 =$$

$$= \frac{F_k^2}{(1-\psi_i)^4}\ \{2(1-\psi_i)F_k^{-1}-1\}$$

$$\frac{\partial^2}{\partial\psi_i\partial\psi_j}\ -\text{Log}\ F_k = \frac{\partial}{\partial\psi_j}\ \frac{F_k}{(1-\psi_i)^2} = -\frac{F_k^2}{(1-\psi_i)^2(1-\psi_j)^2}\ ,\ \text{si}\ i\neq j$$

Luego $-\text{Mat}(k)=\text{Diag}_5(\text{Diag}_6(k)-1_{k-1}1'_{k-1})\text{Diag}_5$, donde Diag_5 y $\text{Diag}_6(k)$ son dos matrices diagonales $(k-1)\times(k-1)$, tales que

$$(\text{Diag}_5)_{ii}=\frac{F_k}{(1-\psi_i)^2},\qquad (\text{Diag}_6(k))_{ii}=2(1-\psi_i)F_k^{-1}$$

Si $\delta<k$:

$$\frac{\partial^2}{\partial\psi_i\partial\psi_j}\ -\text{Log}\ F_\delta = \frac{\partial^2}{\partial\psi_i\partial\psi_j}\ -\left[\text{Log}\ \frac{\psi_\delta}{1-\psi_\delta}+\text{Log}\ F_k\right] =$$

$$= \frac{\partial^2}{\partial\psi_i\partial\psi_j}\ -\text{Log}\ \frac{\psi_\delta}{1-\psi_\delta} + \frac{\partial^2}{\partial\psi_i\partial\psi_j}\ -\text{Log}\ F_k$$

donde el primer sumando es cero a no ser que $i=j=\delta$, entonces:

$$\frac{\partial^2}{\partial \psi_\delta^2} -\text{Log}\,\frac{\psi_\delta}{1-\psi_\delta} = \frac{1-2\psi_\delta}{\psi_\delta^2(1-\psi_\delta)^2}$$

por tanto, $-\text{Mat}(\delta)=\text{Diag}_5(\text{Diag}_6(\delta)-1_{k-1}1'_{k-1})\text{Diag}_5$, donde $\text{Diag}_6(\delta)$ es una matriz diagonal $(k-1)\times(k-1)$ tal que:

$$(\text{Diag}_6(\delta))_{ii} = 2(1-\psi_i)F_k^{-1} \qquad \text{si } i\neq\delta$$

$$= 2(1-\psi_\delta)F_k^{-1}+(1-2\psi_\delta)(1-\psi_\delta)^2(\psi_\delta F_k)^{-2}, \text{ si } i=\delta$$

luego:

$$H = \text{Diag}_4 \underset{\delta|x}{E}\{-\text{Mat}(\delta)\}\text{Diag}_4 =$$

$$= \text{Diag}_4 \underset{\delta|x}{E}\{\text{Diag}_5(\text{Diag}_6(\delta)-1_{k-1}1'_{k-1})\text{Diag}_5\}\text{Diag}_4 =$$

$$= \text{Diag}_4\text{Diag}_5\left\{\underset{\delta|x}{E}\{\text{Diag}_6(\delta)\}-1_{k-1}1'_{k-1}\right\}\text{Diag}_5\text{Diag}_4$$

Ahora bien, $\text{Diag}_1=\text{Diag}_4\text{Diag}_5$, por tanto la proposición estará demostrada si se comprueba que $\text{Diag}_2=E\{\text{Diag}_6(\delta)\}$,

$$\underset{\delta|x}{E}\{(\text{Diag}_6(\delta))_{ii}\} = \sum_{\delta=1}^{k} F_\delta(\text{Diag}_6(\delta))_{ii} =$$

$$= \sum_{\delta=1}^{k-1}\frac{\psi_\delta}{1-\psi_\delta}F_k(\text{Diag}_6(\delta))_{ii} + (\text{Diag}_6(k))_{ii}F_k =$$

$$= \frac{\psi_i}{1-\psi_i}F_k(1-2\psi_i)(1-\psi_i)^2\psi_i^{-2}F_k^{-2} +$$

$$+ \sum_{\delta=1}^{k-1} \left(\frac{\psi_\delta}{1-\psi_\delta} F_k 2(1-\psi_i) F_k^{-1} \right) + 2F_k(1-\psi_i)F_k^{-1} =$$

$$= (1-2\psi_i) \frac{1-\psi_i}{\psi_i} F_k^{-1} + 2(1-\psi_i) \left(1 + \sum_{\delta=1}^{k-1} \frac{\psi_\delta}{1-\psi_\delta} \right) =$$

$$= (1-\psi_i)F_k^{-1}((1-2\psi_i)\psi_i^{-1}+2) = \frac{1-\psi_i}{\psi_i} F_k^{-1} = F_i^{-1} = (\text{Diag}_2)_{ii}$$

■

Como corolario inmediato de esta proposición y de la expresión (2) se obtiene la distribución inicial de referencia para los modelos de Tipo 1:

$$\pi(\theta) \propto \left| \underset{x}{E}\{xx' \otimes H(\theta,x)\} \right|^{\frac{1}{2}}$$

siendo $H(\theta,x)=\text{Diag}_1(\text{Diag}_2-1_{k-1}1'_{k-1})\text{Diag}_1$.

Las dos aproximaciones propuestas anteriormente son:

$$\pi_{MC}(\theta) \propto \left| \sum_{i=1}^{n} \left[x_i x_i' \otimes \left(\text{Diag}_1(\theta,x_i)(\text{Diag}_2(\theta,x_i) - 1_{k-1}1'_{k-1})\text{Diag}_1(\theta,x_i) \right) \right] \right|^{\frac{1}{2}}$$

$$\pi_E(\theta) \propto \left| \text{Diag}_1(\theta,\bar{x})(\text{Diag}_2(\theta,\bar{x}) - 1_{k-1}1'_{k-1})\text{Diag}_1(\theta,\bar{x}) \right|^{m/2}$$

PROPOSICION 3

$$|H(\theta,x)| = \left(\prod_{i=1}^{k} F_i \right) \left(\prod_{i=1}^{k-1} \frac{d}{dt_i} Log \frac{\psi_i}{1-\psi_i} \right)^2$$

DEMOSTRACION

$$|H(\theta,x)| = |\text{Diag}_1(\text{Diag}_2 - 1_{k-1}1'_{k-1})\text{Diag}_1| =$$

$$= |\text{Diag}_1|^2 \, |\text{Diag}_2 - 1_{k-1}1'_{k-1}| =$$

$$= |\text{Diag}_1|^2 \, |\text{Diag}_2| \, |I - \text{Diag}_2^{-1}1_{k-1}1'_{k-1}|$$

Ahora bien, $|I-Aaa'|=(1-a'Aa)$, (ver demostración de la proposición 2.2.2), por tanto:

$$|H(\theta,x)| = |\text{Diag}_1|^2 \, |\text{Diag}_2| \, (1-1'_{k-1}\text{Diag}_2^{-1}1_{k-1}) =$$

$$= |\text{Diag}_1|^2 \, |\text{Diag}_2| \, (1-\text{Traza}(\text{Diag}_2^{-1}))$$

de donde se deduce inmediatamente la proposición. ∎

EJEMPLO 6 *Modelo logístico aditivo* con vector representante formado por un término constante y un indicador dicotómico.

El modelo logístico aditivo es un modelo de Tipo 1 en el que

$$\psi_i(y) = \psi(y) = \exp(y)/(1+\exp(y)), \quad i=1,\ldots,k-1, \quad y \in R$$

por tanto

$$\frac{d}{dt_i} \text{Log} \frac{\psi_i(t_i)}{1-\psi_i(t_i)} = \frac{d}{dt_i} t_i = 1, \quad i=1,\ldots,k-1$$

luego, por la proposición 3,

$$|H(\theta,x)| = \prod_{i=1}^{k} F_i(\theta'x)$$

Por otra parte, en el ejemplo 1 se calculó la distribución inicial de referencia exacta como

$$\pi(\theta) \propto |H(\theta,(1,0)')|^{\frac{1}{2}} \; |H(\theta,(1,1)')|^{\frac{1}{2}}$$

por lo que, para el modelo logístico aditivo,

$$\pi(\theta) \propto \prod_{i=1}^{k} \left\{ F_i((1,0)\theta) \; F_i((1,1)\theta) \right\}^{\frac{1}{2}}$$

Si n_{i1} y n_{i2} representan el número de datos pertenecientes a la clase $\delta=i$, con vector representante igual a $(1,0)'$ ó $(1,1)'$ respectivamente, entonces:

$$V(\theta|D) \propto \prod_{i=1}^{k} \left\{ \{F_i((1,0)\theta)\}^{n_{i1}} \{F_i((1,1)\theta)\}^{n_{i2}} \right\}$$

con lo que $\pi(\theta|D) \propto V(D|\theta) \; \pi(\theta)$,

$$\pi(\theta|D) \propto \prod_{i=1}^{k} \left\{ \{F_i((1,0)\theta)\}^{n_{i1}+\frac{1}{2}} \{F_i((1,1)\theta)\}^{n_{i2}+\frac{1}{2}} \right\}$$

■■

EJEMPLO 7 *Modelo logístico aditivo en general*

Si existen mas de dos vectores representantes distintos, el cálculo exacto de la distribución inicial de referencia no es posible de no conocer la distribución que genera los vectores representantes, e incluso en ese caso puede no conseguirse una forma cerrada para $\pi(\theta)$.

La primera aproximación propuesta en el apartado anterior, aplicada al modelo logístico aditivo, es:

$$\pi_{MC}(\theta) \propto \left| \sum_{i=1}^{n} x_i x_i' \otimes \{\mathrm{Diag}_1(\mathrm{Diag}_2 - 1_{k-1}1_{k-1}')\mathrm{Diag}_1\} \right|^{\frac{1}{2}}$$

donde $(\text{Diag}_1)_{ii}=(1-\psi_i)^{-2}\ F_k \dfrac{d\psi_i}{dt_i} = \dfrac{\psi_i}{1-\psi_i} F_k = F_i$, ya que en el modelo logístico, la derivada de ψ_i con respecto a t_i es $\psi_i(1-\psi_i)$. Por tanto, aplicando la proposición 2,

$$(\ddot{H}(\theta,x_i))_{rs} = \begin{cases} -F_r(\theta'x_i)\ F_s(\theta'x_i) & \text{si } r\neq s \\ F_r(\theta'x_i)(1-F_r(\theta'x_i)) & \text{si } r=s \end{cases}$$

La distribución final obtenida al utilizar $\pi_{MC}(\theta)$ como inicial es:

$$\pi_{MC}(\theta|D) \propto \left|\sum_{i=1}^{n} x_i x_i' \otimes H(\theta,x_i)\right|^{\frac{1}{2}} \prod_x \prod_{i=1}^{k} \left(F_i(\theta'x)\right)^{n_{ix}}$$

siendo n_{ix} el número de datos, incluídos en el banco D, con vector representante x y pertenecientes a la clase i.

El cálculo de la constante de proporcionalidad implica el cálculo de una integral múltiple para la que, de momento, parece no existir solución analítica. Además, las soluciones numéricas son inviables a no ser que el número de parámetros, $m(k-1)$, sea suficientemente pequeño, no mayor que 3 ó 4. Algo similar ocurre con el vector de medias y la matriz de varianzas.

Por el contrario, la moda si puede calcularse utilizando métodos numéricos, en particular Newton-Raphson, pues se trata de maximizar la función:

$$\text{Log}\ \pi_{MC}(\theta|D) = \text{Cte} + \tfrac{1}{2}\text{Log}\left|\sum_{i=1}^{n} x_i x_i' \otimes H(\theta,x_i)\right| + \sum_x \sum_{i=1}^{n} n_{ix}\ \text{Log}\ F_i(\theta'x)$$

en la que tanto el vector gradiente como la matriz de derivadas segundas parciales de Log $F_i(\theta' x)$ son fáciles de calcular; las derivadas relativas a

$$\left|\sum x_i x_i' \otimes H(\theta, x_i)\right|$$

también son calculables utilizando las fórmulas de derivación para determinantes. (Ver por ejemplo Anderson, 1958, lema 5, pag. 347).

Sin embargo, aunque la moda puede ser calculada con la ayuda de un ordenador, las operaciones que conlleva, en particular el enorme número de determinantes involucrados en la derivada de un determinante, aconsejan la utilización de aproximaciones alternativas.

La segunda aproximación propuesta anteriormente es:

$$\pi_E(\theta) \propto \left|H(\theta, \bar{x})\right|^{m/2}$$

En el ejemplo 6 ya se comprobó que para el modelo logístico aditivo,

$$\left|H(\theta, \bar{x})\right| = \sum_{i=1}^{k} F_i(\theta' \bar{x})$$

por tanto,

$$\pi_E(\theta|D) \propto \prod_{i=1}^{k} \left[F_i(\theta' \bar{x})\right]^{m/2} \prod_{x} \prod_{i=1}^{k} \left[F_i(\theta' x)\right]^{n_{ix}}$$

Al igual que ocurría con $\pi_{MC}(\theta|D)$, el vector de medias y la matriz de covarianzas de $\pi_E(\theta|D)$ no pueden, en general, ser calculados. Sin embargo, la moda puede

ser calculada utilizando sin modificación alguna, cualquier programa de ordenador que calcule el estimador maximo verosímil del modelo logístico. Esto es debido a que $\pi_E(\theta|D)$ puede considerarse como una funcion de verosimilitud modificada al haber añadido ciertos datos especiales, los que integran la función $\pi_E(\theta)$. Por el mismo motivo, se puede aplicar el teorema 3.1.1 con lo que $\pi_E(\theta|D)$ pude aproximarse, para tamaños muestrales relativamente grandes, mediante una distribución Normal con media $\tilde{\theta}$, la moda de $\pi_E(\theta|D)$, y con matriz de precisión la matriz $M(\tilde{\theta})$ con elemento genérico:

$$(M(\tilde{\theta}))_{ij} = - \frac{\partial^2}{\partial\theta_i \partial\theta_j} \text{Log } \pi_E(\tilde{\theta}|D)$$

■■

4.2.3 MODELOS DE TIPO 2

Los modelos de clasificación regulares de Tipo 2 se definieron en el apartado 2.2 como aquellos modelos de clasificación tales que:

$$p(\delta=i|x,\theta) = F_i(t), \qquad i=1,\ldots,k$$

siendo,

$$F_1(t) = \psi_1(t_1)$$

$$F_i(t) = \psi_i(t_i) \prod_{j=1}^{i-1} (1-\psi_j(t_j)) \qquad j=2,\ldots,k-1$$

$$F_k(t) = \prod_{j=1}^{k-1} (1-\psi_j(t_j)) \qquad (4)$$

donde $t=t(\theta,x)=\theta'x\in R^{k-1}$ con $t_i=\theta'_i x$ como componente i

del vector t, y siendo $\{\psi_j(.);\ j=1,\ldots,k-1\}$ funciones posiblemente distintas, definidas de R en el intervalo abierto (0,1) y que cumplen las condiciones exigidas a los modelos de clasificación regulares para dos clases.

PROPOSICION 4

Si $F_\delta(t)$ *corresponde a un modelo de clasificación de Tipo 2, expresión (4), entonces* $H(\theta,x)$ *es una matriz diagonal,* $(k-1)\times(k-1)$*, con elemento genérico:*

$$(H)_{ii} = (1-\psi_i)\ F_i \left[\frac{d}{dt_i}\ Log\ \frac{\psi_i}{1-\psi_i}\right]^2$$

DEMOSTRACION

Sea $H_\delta = -D_t^2(\text{Log } F_\delta(t))$, entonces:

$$(H_\delta)_{ij} = -\frac{\partial^2}{\partial t_i \partial t_j} \text{Log } F_\delta(t) =$$

$$= -\frac{\partial^2}{\partial t_i \partial t_j}\left[\sum_{h=1}^{\delta-1} \text{Log } (1-\psi_h(t_h)) + \text{Log } \psi_\delta(t_\delta)\right]$$

$$= -\frac{\partial^2}{\partial t_i \partial t_j} \text{Log } \psi_\delta(t_\delta) - \sum_{h=1}^{\delta-1} \frac{\partial^2}{\partial t_i \partial t_j} \text{Log } (1-\psi_h(t_h))$$

pero esa última expresión vale cero a no ser que $i=j\leq\delta$, por tanto H_δ es una matriz diagonal. Además:

$$\frac{\partial^2}{\partial t_i^2} \text{Log}(1-\psi_i(t_i)) = \frac{\partial}{\partial t_i}\left[-(1-\psi_i)^{-1}\frac{d\psi_i}{dt_i}\right] =$$

$$= -(1-\psi_i)^{-2}\left[\frac{d\psi_i}{dt_i}\right]^2 - (1-\psi_i)^{-1}\ \frac{d^2\psi_i}{dt_i^2}$$

$$\frac{\partial^2}{\partial t_i^2} \operatorname{Log} \psi_i(t_i) = \frac{\partial}{\partial t_i}\left[\frac{1}{\psi_i}\frac{d\psi_i}{dt_i}\right] = -\left(\frac{1}{\psi_i}\frac{d\psi_i}{dt_i}\right)^2 + \frac{1}{\psi_i}\frac{d^2\psi_i}{dt_i^2}$$

con lo que,

$$(H_\delta)_{ii} = \begin{cases} \dfrac{1}{(1-\psi_i)^2}\left[\dfrac{d}{dt_i}\psi_i\right]^2 + \dfrac{1}{1-\psi_i}\dfrac{d^2}{dt_i^2}\psi_i & \text{si } i<\delta \\ \dfrac{1}{\psi_i^2}\left[\dfrac{d}{dt_i}\psi_i\right]^2 - \dfrac{1}{\psi_i}\dfrac{d^2}{dt_i^2}\psi_i & \text{si } i=\delta \\ 0 & \text{si } i>\delta \end{cases}$$

La matriz H es $\underset{\delta}{E}\{H_\delta\} = \sum F_\delta \operatorname{Log} H_\delta$. Por tanto es una matriz diagonal en la que,

$$(H)_{ii} = \sum_{\delta=1}^{k} F_\delta \operatorname{Log} (H_\delta)_{ii} =$$

$$= F_i \operatorname{Log} (H_i)_{ii} + \sum_{\delta=i+1}^{k} F_\delta \operatorname{Log} (H_\delta)_{ii} =$$

$$= F_i\left[\frac{1}{\psi_i^2}\left[\frac{d}{dt_i}\psi_i\right]^2 - \frac{1}{\psi_i}\frac{d^2}{dt_i^2}\psi_i\right] +$$

$$+ \left[\frac{1}{(1-\psi_i)^2}\left[\frac{d}{dt_i}\psi_i\right]^2 + \frac{1}{1-\psi_i}\frac{d^2}{dt_i^2}\psi_i\right] \sum_{\delta=i+1}^{k} F_\delta$$

pero,

$$\sum_{\delta=i+1}^{k} F_\delta = 1 - \sum_{\delta=1}^{i} F_i = 1 - \{\psi_1 + (1-\psi_1)\psi_2 + \ldots + (1-\psi_1)\ldots(1-\psi_{i-1})\psi_i\} =$$

$$= (1-\psi_1)\ldots(1-\psi_i) = \frac{1-\psi_i}{\psi_i} F_i$$

por tanto:

$$(H)_{ii} = \frac{1}{\psi_i} F_i \left\{ \frac{1}{\psi_i} \left[\frac{d}{dt_i} \psi_i \right]^2 - \frac{d^2}{dt_i^2} \psi_i \right\} +$$

$$+ \frac{1}{\psi_i} F_i \left\{ \frac{1}{1-\psi_i} \left[\frac{d}{dt_i} \psi_i \right]^2 + \frac{d^2}{dt_i^2} \psi_i \right\} =$$

$$= \frac{1}{\psi_i} F_i \left\{ \left[\frac{1}{\psi_i} + \frac{1}{1-\psi_i} \right] \left[\frac{d}{dt_i} \psi_i \right]^2 \right\} =$$

$$= (1-\psi_i) F_i \left[\frac{1}{\psi_i (1-\psi_i)} \frac{d}{dt_i} \psi_i \right]^2$$

lo que demuestra la proposición. ∎

Como corolario inmediato, la distribución inicial de referencia para este tipo de modelos es:

$$\pi(\theta) \propto \left| \underset{x}{E} \{ xx' \otimes H(\theta, x) \} \right|^{\frac{1}{2}}$$

donde la matriz H está definida en la proposición anterior y es tal que:

$$|H(\theta,x)| = \prod_{i=1}^{k-1} \left\{ (1-\psi_i) F_i \left[\frac{d}{dt_i} \operatorname{Log} \frac{\psi_i}{1-\psi_i} \right]^2 \right\} =$$

$$= \prod_{i=1}^{k-1} (1-\psi_i) \prod_{i=1}^{k-1} F_i \left[\prod_{i=1}^{k-1} \frac{d}{dt_i} \operatorname{Log} \frac{\psi_i}{1-\psi_i} \right]^2 =$$

$$= \prod_{i=1}^{k} F_i(\theta' x) \left[\prod_{i=1}^{k-1} \frac{d}{dt_i} \operatorname{Log} \frac{\psi_i}{1-\psi_i} \right]^2$$

que coincide con el determinante obtenido para los modelos de Tipo 1.

Una propiedad interesante, consecuencia de la proposición 4, es el carácter diagonal de la matriz $H(\theta,x)$; esta propiedad puede simplificar de forma considerable el cálculo de $\pi(\theta)$. En efecto, aunque el producto de Kronecker no es conmutativo, la diferencia entre $A\otimes B$ y $B\otimes A$ estriba en una simple reordenación de las filas y las columnas, por tanto (Rao, 1973, apartado 1b.2), existen dos matrices no singulares, P_1 y P_2, dependientes exclusivamente de la dimensión de las matrices A y B, tales que $B\otimes A=P_1(A\otimes B)P_2$. Por tanto,

$$xx'\otimes H(\theta,x) = P_1(H(\theta,x)\otimes xx')P_2 \Rightarrow \underset{x}{E}\{xx'\otimes H(\theta,x)\} =$$

$$= P_1\underset{x}{E}\{H\otimes xx'\}P_2 \Rightarrow \left|\underset{x}{E}\{xx'\otimes H\}\right| = |P_1||P_2|\left|\underset{x}{E}\{H\otimes xx'\}\right| \Rightarrow$$

$$\Rightarrow \pi(\theta) \propto \left|\underset{x}{E}\{H(\theta,x)\otimes xx'\}\right|^{\frac{1}{2}}$$

Ahora bien, como la matriz H es diagonal, la matriz $H\otimes xx'$ tendrá todos sus elementos no nulos concentrados en $k-1$ bloques $m\times m$ situados en la diagonal, por tanto:

$$\pi(\theta) \propto \left|\underset{x}{E}\{H(\theta,x)\otimes xx'\}\right|^{\frac{1}{2}} = \prod_{i=1}^{k-1}\left|\underset{x}{E}\{(H(\theta,x))_{ii}xx'\}\right|^{\frac{1}{2}}$$

con lo que se reduce de forma importante la dimensión de los determinantes involucrados.

EJEMPLO 8. *Modelo logístico multiplicativo.*

El modelo logístico multiplicativo es un modelo de Tipo 2 en el que,

$$\psi_i(y) = \psi(y) = \exp(y)/(1+\exp(y)),\ y\in\mathbb{R},\ i=1,\ldots,k-1$$

por tanto, $\text{Log}\dfrac{\psi_i(t_i)}{1-\psi_i(t_i)} = t_i$, luego $\dfrac{d}{dt_i}\text{Log}\dfrac{\psi_i(t_i)}{1-\psi_i(t_i)} = 1.$

En consecuencia, la distribución inicial de referencia para el modelo logístico multiplicativo, es:

$$\pi(\theta) \propto \prod_{i=1}^{k-1} \left| \mathop{E}_{x} \{ (1-\psi_i(\theta'_{\cdot i} x)) F_i(\theta' x) x x' \} \right|^{\frac{1}{2}}$$

Las dos aproximaciones a esta distribución, propuestas en el apartado anterior, son:

$$\pi_{MC}(\theta) \propto \prod_{i=1}^{k-1} \left| \sum_{j=1}^{n} \{ (1-\psi_i(\theta'_{\cdot i} x_j)) F_i(\theta' x_j) x_j x_j' \} \right|^{\frac{1}{2}}$$

$$\pi_E(\theta) \propto \left| H(\theta, \bar{x}) \right|^{m/2} \propto \prod_{i=1}^{k} \left(F_i(\theta' \bar{x}) \right)^{m/2}$$

que proporcionan como distribuciones finales de referencia:

$$\pi_{MC}(\theta|D) \propto \prod_{i=1}^{k-1} \left| \sum_{j=1}^{n} (1-\psi_i) F_i \; x_j x_j' \right|^{\frac{1}{2}} \prod_{x} \prod_{\delta=1}^{k} \left(F_\delta(\theta' x) \right)^{n_{\delta x}}$$

$$\pi_E(\theta|D) \propto \prod_{i=1}^{k} \left(F_i(\theta' \bar{x}) \right)^{m/2} \prod_{x} \prod_{\delta=1}^{k} \left(F_\delta(\theta' x) \right)^{n_{\delta x}}$$

donde $n_{\delta x}$ se define como el número de datos en D que, pertenecientes a la clase δ, tienen vector representante x.

Al igual que ocurría con el modelo logístico aditivo, los momentos de $\pi_{MC}(\theta|D)$ y $\pi_E(\theta|D)$ no son calculables a no ser que la dimensión del espacio paramétrico, Θ, sea muy pequeña, no mayor que 3 ó 4. Además, el cálculo de la moda de $\pi_{MC}(\theta|D)$ requiere un número de operaciones que, aunque muy inferior que en el modelo logístico aditivo, sigue siendo excesivo para su aplicación práctica.

Por el contrario, la moda de la distribución $\pi_E(\theta|D)$ si que es calculable de forma sencilla utilizando Newton-Raphson. En efecto, $\pi_E(\theta|D)$ puede expresarse como

$$\pi_E(\theta|D) \propto \prod_x \prod_{\delta=1}^{k} \left\{ F_\delta(\theta' x) \right\}^{n_{\delta x}}$$

donde el productorio con respecto a x incluye a todos los vectores representantes del banco de datos así como también a $\bar{x}$; por su parte, $n_{\delta x}$ se define como anteriormente excepto $n_{\delta \bar{x}}$ que valdrá $m/2$ mas el número de datos con vector representante $\bar{x}$, pertenecientes a la clase δ.

Con esta notación:

$$\text{Log } \pi_E(\theta|D) = C + \sum_x \sum_{\delta=1}^{k} n_{\delta x} \text{ Log } F_\delta(\theta' x)$$

luego:

$$\frac{\partial}{\partial \theta_{ij}} \text{Log } \pi_E(\theta|D) = \sum_x \sum_{\delta=1}^{k} n_{\delta x} \frac{\partial}{\partial \theta_{ij}} \text{Log } F_\delta(\theta' x)$$

$$\frac{\partial^2}{\partial \theta_{i_1 j_1} \partial \theta_{i_2 j_2}} \text{Log } \pi_E(\theta|D) = \sum_x \sum_{\delta=1}^{k} n_{\delta x} \frac{\partial^2}{\partial \theta_{i_1 j_1} \partial \theta_{i_2 j_2}} \text{Log } F_\delta(\theta' x)$$

Aprovechando los cálculos utilizados en la demostración de la proposición 4, y aplicándolos al modelo logístico multiplicativo:

$$\frac{\partial}{\partial \theta_{ij}} \text{Log } F_\delta(\theta' x) =$$

$$= \begin{cases} -x_i \left[1+\exp(\theta'_{\cdot j} x)\right]^{-1} & \text{si } j<\delta \\ x_i \left[\exp(\theta'_{\cdot j} x)(1+\exp(\theta'_{\cdot j} x))\right]^{-1} & \text{si } j=\delta \\ 0 & \text{si } j>\delta \end{cases}$$

mientras que,

$$\frac{\partial^2}{\partial \theta_{i_1 j_1} \partial \theta_{i_2 j_2}} \text{Log } F_\delta(\theta' x) =$$

$$= \begin{cases} x_{i_1} x_{i_2} \left[1+\exp(\theta'_{\cdot j} x)\right]^{-1} \exp(\theta'_{\cdot j} x) & \text{si } j_1=j_2<\delta \\ -x_{i_1} x_{i_2} \left[\exp(\theta'_{\cdot j} x)\right]^{-1} \left[1+\exp(\theta'_{\cdot j} x)\right]^{-2} (1+2\exp(\theta'_{\cdot j} x)) & \text{si } j_1=j_2=\delta \\ 0 \qquad \text{en otro caso} & \end{cases}$$

con lo que el método de Newton-Raphson es fácilmente aplicable.

■■

CAPITULO 5

COMPARACION CON METODOS ALTERNATIVOS

Los resultados teóricos obtenidos en el capítulo cuatro se comparan con algunos de los métodos alternativos mas utilizados en la práctica. Para la comparación de métodos alternativos es necesario definir unos criterios de evaluación que permitan medir la bondad de las predicciones obtenidas por los diferentes métodos. Se han utilizado cuatro criterios distintos: dos funciones de evaluación propia, cuadrática y logarítmica; y dos funciones no propias pero ampliamente difundidas, el porcentaje de clasificaciones correctas y la probabilidad media asignada a la categoría correcta.

Se estudian tres ejemplos numéricos concretos. En el primero, datos simulados a partir de un modelo logístico, todas las componentes del vector representante son discretas. En el segundo, datos demográficos estudiados por Press and Wilson (1978), existen indicadores discretos y continuos. Por último, se estudian los datos proporcionados por Fisher (1936), datos en los que el vector representante puede considerarse multinormal.

5.1 METODOS ALTERNATIVOS

En el capítulo anterior se ha propuesto $\pi_E(\theta|D)$ como distribución final de referencia para el modelo logístico aditivo. (Ejemplo 4.2.7). La distribución de clasificación se obtiene entonces calculando la distribución predictiva, esto es,

$$p(\delta|x,D) = \int p(\delta|x,\theta)\ \pi_E(\theta|D)\ d\theta$$

Ahora bien, esa integral no puede ser calculada de forma analítica. De nuevo es necesaria la búsqueda de aproximaciones.

Una primera aproximación, denominada en el resto del capítulo *clasificación logística de referencia 1*, RLC1, consiste en sustituír el parámetro desconocido θ, por la moda de la distribución $\pi_E(\theta|D)$, θ^*. Así,

$$p(\delta|x,D) \cong p(\delta|x,\theta^*)$$

Sin embargo, argumentos similares a los que proporcionan el teorema 3.1.1. sobre normalidad asintótica, pueden ser invocados para proponer la moda de $\pi_E(\theta|D)$, θ^* como un estimador del valor esperado de θ, mientras que la matriz $(D^2_\theta(-\text{Log}\pi_E(\theta^*|D)))^{-1}$ puede ser considerada, al menos aproximadamente, como un estimador de la varianza de θ. Sustituyendo los verdaderos valores por estas aproximaciones y utilizando la generalización multidimensional del teorema 3.4.1 en Lindley (1965):

$$p(\delta|x,D) = \int p(\delta|x,\theta)\ \pi_E(\theta|D)\ d\theta = \underset{\theta|D}{E}\{p(\delta|x,\theta)\}$$

$$\cong p(\delta|x,E(\theta)) + \tfrac{1}{2}tr\left\{D^2_\theta(p(\delta|x,E(\theta)))\ VAR(\theta)\right\}$$

$$\cong p(\delta|x,\theta^*) + \tfrac{1}{2}tr\left\{D^2_\theta(p(\delta|x,\theta^*))\ (D^2_\theta(-\text{Log}\pi_E(\theta^*|D))^{-1}\right\}$$

Para el modelo logístico aditivo, como ya se discutió en el ejemplo 4.2.7, los valores de θ^* y $D^2_\theta(-\text{Log}\ \pi_E(\theta^*|D))$ se pueden conseguir mediante el método de Newton-Raphson, mientras que la matriz $D^2_\theta(p(\delta|x,\theta^*))$ es fácilmente calculable. Por tanto, este método, *clasificación logística de referencia 2* (RLC2) puede mejorar considerablemente los resultados del método RLC1 sin añadir demasiadas complicaciones de cálculo.

En este capítulo se comparan el RLC1 y el RLC2 con los dos métodos clásicos mas importantes, *clasificación logística maximo verosímil* (LML) y *análisis lineal discriminante* (LDA), y con dos métodos bayesianos, *clasificación normal bayesiana* (BNC) y *clasificación probabilística bayesiana* (BPC), método propuesto recientemente por J. M. Bernardo.

La clasificación logística maximo verosímil (Anderson, 1972) consiste en el cálculo del máximo de la función de verosimilitud obtenida suponiendo el modelo logístico, $\hat{\theta}$, y la utilización de este estimador como si fuese el verdadero valor del parámetro, esto es,

$$p(\delta|x,D) \cong p(\delta|x,\hat{\theta})$$

El análisis linear discriminante (Fisher, 1936; Anderson, 1958) propone, como estimadores de los parámetros θ, a los valores $\hat{\theta}$ calculados como:

$$\hat{\theta}_{i1} = -\tfrac{1}{2}(\bar{y}_i - \bar{y}_k)' \; S^{-1}(\bar{y}_i + \bar{y}_k)$$

$$\hat{\theta}_i = (\bar{y}_i - \bar{y}_k)' \; S^{-1}$$

donde $\bar{y}_i$, $i=1,\ldots,k$, son las medias muestrales de los vectores de indicadores en cada una de las k clases y S es la matriz de varianzas-covarianzas de todos los vectores de indicadores presentes en el banco de datos. La distribución de clasificación propuesta por el método es,

$$p(\delta|x,D) \cong p(\delta|x,\hat{\theta})$$

La clasificación normal bayesiana (Geisser, 1964; Bernardo,1978) se basa en la hipótesis de multinormalidad para el vector representante, en cada una de las clases. Es, por tanto, un método encuadrado en el enfoque muestral. Una vez calculadas las distribuciones predictivas de los vectores representantes en cada una de las clases, $p(x|\delta,D)$, la distribución de clasificación se obtiene como

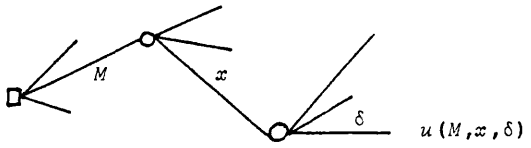
$$p(\delta|x,D) \propto p(x|\delta,D)\; p(\delta)$$

Por último, el método de clasificación probabilística bayesiana, propuesto por Bernardo (1983), debilita la hipótesis de multinormalidad asumida por BNC suponiendo la normalidad solo para determinada combinación lineal de los vectores representantes, $\hat{\theta}'x$, siendo $\hat{\theta}$ precisamente el estimador proporcionado por LDA. Suponiendo que esa combinación lineal es un estadístico suficiente de los datos para el problema de clasificación, se puede obtener la distribución de clasificación exactamente igual que BNC pero utilizando esa combinación lineal como único elemento del vector representante.

5.2 CRITERIOS DE EVALUACION

El problema de comparar métodos estadísticos alternativos se enmarca de forma natural dentro de la estructura bayesiana de la teoría de la decisión. Así, una herramienta para medir la bondad de los diferentes métodos puede encontrarse en el Teorema de Maximización de la Utilidad Esperada.

En particular, la evaluación de distintos métodos de clasificación puede plantearse a través del siguiente árbol de decisión:



Una vez especificadas las utilidades y probabilidades que intervienen en el problema, cada método M vendrá evaluado mediante su utilidad esperada,

$$u^*(M) = \underset{x}{E}\ \underset{\delta|x}{E}\ u(M,x,\delta).$$

Considérese las siguientes funciones de utilidad, donde $p(\delta|x,M)$ representa la distribución de clasificación proporcionada por M:

(i) *Porcentaje de clasificaciones correctas:*

$$u_1(M,x,\delta) = 1 \quad \text{si} \quad p(\delta|x,M) = \max_i\{p(i|x,M)\}$$

$$= 0 \quad \text{en otro caso}$$

(ii) *Probabilidad asignada a la categoría correcta:*

$$u_2(M,x,\delta) = p(\delta|x,M)$$

(iii) *Utilidad cuadrática:*

$$u_3(M,x,\delta) = \frac{k}{k-1}\left[2p(\delta|x,M) - \| p(\cdot|x,M) \|^2\right] - \frac{1}{k-1}$$

$$\text{siendo} \quad \| p(\cdot|x,M) \|^2 = \sum_{i=1}^{k} \left[p(i|x,M)\right]^2$$

(iv) *Utilidad logarítmica acotada:*

$$u_4(M,x,\delta) = 1 + \frac{\text{Log } p(\delta|x,M)}{\text{Log } k} \quad \text{si} \quad p(\delta|x,M) > \varepsilon > 0$$

$$= 1 + \text{Log}\varepsilon \,/\, \text{Log } k \quad \text{en otro caso}$$

La función (i) puede encontrarse en la literatura definida de manera ligeramente distinta. En efecto, diversos autores asignan utilidad cero a aquellos casos en los que $p(\delta|x,M) = \max\{p(i|x,M)\}$ pero existe alguna otra clase cuya probabilidad también alcanza ese máximo; por el contrario, otros autores asignan, en ese tipo de situaciones, utilidad uno partido por el número de clases para las que se alcanza el máximo. Por otra parte, las funciones (iii) y (iv) pueden definirse, con mas genera-

lidad, en función de ciertas constantes de normalización.

Los valores empleados en la definición anterior para dichas constantes han sido elegidos de manera que se obtiene utilidad uno si se asigna probabilidad uno a la clase correcta, mientras que se obtiene utilidad cero con una distribución de clasificación uniforme.

Las dos primeras funciones de utilidad no son funciones de evaluación (*scoring rules*) propias, por tanto su uso no garantiza que el método que maximiza la utilidad esperada sea el que proporcione la *verdadera* distribución de clasificación. Debido a su amplia difusión, se han calculado en los ejemplos numéricos de este capítulo; sin embargo, el énfasis en la discusión de los resultados se sitúa en las otras dos funciones de utilidad.

Tanto la utilidad cuadrática (Brier, 1950) como la logarítmica *no* acotada (Good, 1950) son funciones *propias*. A un nivel práctico es necesario acotar la función logarítmica pues, aunque teóricamente $p(\delta|x,M)$ no puede valer cero, los errores de redondeo proporcionan con frecuencia valores cero para $p(\delta|x,M)$, siendo en tal caso imposible el cálculo de Log $p(\delta|x,M)$. La cota utilizada en esta memória ha sido $\varepsilon=0.0005$.

La utilidad cuadrática proporciona una función de utilidad acotada, tomando valores en el intervalo $\left(-\frac{k+1}{k-1}, 1\right]$; en particular, para dos clases, $k=2$, ese intervalo es $\left(-3,1\right]$. El valor mas bajo, $-\frac{k+1}{k-1}$, corresponde a una dis-

tribución degenerada que asigna probabilidad uno a una clase incorrecta; mientras que el valor 1 se obtiene con una clasificación perfecta, esto es, asignando probabilidad uno a la clase correcta. Como resultado de referencia, la distribución uniforme proporciona una utilidad cuadrática cero.

La utilidad logarítmica acotada, u_4, tiene un comportamiento similar al de la cuadrática, comentado en el párrafo anterior. La única diferencia aparece en la cota inferior de u_4, que es 1+Log ε /Log k. Con ε=0.0005 y dos clases, k=2, esta cota es -9.9658.

La utilidad logarítmica es una función de evaluación *local*, esto es, solo depende de la probabilidad asociada a la categoría correcta. Por el contrario, la utilidad cuadrática también depende de la manera en la que se distribuye el resto de la unidad de probabilidad entre las demás categorías.

Una vez especificada la función de utilidad, si la distribución conjunta $p(x,\delta)$ fuese conocida, la utilidad esperada $u^*(M)$ podría ser fácilmente calculable. Sin embargo, si se desea elegir un método de clasificación es porque se desconoce la distribución $p(\delta|x)$, y por tanto también la distribución conjunta $p(x,\delta)$. Bernardo (1983), en el contexto de comparación de métodos de clasificación, y Bernardo and Bermúdez (1984), en selección de variables, proponen dividir el banco de datos, D, en dos subbancos, D_1 y D_2. Con el primero, D_1, se calculan las distribuciones de clasificación relativas a los distintos métodos, resevando D_2 para su evaluación. Así, si los datos provienen de un muestreo prospectivo, D_2 es una muestra aleatoria de la distribución $P(x,\delta)$, y por tanto puede

utilizarse para obtener, mediante integración por Monte-Carlo,

$$u^*(M) = \mathop{E}_{x,\delta} u(x,\delta,M) \cong \frac{1}{n_2} \sum_{i=1}^{n_2} u(x_i,\delta_i,M)$$

siendo n_2 el número de datos en $D_2 \equiv \{(x_i,\delta_i),\ i=1,\ldots,n_2\}$. Por el contrario, si los datos provienen de un muestreo retrospectivo, entonces D_2 está formado por muestras aleatorias de las distribuciones $p(x|\delta)$, por tanto, si se conocen las probabilidades $p(\delta)$, la aproximación por Monte-Carlo proporciona:

$$u^*(M) \cong \sum_{j=1}^{k} p(\delta=j) \frac{1}{n_{2j}} \sum_{i=1}^{n_{2j}} u(x_{ij},\delta=j,M)$$

siendo n_{2j} el número de datos en D_2 que pertenecen a la clase j, y $\{x_{ij},\ i=1,\ldots,n_{2j}\}$ los vectores representantes de dichos datos.

De forma totalmente similar se pueden calcular la esperanza de los cuadrados de las utilidades, y con ello la varianza de la utilidad.

Sin embargo, estas estimaciones pueden mejorarse sensiblemente sin un incremento excesivo en el cálculo. En efecto, considerando el muestreo prospectivo, las utilidades obtenidas por los n_2 elementos del banco D_2 pueden ser consideradas una muestra aleatoria de tamaño n_2 de la población *unidimensional* de utilidades. Por tanto, la utilidad esperada puede obtenerse haciendo inferencias sobre la media de una población unidimensional,

en lugar de estudiar la esperanza, con respecto a la cantidad aleatoria multidimensional (x,δ), de la función de utilidad.

Así por ejemplo, el porcentaje de clasificación correcta, n_1, solo toma los valores 0 y 1. Por tanto, los n_2 valores producidos por los elementos del banco D_2 pueden considerarse una muestra aleatoria de una distribución binomial del parámetro u_1^*, esto es, la utilidad esperada que se desea estudiar. La distribución de referencia sobre el parámetro de una distribución binomial, (ver por ejemplo, Bernardo, 1981), es Beta con parámetros $r+\frac{1}{2}$ y $n_2-r+\frac{1}{2}$, siendo r el número de aciertos, $u_1=1$, obtenidos entre los datos de D_2.

De esta forma no solo puede proporcionarse un estimador de la utilidad esperada, la media de esa distribución beta, $r+\frac{1}{2}/(n_2+1)$, sino también la varianza de la utilidad esperada, $\left((r+\frac{1}{2})\ (n_2-r+\frac{1}{2})\right)/\left((n_2+1)^2\ (n_2+2)\right)$.

Es de destacar que el estimador de la utilidad esperada, u_1^*, así obtenido, es el estimador bayesiano usual de la probabilidad de un suceso. El obtenido por Monte-Carlo es r/n_2, que es el estimador maximo verosímil de la probabilidad de un suceso. Por otra parte, por Monte-Carlo se obtenía un estimador de la varianza de la población de utilidades, mientras que el estudio directo de esa población permite obtener la varianza de la utilidad esperada, dando con ello una medida de la fiabilidad del resultado obtenido.

Al estudiar cualquiera de las otras tres funciones de utilidad, u_2, u_3 o u_4, el problema se complica pues

es necesario modelizar una cantidad aleatoria continua. Además, las aproximaciones asumidas por los distintos métodos producen una ficticia eliminación de incertidumbre; el ejemplo mas llamativo lo constituye, sin duda alguna, la sustitución de un parámetro desconocido por un estimador suyo, actuando *como si* se conociera con certeza ese parámetro. Por este motivo es de esperar unas probabilidades de clasificación mas *seguras*, esto es, mas alejadas de la distribución uniforme, de lo que los datos pueden garantizar. En consecuencia, es de esperar que la población de utilidades posea un carácter bimodal, una moda correspondiendo a aciertos importantes y la otra a errores graves.

Una forma de estudiar una población bimodal consiste en dividirla en dos subpoblaciones y estudiar cada una por separado. Así en este caso, si ξ es la cantidad aleatoria dicotómica que representa los aciertos ($\xi=1$) y los fracasos ($\xi=0$), se podría calcular la media y la varianza de la cantidad aleatoria $u^*|\xi$, esto es, la utilidad esperada condicionada a éxito o fracaso. De hecho, la cantidad aleatoria ξ es precisamente la utilidad u_1 estudiada con antelación. Las propiedades predictivas de los operadores esperanza y varianza, permiten calcular:

$$E(u^*) = \underset{\xi}{E}\, E\,(u^*|\xi) = E(u^*|\xi=1)\; p(\xi=1) + E(u^*|\xi=0)\; p(\xi=0)$$

$$= E(u^*|\xi=0) + (E(u^*|\xi=1) - E(u^*|\xi=0)$$

$$p(\xi=1)) \qquad (1)$$

$$VAR(u^*) = \underset{\xi}{E}\, VAR(u^*|\xi) + \underset{\xi}{VAR}\, E(u^*|\xi) = VAR(u^*|\xi=1)\; p(\xi=1) +$$

$$+ \mathrm{VAR}(u^*|\xi=0)\ p(\xi=0) + (\mathrm{E}(u^*|\xi=1) - \mathrm{E}(u^*))^2$$

$$p(\xi=1) + (\mathrm{E}(u^*|\xi=0) - \mathrm{E}(u^*))^2\ p(\xi=0)$$

$$= \mathrm{VAR}(u^*|\xi=1)\ p(\xi=1) + \mathrm{VAR}(u^*|\xi=0)\ p(\xi=0) +$$

$$+ (\mathrm{E}(u^*|\xi=1) - \mathrm{E}(u^*|\xi=0))^2\ \ p(\xi=1)\ p(\xi=0) \qquad (2)$$

Donde $p(\xi=1)$ es prescisamente u_1^*, luego esas expresiones dependen del valor u_1^*; por tanto, las fórmulas (1) y (2) corresponden en realidad a $\mathrm{E}(u^*|u_1^*)$ y $\mathrm{VAR}(u^*|u_1^*)$. Una nueva aplicación de las propiedades de los operadores esperanza y varianza proporciona:

$$\mathrm{E}(u^*) = \underset{u_1^*}{\mathrm{E}}\ \mathrm{E}(u^*|u_1^*) =$$

$$= \mathrm{E}(u^*|\xi=1)\ \mathrm{E}(u_1^*) + \mathrm{E}(u^*|\xi=0)\ (1-\mathrm{E}(u_1^*)) \qquad (3)$$

$$\mathrm{VAR}(u^*) = \underset{u_1^*}{\mathrm{E}}\ \mathrm{VAR}(u^*|u_1^*) + \underset{u_1^*}{\mathrm{VAR}}\ \mathrm{E}(u^*|u_1^*) =$$

$$= \mathrm{VAR}(u^*|\xi=1)\ \mathrm{E}(u_1^*) + \mathrm{VAR}(u^*|\xi=0)\ (1-\mathrm{E}(u_1^*) +$$

$$+ (\mathrm{E}(u^*|\xi=1) - \mathrm{E}(u^*|\xi=0))^2\ \mathrm{E}(u_1^*(1-u_1^*)) +$$

$$+ (\mathrm{E}(u^*|\xi=1) - \mathrm{E}(u^*|\xi=0))^2\ \mathrm{VAR}(u_1^*)$$

$$= \mathrm{VAR}(u^*|\xi=1)\ \mathrm{E}(u_1^*) + \mathrm{VAR}(u^*|\xi=0)\ (1-\mathrm{E}(u_1^*)) +$$

$$+ (\mathrm{E}(u^*|\xi=1) - \mathrm{E}(u^*|\xi=0))^2\ \mathrm{E}(u_1^*)\ (1-\mathrm{E}(u_1^*)) \qquad (4)$$

En consecuencia, estudiando por separado las pobla-

ciones de aciertos y fracasos, se obtiene la esperanza y la varianza de la utilidad esperada mediante unas fórmulas de fácil cálculo. Sin embargo, al tratarse de una distribución bimodal, su varianza puede resultar menos informativa que las varianzas de las dos subpoblaciones de aciertos y fracasos, por lo que puede resultar aconsejable no solamente proporcionar $E(u^*)$ y $VAR(u^*)$ sino también $VAR(u^*|\xi=0)$ y $VAR(u^*|\xi=1)$.

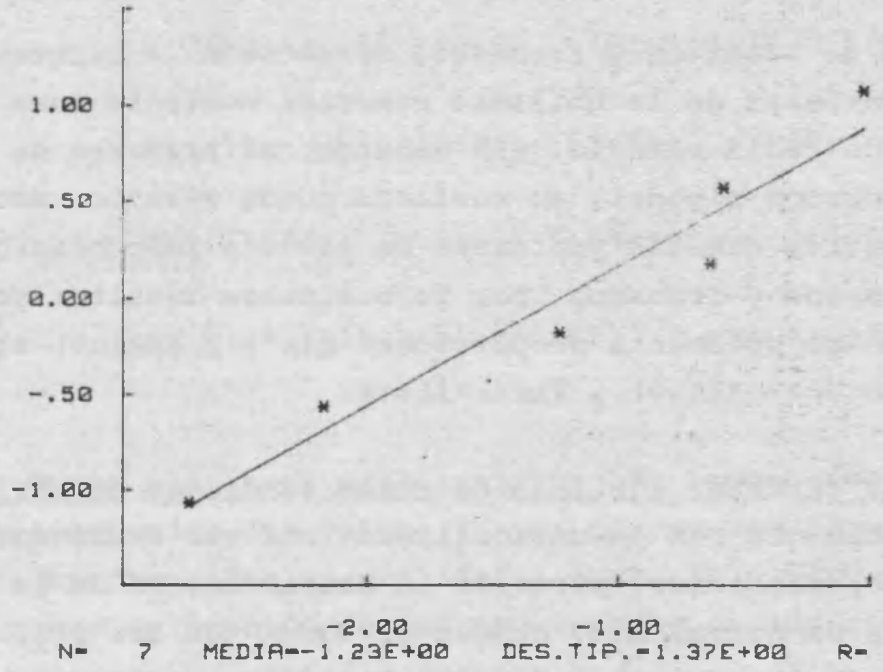
El carácter continuo de estas funciones de utilidad conjuntamente con su unimodalidad, una vez separadas las dos subpoblaciones, permiten la consideración de la hipótesis de normalidad. Aunque el rango de las utilidades es finito, el test de normalidad en papel probabilístico, ver figuras 1 a 4, no permite rechazar la hipótesis de normalidad en los casos en los que se ha contrastado, bancos de datos correspondientes a los ejemplos del apartado 5.3. Posiblemente, con otros conjuntos de datos podría aconsejarse una transformación normalizadora de las utilidades.

Una vez asumida la hipótesis de normalidad, la distribución final de referencia sobre la media de una población normal (ver, por ejemplo, Bernardo, 1981), proporciona:

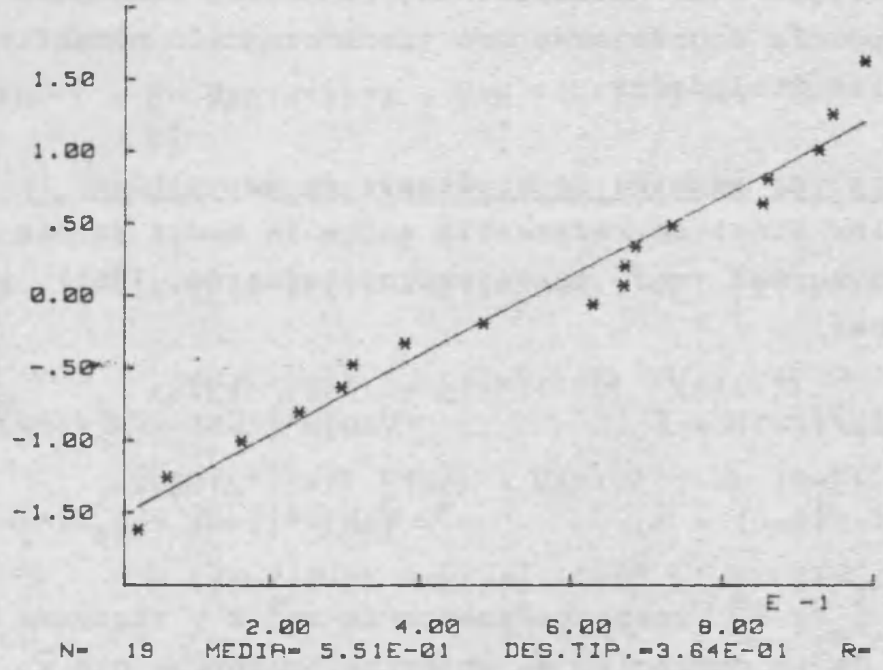
$$E(u^*|\xi=1) = \bar{u}_a \qquad VAR(u^*|\xi=1) = s_a^2/(r-3)$$

$$E(u^*|\xi=0) = \bar{u}_f \qquad VAR(u^*|\xi=0) = s_f^2/(n_2-r-3)$$

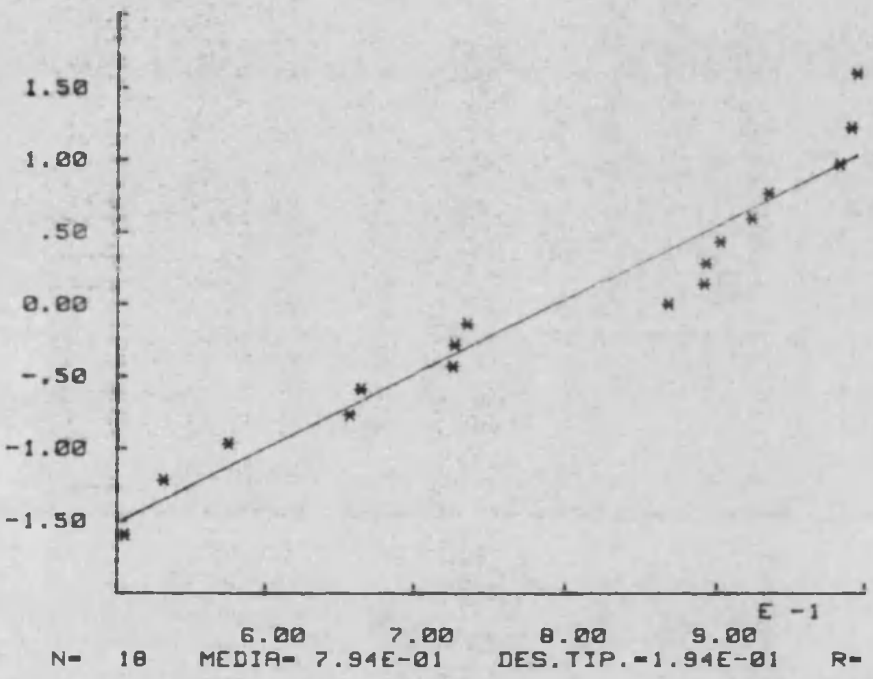
siendo $\bar{u}_a$ y s_a^2, respectivamente la media y varianza muestrales de la población de aciertos, mientras que $\bar{u}_f$ y s_f^2 representan los de la población de fracasos. Al igual que



- Fig. 1 -
-Test de normalidad para las utilidades u_3 correspondientes a fracasos del método LML aplicado a los datos del ejemplo 5.3.2.-

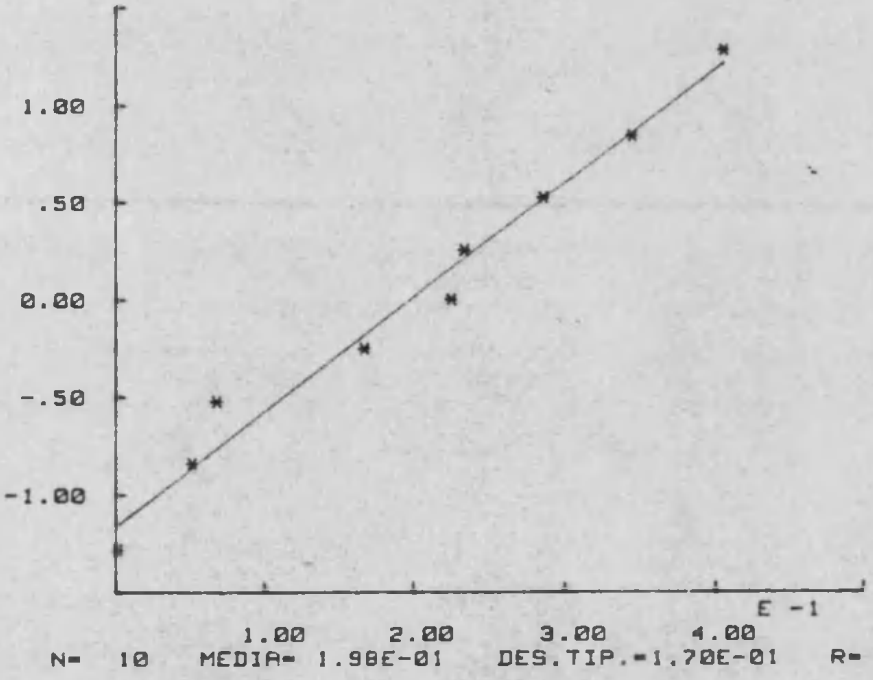


- Fig. 2 -
-Test de normalidad para las utilidades u_4 correspondientes a aciertos del método LDA aplicado a los datos del ejemplo 5.3.2.-



- Fig. 3 -

-Test de normalidad para las utilidades u_2 correspondientes a aciertos del método RLC2 aplicado a los datos del ejemplo 5.3.2.-



- Fig. 4 -

-Test de normalidad para las utilidades u_2 correspondientes a fracasos del método BNC aplicado a los datos del ejemplo 5.3.2.-

anteriormente, n_2 es el número de datos en D_2 y r, el número de aciertos.

Trasladando estos resultados a las fórmulas (3) y (4)

$$E(u^*) = \bar{u}_a \frac{r+\frac{1}{2}}{n_2+1} + \bar{u}_f \frac{n_2-r+\frac{1}{2}}{n_2+1} \qquad (5)$$

$$VAR(u^*) = \frac{S_a^2}{r-3} \frac{r+\frac{1}{2}}{n_2+1} + \frac{S_b^2}{n_2-r-3} \frac{n_2-r+\frac{1}{2}}{n_2+1} + (\bar{u}_a-\bar{u}_f) \frac{(r+\frac{1}{2})(n_2-r+\frac{1}{2})}{(n_2+1)^2} \qquad (6)$$

La utilidad esperada conseguida por Monte-Carlo es:

$$u^* \cong \frac{1}{n_2} \sum_{i=1}^{n_2} u_i = \frac{1}{n_2} (\bar{u}_a r + \bar{u}_f (n_2-r)) =$$

$$= \bar{u}_a \frac{r}{n_2} + \bar{u}_f \frac{n_2-r}{n_2} \qquad (7)$$

que es muy parecida a la proporcionada por (5), la única diferencia reside en el uso, en la expresión (7), del estimador maximo verosímil, r/n_2, en lugar de la esperanza de la distribución de referencia, $(r+\frac{1}{2})/(n_2+1)$.

Las expresiones (5) y (6) corresponden a un muestreo prospectivo. Si el muestreo fuese retrospectivo, sería necesario calcular (5) y (6) en cada clase, δ, por separado. Posteriormente, una nueva aplicación de la esperanza y varianza predictivas proporcionaría $E(u^*)$ y $VAR(u^*)$.

5.3 EJEMPLOS NUMERICOS

En este apartado se comparan los seis métodos de clasificación introducidos en el apartado 5.1, (RLC1, RLC2, LML, LDA, BPC y BNC), utilizando para ello los métodos de evaluación comentados en el apartado anterior.

EJEMPLO 1. *Datos simulados.*

En el banco 1 del apéndice 2 se recogen 100 datos correspondientes a un problema de clasificación en el que existen dos clases alternativas, $k=2$, y siete indicadores dicotómicos; así, considerando que el vector representante, x, está formado por los indicadores mas un término independiente, el número de componentes de x es ocho, $m=8$.

Estos datos corresponden a una muestra simulada, obtenida a partir de una distribución uniforme para los vectores representantes, $p(x) = 2^{-7} = \text{cte.}$, y una distribución de clasificación, $p(\delta|x,\theta)$, logística con parámetro $\theta \equiv (0,3,2,1,-3,-2,-1,0)'$.

Se realizaron tres estudios con tamaños muestrales diferentes para comprobar el comportamiento de los distintos métodos en relación con el número de datos. Así, en el primer estudio solo se utilizaron los 50 primeros datos, $n=50$; el segundo estudio se realizó con los 75 primeros datos, $n=75$; y por último, en el tercer estudio se trabajó con el banco completo, $n=100$.

Como la distribución de los vectores representantes, $p(x)$, es conocida, las utilidades esperadas para cada método concreto pueden calcularse exactamente. No es pre-

ciso recurrir a las estimaciones propuestas en el apartado 5.2.

La tabla 1 recoge los resultados obtenidos en los tres estudios, así como las utilidades esperadas que se obtienen utilizando la distribución de clasificación correcta, esto es, la distribución logística con parámetro $\theta \equiv (0,3,2,1,-3,-2,-1,0)'$.

	n=50				n=75				n=100			
	u_1^*	u_2^*	u_3^*	u_4^*	u_1^*	u_2^*	u_3^*	u_4^*	u_1^*	u_2^*	u_3^*	u_4^*
RLC1	.7899	.7773	.3491	-.0022	.7960	.7543	.3891	.2870	.8030	.7576	.4537	.3823
RLC2	.7899	.7670	.3759	.1430	.7960	.7474	.4014	.3157	.8030	.7517	.4580	.3908
LML	.8030	.7917	.3389	-.2762	.7948	.7566	.3870	.2780	.8030	.7594	.4386	.3592
LDA	.7911	.7707	.3455	.1256	.7839	.7590	.3702	.2429	.8030	.7564	.4323	.3555
BPC	.7911	.7653	.3613	.2126	.7852	.7542	.3708	.2556	.7911	.7520	.4230	.3360
BNC	.7579	.7362	.2979	.1519	.7723	.7274	.2864	.1326	.7495	.7187	.2893	.1734
Distribución correcta					.8175	.7564	.5128	.4503				

- Tabla 1 -

En los dos últimos estudios, n=75 y n=100, las utilidades cuadrática y logarítmica máximas corresponden a RLC2, seguido por RLC1. En concreto, para n=100, las utilidades obtenidas con RLC2 solo distan seis centésimas de las proporcionadas por la distribución correcta.

En el primer estudio, n=50, es de destacar el buen comportamiento de BNC en relacción con los demás métodos.

Sobre todo habida cuenta de que no se cumplen las hipótesis en las que se basa el método BNC, multinormalidad del vector representante. La utilidad logarítmica, u_4^*, de BNC solo es superada por la de BPC quedando RLC2 en tercer lugar. Esto es debido a que las probabilidades proporcionadas por BNC son mas conservadoras, i.e. mas cercanas a la distribución uniforme, que las proporcionadas por los demás métodos; así, aunque su porcentaje esperado de aciertos, u_1^*, es menor, sus fallos son penalizados también menos fuertemente que los de los demás métodos, obteniendo una utilidad esperada, u_4^*, relativamente alta. La ventaja proporcionada por este conservadurismo del método BNC desaparece en cuanto crece el número de datos; esto es debido a un aumento, al aumentar el número de datos, de la bondad de las aproximaciones utilizadas en los demás métodos. Así, las utilidades conseguidas por BNC con n=75 y n=100 son prácticamente iguales que las conseguidas con n=50. Por el contrario, los demás métodos aumentan considerablemente sus utilidades esperadas al aumentar n.

Otro detalle a subrayar en el estudio realizado con n=50 es el considerable incremento en utilidad logarítmica obtenido al pasar de RLC1 a RLC2. Esto es debido a que la información proporcionada por los cincuenta primeros datos es mas bien escasa, por ello se incrementan los incovenientes de un método estimativo, RLC1, frente a un método *quasi*-predictivo, RLC2.

■■

EJEMPLO 2. *Datos demográficos.*

En el banco 2 del apéndice 2 se recogen 50 datos correpondientes a un estudio demográfico realizado en USA. Press and Wilson (1978) utilizaron estos datos pa-

ra comparar los métodos LML y LDA.

El número de clases consideradas es dos, k=2, y existen cinco indicadores, tres de ellos contínuos y dos dicotómicos. Considerando un término independiente, el número de componentes del vector representante es seis, m=6.

Con el fin de evaluar los resultados obtenidos, el banco de datos se divide, de forma aleatoria, en dos partes iguales. Las distribuciones de clasificación se calculan con 25 datos y se evalúan con los restantes 25 datos. Para disminuir la influencia de la partición evaluada, los cálculos se repiten con cinco particiones distintas, todas ellas seleccionadas de forma aleatoria.

Los resultados obtenidos en las dos primeras evaluaciones se presentan en la tabla 2, mientras que la tabla 3 recoge la media de las cinco evaluaciones.

En el estudio realizado por Press and Wilson (1978), se evaluaban los métodos calculando el porcentaje de aciertos, u_1. Las conclusiones de su artículo eran favorables al método LML frente al LDA. Los resultados recogidos en la tabla 3 confirman que la utilidad u_1 proporcionada por LML es mayor que la obtenida por LDA; sin embargo, tanto en utilidad cuadrática, u_3, como en logarítmica, u_4, son mejores los resultados de LDA.

En porcentaje de aciertos, u_1, y en utilidad cuadrática, u_3, los mejores resultados obtenidos corresponden al método RLC2, quedando este en segundo lugar en utilidad logarítmica, u_4. Es de destacar los pobres resultados obtenidos por BNC.

		Repetición 1				Repetición 2			
		u_1^*	u_2^*	u_3^*	u_4^*	u_1^*	u_2^*	u_3^*	u_4^*
RLC1	Media	.6305	.6254	-.0226	-.0956	.7292	.7118	.2533	.1867
	Des.Tip.	.1288	.3152	1.1592	1.2037	.1556	.3024	1.1017	1.0815
RLC2	Media	.6305	.6199	.0224	-.0284	.7292	.7006	.2904	.2525
	Des.Tip.	.1288	.2899	1.0667	1.0489	.1556	.2743	.9775	.8805
LML	Media	.6305	.6256	-.0245	-.1052	.7292	.7123	.2536	.1905
	Des.Tip.	.1288	.3165	1.1668	1.2295	.1556	.3018	1.0952	1.0696
LDA	Media	.5920	.5897	-.1593	-.2280	.7292	.6867	.2400	.1887
	Des.Tip.	.1300	.3248	1.2238	1.2173	.1552	.2769	1.0195	.9653
BPC	Media	.5920	.5796	-.0840	-.0975	.6875	.6442	.2177	.1760
	Des.Tip.	.1300	.2860	1.0821	.9595	.1875	.2379	.8650	.7481
BNC	Media	.5948	.5805	-.2158	-.5883	.6458	.6084	-.0135	-.2115
	Des.Tip.	.1409	.3297	1.2820	1.6930	.1882	.2847	1.0851	1.3930

- Tabla 2 -

	u_1^*	u_2^*	u_3^*	u_4^*
RLC1	.6705	.6614	.0828	-.0634
RLC2	.6776	.6545	.1237	.0385
LML	.6705	.6609	.0676	-.1116
LDA	.6551	.6402	.1115	.0533
BPC	.6478	.6040	.0647	.0179
BNC	.5969	.5855	-.1212	-.3948

- Tabla 3 -

Como muestra de la necesidad de utilizar un banco auxiliar para la avaluación de resultados, en la tabla 4 se recogen las medias y desviaciones típicas obtenidas al utilizar el banco completo, los cincuenta datos, tanto para la obtención de los métodos como para su evaluación. Los métodos que obtuvieron peores resultados en el estudio anterior, tabla 3, son ahora los que se destacan, mostrando con ello una mayor capacidad de ajuste a unos datos concretos, pero un menor poder predictivo.

		u_1^*	u_2^*	u_3^*	u_4^*
RLC1	Media	.7885	.6990	.3887	.3355
	Des.Tip.	.0872	.2046	.7178	.6304
RLC2	Media	.7885	.6919	.3877	.3550
	Des.Tip.	.0872	.1954	.6824	.5884
LML	Media	.7885	.6990	.3886	.3354
	Des.Tip.	.0872	.2045	.7171	.6304
LDA	Media	.7885	.6512	.3370	.2814
	Des.Tip.	.0872	.1619	.5808	.4813
BPC	Media	.7885	.6404	.3311	.2712
	Des.Tip.	.0872	.1507	.5459	.4419
BNC	Media	.8269	.7135	.5004	.4268
	Des.Tip.	.1049	.1687	.5709	.4764

- Tabla 4 -

■■

EJEMPLO 3. *Datos de botánica.*

El banco 3 del apéndice 2 consiste en los 150 datos presentados por Fisher (1936), en el primer artículo aparecido sobre clasificación estadística.

El número de clases consideradas es tres, $k=3$, y existen cuatro indicadores, todos ellos continuos. Considerando un término independiente, el número de componentes del vector representante es cinco, $m=5$.

En su artículo, Fisher solo estudió las dos primeras categorías, Iris setosa e Iris versicolor, obteniendo que la combinación lineal

$$y_1 + 5.9037y_2 - 7.1299y_3 - 10.1036y_4$$

separaba completamente las dos clases. En efecto, esa combinación lineal da un resultado negativo para los cincuenta datos correspondientes a Iris setosa, mientras que es positiva en los cincuenta casos de Iris versicolor.

La separación completa del banco de datos va emparejada de la no existencia de estimador maximo verosímil, por tanto la función de verosimilitud no tiene forma acampanada y, en consecuencia, su medida de Lebesgue es infinita.

En tal caso, el uso de iniciales impropias puede no ser válido pues la distribución final puede ser a su vez, impropia. Este es el caso al considerar los 150 datos del banco 3.

Por el contrario, entre las categorías versicolor y virgínica no existe separación completa si se consideran los cien datos correspondientes a esas dos categorías.

Por estos motivos, el estudio aquí presentado se reduce a la clasificación entre Iris versicolor e Iris virgínica. Al igual que en el ejemplo anterior, se realizaron cinco repeticiones en las que los métodos fueron calculados con ochenta datos y evaluados con los veinte restantes. Sin embargo, solo en dos de esas repeticiones se obtuvieron datos que no presentasen separación completa, por ello las otras tres evaluaciones fueron rechazadas (la distribución final $\pi_E(\theta|D)$ era impropia). En la tabla 5 se presentan esas dos repeticiones, mientras que su media aparece en la tabla 6.

		Repetición 1				Repetición 2			
		u_1^*	u_2^*	u_3^*	u_4^*	u_1^*	u_2^*	u_3^*	u_4^*
RLC1	Media	.9542	.9511	.9536	.9497	.9530	.9355	.9281	.9215
	Des.Tip.	.0609	.2085	.2090	.2083	.0634	.2100	.2114	.2120
RLC2	Media	.9542	.9428	.9460	.9361	.9530	.9328	.9261	.9171
	Des.Tip.	.0609	.2076	.2080	.2077	.0634	.2096	.2117	.2119
LML	Media	.9542	.9512	.9536	.9497	.9530	.9337	.9222	.9171
	Des.Tip.	.0609	.2085	.2090	.2083	.0634	.2101	.2129	.2134
LDA	Media	.9542	.9451	.9525	.9408	.9530	.9247	.9195	.9036
	Des.Tip.	.0609	.2073	.2088	.2066	.0634	.2083	.2127	.2111
BPC	Media	.9542	.9422	.9522	.9366	.9530	.9225	.9279	.9031
	Des.Tip.	.0609	.2067	.2087	.2056	.0634	.2066	.2101	.2063
BNC	Media	.9542	.9480	.9538	.9452	.9530	.9215	.9281	.9021
	Des.Tip.	.0609	.2078	.2090	.2072	.0634	.2073	.2090	.2077

- Tabla 5 -

	u_1^*	u_2^*	u_3^*	u_4^*
RLC1	.9536	.9433	.9409	.9356
RLC2	.9536	.9378	.9361	.9266
LML	.9536	.9425	.9379	.9334
LDA	.9536	.9349	.9360	.9222
BPC	.9536	.9324	.9401	.9199
BNC	.9536	.9348	.9410	.9236

- Tabla 6 -

A pesar de los problemas que conlleva la separación completa del banco de datos, este ejemplo es muy interesante pues muestra el comportamiento de los métodos RLC1 y RLC2 en una situación que, en teoría, les es totalmente adversa. En efecto, el banco de datos 3 parece ajustarse perfectamente al enfoque muestral y la hipótesis de multinormalidad de los indicadores dada la clase parece correcta.

Los resultados obtenidos por el método RLC1 son sorprendentemente buenos, superando en utilidad logarítmica u_4, incluso al método BNC, el gran favorito a priori. En utilidad cuadrática, u_3, RLC1 queda en segundo lugar, siendo superado por BNC. Por el contrario, el método RLC2, aunque proporcionando también utilidades muy altas, se queda algo rezagado aunque superando siempre al método LDA.

■■

CAPITULO 6

DISCUSION Y AREAS DE INVESTIGACION FUTURA

El tema principal de esta memoria, clasificación estadística, constituye una de las áreas de estadística aplicada con mayor bibliografía, sin embargo, no existen apenas referencias en la literatura especializada sobre metodología bayesiana aplicada al enfoque clasificatorio. Esta memoria quiere ser un primer intento en el desarrollo de ese campo de investigación, totalmente novedoso y con un gran atractivo dentro de la estadística aplicada.

Partiendo de un concepto original expuesto por primera vez en esta tesis, el concepto de modelos de clasificación regulares, se desarrollan las bases y una incipiente estructura de lo que, aunque todavía en fase de experimentación, parece ser una herramienta importante en estadística aplicada y especialmente en diagnosis automática. Algunos de los resultados propuestos en el capítulo 4, en concreto la distribución $\pi_E(\theta|x,D)$, no pretenden ser *la* solución bayesiana al problema, sino que se han presentado solamente como primeros frutos de la estructura general defendida en esta memoria. Sin embargo, los ejemplos expuestos en los capítulos 3 y 5 muestran que esos primeros frutos son altamente prometedores.

Muchos son los problemas y cuestiones a las que todavía es necesario contestar. Unos son teóricos, importantes en el perfeccionamiento matemático de la estructura global; otros, los mas, se refieren a posibles áreas de investigación que complementen la teoría aquí presentada y amplíen el ámbito de aplicaciones.

Entre los primeros hay dos en concreto que requieren un interés especial.

(i) Un complemento al capítulo 2 lo constituiría el enunciado de una proposición que, recogiendo un sistema de condiciones suficientes, demostrase que la existencia del estimador maximo verosímil es suficiente para que las colas de la función de verosimilitud converjan a cero. La importancia de ese resultado radica en que, dado un conjunto de datos específico, es bastante mas sencillo demostrar si existe o no el estimador maximo verosímil que comprobar si las colas de la función de verosimilitud tienden a cero.

Un resultado todavía mas interesante consistiría en encontrar las condiciones bajo las cuales la existencia del estimador maximo verosímil es suficiente para que la función de verosimilitud posea medida finita. Ese resultado permitiría conocer bajo que condiciones puede utilizarse una inicial impropia con la seguridad de que la distribución final será propia.

Este punto está íntimamente relaccionado con un pro-

blema práctico de gran importancia: separación completa en el banco de datos. En el ejemplo 5.3.3. se comprobaba que, para algunas repeticiones, la distribución final no era propia debido a que el banco de datos presentaba separación completa y por tanto la función de verosimilitud no tenía máximo. Sin embargo, esas son las situaciones en las que la clasificación debería ser mas sencilla. El estudio de este problema puede aconsejar alguna modificación de la distribución inicial.

(ii) En el capítulo 4 se utiliza la regla de Jeffreys para la búsqueda de distribuciones de referencia. Su utilización se justifica a través de los resultados obtenidos por Bernardo (1979). Sin embargo, como apunta Ferrandiz (1981), los resultados de Bernardo necesitan la convergencia en media, en lugar de la convergencia en distribución obtenida en el teorema 3.1.1. Sería conveniente, por tanto, un estudio de las condiciones bajo las cuales se obtiene la convergencia en media.

Entre las lineas de investigación orientadas hacia un perfeccionamiento que incremente las posibilidades de aplicación en problemas concretos, cabe destacar:

(i) El estudio de la función de verosimilitud correspondiente a un muestreo retrospectivo. En efecto, como ya se comentó en el apartado 1.2, el enfoque clasificatorio se ajusta muy bien a datos prospectivos, sin embargo con datos retrospectivos parece necesario, desde el punto de vista teórico, la modelización de la distribución del vector represen-

tante. Habida cuenta de la supremacía, en aplicaciones, de datos retrospectivos frente a prospectivos, este punto requiere un interés especial.

(ii) El tratamiento de indicadores categóricos no ordenados . Si estos indicadores solo poseen dos categorias no aparece ningún problema, pues existe una cantidad aleatoria dicotómica, (por ejemplo 0,1), que los representa de forma natural. Ese no es el caso si el número de categorías es mayor de 2, entonces el orden artificial que se les asigna al representarlos mediante una cantidad aleatoria puede influír de forma radical en los resultados.

(iii) La asignación de distribuciones iniciales informativas. Este es un problema inherente a toda aplicación práctica de la metodología bayesiana. Por tanto, aunque verdaderamente importante en el contexto de clasificación, no es una linea de investigación específica del mismo. Cualquier resultado obtenido sobre asignación de distribuciones iniciales multivariantes en cualquier otra área de investigación estadística podría ser, sin excesivos problemas, utilizado en clasificación.

(iv) Por último, aunque no por ello lo menos prioritario hay que comentar la necesidad de rutinas de ordenador que posibiliten el uso de los modelos normal acumulado y logístico multiplicativo. Obteniendo, de esta forma, alternativas al modelo logístico aditivo y la posibilidad de contrastar los resultados obtenidos en cualquier aplicación práctica.

APENDICE 1

UNA APORTACION AL COMPORTAMIENTO ASINTOTICO DE LA DISTRIBUCION FINAL

En los estudios realizados sobre el comportamiento asintótico del estimador máximo-verosímil, o de los estimadores bayes, cuando se considera que el espacio paramétrico, Θ, puede no se acotado, se añade una condición de regularidad, (ver por ejemplo la condición 5 en Walker, 1969), prácticamente la misma en todos los artículos consultados por el autor, que no es satisfecha por los modelos de clasificación diagnóstica regulares. En este apéndice se modifica la axiomática de Walker proponiendo una condición alternativa, satisfecha por los modelos de clasificación diagnóstica regulares, estrictamente mas débil que su condición 5.

Se comprueba que la sustitución de la condición 5 por esa condición alternativa en la axiomática de Walker sigue proporcionando una axiomática suficiente para la aproximación asintótica normal a la distribución final.

A1.1 CONDICIONES DE REGULARIDAD

Dada una muestra aleatoria de tamaño n, $x^{(n)} \equiv (x_1,\ldots,x_n)$, tomada de una población cuya distribución se ha parametrizado a través del modelo probabilístico $p(x|\theta)$, el Teorema de Bayes asegura que

$$p(\theta|x^{(n)}) \propto p(x^{(n)}|\theta)\ p(\theta) = p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

Si la distribución inicial, $p(\theta)$, no se anula en ningún punto del espacio paramétrico Θ, y si la función de verosimilitud, $p(x^{(n)}|\theta)$, se va concentrando mas y mas en torno a su máximo conforme n crece, es sencillo dar una demostración heurística, utilizando series de Taylor, de que la distribución final, $p(\theta|x^{(n)})$, es aproximadamente normal con media el estimador máximo verosímil, $\hat{\theta}$, y con matriz de precisión la matriz de Información de Fisher. (Ver por ejemplo: Lindley, 1965, teoremas 7.1 y 7.2; Bernardo, 1981, teoremas 6.4.2 y 6.4.3).

Sin embargo, una demostración rigorosa de este resultado exige la introducción de un conjunto de condiciones de regularidad que garanticen tanto la consistencia del estimador máximo verosímil, lo que conlleva la concentración de la función de verosimilitud en torno a su máximo, como la normalidad asintótica de la distribución en el muestreo del mismo.

Las condiciones de regularidad mas conocidas son, sin lugar a dudas, las propuestas por Walker(1969):

CR.1

$\Theta \in R^s$, es un conjunto cerrado

CR.2

El conjunto de puntos $\chi \equiv \{x; p(x|\theta) > 0\}$ es independiente de θ.

CR.3

Si θ_1 y θ_2 son dos puntos distintos de Θ, el conjunto de puntos $x \in \chi$ tales que $p(x|\theta_1) \neq p(x|\theta_2)$ tiene medida no nula.

CR.4

Sea $x \in \chi$ y $\theta^* \in \Theta$. Dado $\varepsilon > 0$, suficientemente pequeño, $\forall \theta$ tal que $\|\theta - \theta^*\| < \varepsilon$, $|\text{Log } p(x|\theta) - \text{Log } p(x|\theta^*)|$ está acotado por una función $G_\varepsilon(x, \theta^*)$ tal que,

$$\lim_{\varepsilon \to 0} G_\varepsilon(x, \theta^*) = 0$$

y $\forall \theta_0 \in \Theta$, $\lim_{\varepsilon \to 0} \int G_\varepsilon(x, \theta^*)\, p(x|\theta_0)\, d\mu = 0$.

CR.5

Si el conjunto Θ no está acotado entonces, dado $\theta_0 \in \Theta$ y $\Delta \in R$, constante positiva suficientemente grande:

$$\forall \theta,\ \|\theta\| > \Delta,\ \text{Log } p(x|\theta) - \text{Log } p(x|\theta_0) < K_\Delta(x, \theta_0)$$

con $\lim_{\Delta \to \infty} \int K_\Delta(x, \theta_0)\, p(x|\theta_0)\, d\mu < 0$.

Este límite puede no ser finito.

En las siguientes condiciones θ_0 representa un punto interior del espacio paramétrico Θ.

CR.6

Log $p(x|\theta)$ es dos veces diferenciable con respecto a θ, en algún entorno de θ_0.

CR.7

Sea $J(\theta_0)$ la matriz $s\times s$ con elemento típico

$$J_{ij}(\theta_0) = \int \frac{\partial \text{ Log } p_0}{\partial \theta_{0,i}} \frac{\partial \text{ Log } p_0}{\partial \theta_{0,j}} p_0 \, d\mu$$

donde $p_0=p(x|\theta_0)$. Entonces $|J_{ij}(\theta_0)|<+\infty \ \forall i,j$ y la matriz $J(\theta_0)$ es definida positiva.

CR.8

$$\int \frac{\partial p_0}{\partial \theta_{0,i}} d\mu = \int \frac{\partial^2 p_0}{\partial \theta_{0,i} \partial \theta_{0,j}} d\mu = 0, \quad \forall i,j=1,\ldots,s$$

CR.9

Si $\|\theta-\theta_0\|<\varepsilon$, siendo ε una constante positiva suficientemente pequeña, entonces:

$$\left| \frac{\partial^2 \text{ Log } p(x|\theta)}{\partial \theta_i \, \partial \theta_j} - \frac{\partial^2 \text{ Log } p(x|\theta_0)}{\partial \theta_{0i} \partial \theta_{0j}} \right| < m_{ij}$$

donde m_{ij} es el elemento típico de la matriz $M_\varepsilon(x,\theta_0)$. Además,

$$\lim_{\varepsilon\to 0} \int M_\varepsilon(x,\theta_0) \, p(x|\theta_0) \, d\mu = 0$$

CR.10

La distribución inicial sobre θ, $\pi(\theta)$, es continua en $\theta=\theta_0$ y $\pi(\theta_0)>0$.

Con las condiciones CR.1 a CR.5 se puede demostrar la consistencia del estimador máximo verosímil. Sin embargo, la conclusión que se pretende obtener con esas condiciones en la demostración dada por Walker(1969) es:

> Sea $N_0(\varepsilon) \equiv \{\theta \in \Theta : \|\theta - \theta_0\| < \varepsilon\}$ un entorno de θ_0. Existe un numero positivo $k(\varepsilon) > 0$, dependiente de ε, tal que,

$$\lim_{n \to \infty} P\left[\sup_{\theta \in \Theta - N_0(\varepsilon)} n^{-1}\{p(x^{(n)}|\theta) - p(x^{(n)}|\theta_0)\} < -k(\varepsilon)\right] = 1 \qquad (1)$$

Las condiciones CR.6 a CR.9 son, según Walker, "...las convenientes para asegurar que la distribución asintótica en el muestreo del estimador máximo verosímil sea Normal". Por último, CR.10 conlleva la no anulación de la distribución inicial en un entorno alrededor del verdadero valor del parámetro.

A1.2 ESTUDIO DE LA CONDICION CR.5

La condición CR.5 permite trabajar con un conjunto Θ no acotado, por ello no aparece en ninguno de los trabajos en los que, explícita o implícitamente, se considera que $\Theta \subset R^s$ es un conjunto acotado. (Ver por ejemplo: Cramer, 1946; Huzurbazar, 1948; Chanda, 1954; Bradley and Gart, 1962).

Por el contrario, en todos los trabajos consultados por el autor sobre el comportamiento asintótico

del estimador máximo verosímil, cuando se considera un espacio paramétrico, Θ, posiblemente no acotado, se utiliza la condición CR.5 o alguna otra equivalente. (Ver por ejemplo: Wald, 1949, hipótesis 5; Kiefer and Wolfowitz, 1956, hipótesis 2; Chao, 1970, hipótesis A9; Johnson, 1970, hipótesis 6; Dawid, 1970, hipótesis C7).

Sin embargo, los modelos de clasificación regulares no satisfacen dicha condición pues,

Dado $\theta_0 \in \Theta$, para todo $\Delta>0$, si existe $K_\Delta(x,\theta_0)$ tal que $\text{Log } p(x|\theta)-\text{Log } p(x|\theta_0)<K_\Delta(x,\theta_0)$ para todo $\theta\in\Theta$ con $\|\theta\|>\Delta$, entonces,

$$K_\Delta(x,\theta_0) \geq \sup_{\theta:\|\theta\|>\Delta} \left\{\text{Log } p(x|\theta)-\text{Log } p(x|\theta_0)\right\} =$$

$$= - \text{Log } p(x|\theta_0) \geq 0$$

ya que, por la propiedad P3, el supremo de la función de verosimilitud de un dato es uno, y se alcanza de forma asintótica, esto es, mediante una sucesión $\{\theta_i; i=1,...\}$ para la que $\|\theta_i\|\to+\infty$ cuando $i\uparrow\infty$.

Por tanto la esperanza de $K_\Delta(x,\theta_0)$ es no negativa para todo Δ, lo que contradice la condición C1.5.

Este hecho hace necesario un estudio profundo de las implicaciones de CR.5 necesarias en la demostración de la normalidad asintótica, que permita encontrar condiciones alternativas mas débiles.

Si se considera

$$K_\Delta(x,\theta_0) = \sup_{\theta:\,||\theta||>\Delta} \text{Log } p(x|\theta) - \text{Log } p(x|\theta_0)$$

entonces la condición $\lim_{\Delta\to\infty} \int K_\Delta(x,\theta_0)\ p(x|\theta_0)\ d\mu < 0$ es equivalente a

$$\lim_{\Delta\to\infty} \int \left(\sup_{\theta:\,||\theta||>\Delta} \text{Log } p(x|\theta) - \text{Log } p(x|\theta_0) \right) p(x|\theta_0)\ d\mu < 0.$$

Por tanto, existe Δ_0 tal que para todo $\Delta>\Delta_0$:

$$\int \left(\sup_{\theta:\,||\theta||>\Delta} \text{Log } p(x|\theta) - \text{Log } p(x|\theta_0) \right) p(x|\theta_0)\ d\mu < 0 \qquad (2)$$

Para que una función medible tenga medida negativa, posiblemente $-\infty$, debe ocurrir que su parte positiva sea finita. Es mas, si $\mu(f)<+\infty$ y $f=g+h$ entonces $\mu(g)<+\infty$ y $\mu(h)<+\infty$. Aplicando este resultado elemental de teoría de la medida a la desigualdad (2), debe ocurrir que:

$$\int \left(\sup_{\theta:\,||\theta||>\Delta} \text{Log } p(x|\theta) \right) p(x|\theta_0)\ d\mu < +\infty \ \forall\ \Delta>\Delta_0$$

y

$$-\int \left(\text{Log } p(x|\theta_0) \right) p(x|\theta_0)\ d\mu = H\{p(x|\theta_0)\} < +\infty$$

siendo $H\{p(x|\theta_0)\}$ la entropía de $p(x|\theta_0)$.

Por otra parte (2) puede ser escrita como:

$$\int \left(\text{Log } \frac{\sup p(x|\theta)}{p(x|\theta_0)} \right) p(x|\theta_0)\ d\mu < 0$$

lo que en cierto modo, conlleva un comportamiento decreciente de las colas de la función de verosimilitud.

Algunos autores exigen explícitamente ese comportamiento *acampanado* de la verosimilitud; así por ejemplo, Wald (1949) y Johnson (1970) suponen que si $\{\theta_i\}$ es una sucesión con Lim $\|\theta_i\|=+\infty$, entonces Lim $p(x|\theta_i)=0$.

Sin embargo, en la demostración propuesta por Walker (1969) solo se utiliza CR.5 para comprobar la condición (1). Mas concretamente, para demostrar que:

Existe $\Delta>0$ tal que

$$\lim_{n\to\infty} P\left[\sup_{\theta\in S(\Delta)} n^{-1}\{\text{Log } p(x^{(n)}|\theta)-\text{Log } p(x^{(n)}|\theta_0)\}<-C_\Delta<0\right]=1 \qquad (3)$$

siendo $S(\Delta)\equiv\{\theta\in\Theta:\|\theta\|>\Delta\}$.

A1.3 UNA CONDICION ALTERNATIVA A CR.5

Considérese la condición alternativa CA desglosada de la siguiente forma:

CA.1

$\forall\,\theta_0$, punto interior de Θ, $\exists\,\Delta_1>0$ tal que $\forall\,x\in\chi$,

$\sup_{\|\theta\|^2>\Delta_1}$ Log $p(x|\theta) \le B_1(x)$, siendo $B_1(x)$ tal que

$$\int B_1(x)\, p(x|\theta_0)\, d\mu = B_1(\theta_0) < +\infty$$

CA.2

$\forall\,\theta_0$ punto interior de Θ,

$$H\{p(x|\theta_0)\}=-\int p(x|\theta_0)\ \text{Log}\ p(x|\theta_0)\ d\mu=B_2(\theta_0)<+\infty$$

CA.3

Sea $C_n \equiv \{(x_1,\ldots,x_n)\in\chi^n : \lim_{\Delta\to\infty} \sup_{\|\theta\|^2>\Delta} p(x_1,\ldots,x_n|\theta)=0\}$

entonces $\lim_{n\to\infty} P(C_n)=1$

Las condiciones CA.1 y CA.2 podrían resumirse en la condición equivalente

$\forall\theta_0$ punto interior de Θ, y dado $\Delta>0$ suficientemente grande, existe $M_\Delta(x,\theta_0)$ tal que,

$\forall\theta\in\Theta$ con $\|\theta\|>\Delta$, Log $p(x|\theta)$-Log $p(x|\theta_0)<M_\Delta(x,\theta_0)$

con $\lim_{\Delta\to\infty} \int M_\Delta(x,\theta_0)\ p(x|\theta_0)\ d\mu \le M(\theta_0) < +\infty$

condición muy similar a CR.5 pero mas débil, pues se admite que el límite pueda tomar cualquier valor distinto de $+\infty$. Sin embargo se ha preferido desglosar esta condición con el fin de mostrar las implicaciones que conlleva. Así CA.1 exige que, en el complementario de la bola de centro el origen y radio Δ_1, el logaritmo de la función de verosimilitud esté acotado por una función con medida finita. CA.2 exige que, para todo punto interior $\theta_0\in\Theta$, la entropía de la función $p(x|\theta_0)$ esté acotada por una constante finita.

Por último CA.3 exige que, con probabilidad tendiendo a uno conforme el número de datos crece, las colas de la función de verosimilitud se acercan a cero, alcanzando ese valor en el límite.

Obviamente las condiciones CA.1 y CA.2 son necesarias para que se cumpla la condición CR.5. El siguiente

resultado demuestra que lo mismo es cierto para CA.3.

PROPOSICION 1

La condición CA.3 es necesaria para que se cumpla CR.5.

DEMOSTRACION

CR.5 implica que, dado θ_0 punto interior de Θ, existe Δ^* tal que:para todo $\Delta \geq \Delta^*$ existe $c_\Delta > 0$,

$$\int K_\Delta(x,\theta_0)\ p(x|\theta_0)\ d\mu \leq -2c_\Delta < 0.$$

Además, como $\sup_{||\theta|| > \Delta}$ Log $p(x|\theta)$-Log $p(x|\theta_0) \leq K_\Delta(x,\theta_0)$, entonces

$$\int \left\{ \sup_{||\theta|| > \Delta} \text{Log } p(x|\theta)-\text{Log } p(x|\theta_0) \right\} p(x|\theta_0)\ d\mu \leq -2c_\Delta$$

Aplicando la ley débil de los grandes números a la cantidad aleatoria $\sup_{||\theta|| > \Delta}$ Log $p(x|\theta)$-Log $p(x|\theta_0)$, cuya esperanza es menor o igual que $-2c_\Delta < 0$:

$$\lim_{n\to\infty} P\left\{ \sup_{||\theta|| > \Delta} n^{-1} \sum_{i=1}^{n} \{\text{Log } p(x_i|\theta)-\text{Log } p(x_i|\theta_0)\} < -c_\Delta < 0 \right\} = 1$$

que es precisamente la condición (3).

La cantidad aleatoria

$$n^{-1}\left\{ \sup_{||\theta|| > \Delta} \text{Log } p(x^{(n)}|\theta)-\text{Log } p(x^{(n)}|\theta_0) \right\}$$

converge estocásticamente a un número negativo, por tanto, la cantidad aleatoria

$$\text{Log} \quad \frac{\sup\, p(x^{(n)}|\theta)}{p(x^{(n)}|\theta_0)}$$

converge estocásticamente a $-\infty$, lo que implica que

$$\left\{\sup\, p(x^{(n)}|\theta)\right\}/p(x^{(n)}|\theta_0)$$

converge estocásticamente a cero. Por último, eso implica la convergencia estocástica a cero de la cantidad aleatoria

$$\sup_{\|\theta\|>\Delta} \quad p(x^{(n)}|\theta), \quad \forall\ \Delta\geq\Delta^*$$

por tanto

$$\lim_{n\to\infty} P\left\{\lim_{\Delta\to\infty} \sup_{\|\theta\|>\Delta} p(x^{(n)}|\theta)=0\right\} = 1 \quad \blacksquare$$

Una consecuencia inmediata de este resultado es que la condición CA es mas débil que CR.5. Es mas, CA es estrictamente mas débil que CR.5 puesto que, por ejemplo, los modelos de clasificación regulares cumplen CA pero no cumplen CR.5.

CA puede ser utilizada en lugar de CR.5 en la demostración de la normalidad asintótica de la distribución final, como prueba el siguiente resultado.

TEOREMA 1

Si se cumple CR.2 entonces CA es suficiente para que se cumpla (3),i.e., CA es suficiente para que exista $\Delta>0$ tal que si $S(\Delta)\equiv\{\theta\in\Theta: \|\theta\|>\Delta\}$ entonces

$$\lim_{n\to\infty} P\left\{\sup_{\|\theta\|>\Delta} n^{-1}\{Log\ p(x^{(n)}|\theta)-Log\ p(x^{(n)}|\theta_0)\}<-c_\Delta<0\right\} = 1$$

Si se define $Q_{\Delta}(x^{(n)},\theta_0)$ como

$$Q_{\Delta}(x^{(n)},\theta_0) = n^{-1}\left[\sup_{\|\theta\|>\Delta} \text{Log } p(x^{(n)}|\theta) - \text{Log } p(x^{(n)}|\theta_0)\right]$$

entonces la expresión (3) especifica que

$$\exists\, \Delta^*>0 \text{ tal que } \lim_{n\to\infty} P\left\{Q_{\Delta^*}(x^{(n)},\theta_0) < -c_{\Delta^*} < 0\right\} = 1$$

La demostración del teorema 1 aquí presentada tiene dos partes claramente diferenciadas. En la primera se demuestra la existencia de tres constantes finitas $\Delta^*>0$, $c_{\Delta^*}>0$ y $m_0>0$, tales que

$$E\left(Q_{\Delta^*}(x^{(m_0)},\theta_0)\right) \leq -4c_{\Delta^*} < 0.$$

En la segunda parte se utiliza la ley débil de los grandes números aplicada a las cantidades aleatorias $\{Q_{\Delta^*}(x^{(n)},\theta_0);\ n=1,\ldots\}$ para demostrar que

$$\lim_{n\to\infty} P\left\{Q_{\Delta^*}(x^{(n)},\theta_0) < -c_{\Delta^*} < 0\right\} = 1.$$

DEMOSTRACION DEL TEOREMA 1

PARTE A

Sea $B_3(x,\theta_0) = \text{Max}\left\{B_1(x) - \text{Log } p(x|\theta_0),\ 0\right\}$, donde $B_1(x)$ está definido en CA.1. Sea

$$B_3(\theta_0) = \int B_3(x,\theta_0)\ p(x|\theta_0)\ d\mu \geq$$

$$\geq \int\left\{B_1(x) - \text{Log } p(x|\theta_0)\right\}\ p(x|\theta_0)\ d\mu =$$

$$= \int B_1(x)\ p(x|\theta_0)\ d\mu - \int \text{Log}\ p(x|\theta_0)\ p(x|\theta_0)\ d\mu =$$

$$= B_1(\theta_0) + B_2(\theta_0)$$

Con esta definición $B_3(x,\theta_0)$ es la parte positiva de la función $B_1(x)$-Log $p(x|\theta_0)$ que tiene medida finita

$$\mu\{B_1(x)-\text{Log}\ p(x|\theta_0)\}=B_1(\theta_0)+B_2(\theta_0)<+\infty$$

por tanto $0\leq\mu\{B_3(x,\theta_0)\}=B_3(\theta_0)<+\infty$.

En consecuencia, $\forall\ \Delta\geq\Delta_1$ (Δ_1 definido en CA.1), $\forall\ n>0$ y $\forall\ C\subset\chi^n$:

$$\int_C Q_\Delta(x^{(n)},\theta_0)\ p(x^{(n)}|\theta_0)\ d\mu =$$

$$=\int_C n^{-1}\left\{\sup_{\|\theta\|>\Delta} \text{Log}\ p(x^{(n)}|\theta)-\text{Log}\ p(x^{(n)}|\theta_0)\right\}p(x^{(n)}|\theta_0)d\mu\leq$$

$$\leq\int_C n^{-1}\sum_{i=1}^{n}\left\{\sup_{\|\theta\|>\Delta} \text{Log}\ p(x_i|\theta)-\text{Log}\ p(x_i|\theta_0)\right\}p(x^{(n)}|\theta_0)d\mu\leq$$

$$\leq\int_C n^{-1}\sum_{i=1}^{n}\left\{B_1(x_i)-\text{Log}\ p(x_i|\theta_0)\right\}p(x^{(n)}|\theta_0)\ d\mu\leq$$

$$\leq\int_C n^{-1}\sum_{i=1}^{n} B_3(x_i,\theta_0)\ p(x^{(n)}|\theta_0)\ d\mu =$$

$$=n^{-1}\sum_{i=1}^{n}\int_C B_3(x_i,\theta_0)\ p(x^{(n)}|\theta_0)\ d\mu\leq$$

$$\leq n^{-1}\sum_{i=1}^{n}\int B_3(x_i,\theta_0)\ p(x^{(n)}|\theta_0)\ d\mu =$$

$$=n^{-1}\sum_{i=1}^{n}\int B_3(x_i,\theta_0)\ p(x_i|\theta_0)\ d\mu\leq B_3(\theta_0)<+\infty \qquad (4)$$

donde el rango de integración solo se ha especificado si este era solo parte del espacio total.

Por otra parte, sean $\{C_n;\ n=1,\ldots\}$ los conjuntos definidos en CA.3. Como Lim $P(C_n)=1$, dado $\varepsilon>0$ existe $m(\varepsilon)$ talque para todo $n\geq m(\varepsilon)$ $p(C_n)>1-\varepsilon$. Además, por definición de C_n,

$$\forall\ x^{(n)}\in C_n,\ \lim_{\Delta\to\infty}\ \sup_{\|\theta\|>\Delta}\ p(x^{(n)}|\theta) = 0$$

siendo $p(x^{(n)}|\theta_0) = \prod_{i=1}^{n} p(x_i|\theta_0) > 0$ ya que $x_i\in\chi\equiv\{x;$ $p(x|\theta)>0\}$, y por CR.2 χ es independiente de θ. Por tanto:

$$\forall\ x^{(n)}\in C_n,\ \lim_{\Delta\to\infty}\ \frac{\sup p(x^{(n)}|\theta)}{p(x^{(n)}|\theta_0)} = 0 \Rightarrow$$

$$\Rightarrow \lim_{\Delta\to\infty} n^{-1}\left\{\sup \text{Log}\ p(x^{(n)}|\theta)-\text{Log}\ p(x^{(n)}|\theta_0)\right\} =$$

$$= \lim_{\Delta\to\infty} Q_\Delta(x^{(n)},\theta_0) = -\infty$$

puesto que Log es una función estrictamente creciente y n es constante. Por tanto $\exists\Delta_2(n)$ tal que $\forall\Delta\geq\Delta_2(n)$ y $\forall x^{(n)}\in C_n$:

$$Q_\Delta(x^{(n)},\theta_0) < -(B_3(\theta_0)+1) \qquad (5)$$

Sea $0<\varepsilon_0<(B_3(\theta_0)+1)^{-1}$, $m_0=m(\varepsilon)$ y $\Delta^*=\text{Max}(\Delta_1,\Delta_2(m_0))$, entonces:

$$E\left\{Q_{\Delta^*}(x^{(m_0)},\theta_0)\right\} = \int Q_{\Delta^*}(x^{(m_0)},\theta_0)\ p(x^{(m_0)}|\theta_0)\ d\mu =$$

$$= \int_{C_{m_0}} Q_{\Delta^*}(x^{(m_0)},\theta_0)\ p(x^{(m_0)}|\theta_0)d\mu + \int_{\bar{C}_{m_0}} Q_{\Delta^*}(x^{(m_0)},\theta_0)p(x^{(m_0)}| \qquad (6)$$

donde $\overline{C}_{m_0}$ es el complementario de C_{m_0} con respecto a χ^m .

Por (5), la primera integral en (6) se reduce a

$$\int_{C_{m_0}} Q_{\Delta^*}(x^{(m_0)},\theta_0)\ p(x^{(m_0)}|\theta_0)\ d\mu <$$

$$< \int_{C_{m_0}} -(B_3(\theta_0)+1)\ p(x^{(m_0)}|\theta_0)\ d\mu =$$

$$= -(B_3(\theta_0)+1)\ P(C_{m_0}) < -(B_3(\theta_0)+1)(1-\varepsilon_0) =$$

$$= \varepsilon_0(B_3(\theta_0)+1)-B_3(\theta_0)-1 < 1-B_3(\theta_0)-1 = -B_3(\theta_0)$$

por construcción de C_{m_0} y ε_0.

La segunda integral en (6) puede ser acotada utilizando (4):

$$\int_{\overline{C}_{m_0}} Q_{\Delta^*}(x^{(m_0)},\theta_0)\ p(x^{(m_0)}|\theta_0)\ d\mu \leq B_3(\theta_0)$$

por tanto,

$$E\left\{Q_{\Delta^*}(x^{(m_0)},\theta_0)\right\} < -B_3(\theta_0)+B_3(\theta_0) = 0$$

como la desigualdad es estricta, $\exists\ c_{\Delta^*}>0$ tal que

$$E\left\{Q_{\Delta^*}(x^{(m_0)},\theta_0)\right\} < -4c_{\Delta^*} < 0.$$

PARTE B

Para todo $n\geq m_0$, existen $r,s\in\mathbb{N}$ tales que $n=rm_0+s$, con $r\geq 1$, y $s<m_0$. Por tanto:

$$Q_{\Delta^*}(x_1,\ldots,x_n,\theta_0) =$$

$$= n^{-1}\left[\sum_{i=1}^{r-1} m_0 Q_{\Delta^*}(x_{im_0+1},\ldots,x_{im_0+m_0},\theta_0) + \sum_{j=1}^{s} Q_{\Delta^*}(x_{rm_0+j},\theta_0)\right]$$

Sea $z_{i+1}=Q_{\Delta^*}(x_{im_0+1},\ldots,x_{im_0+m_0},\theta_0)$, con lo que las cantidades aleatorias $\{z_i;\ i=1,\ldots,r\}$ serán independientes e idénticamente distribuídas, con

$$E(z_i) = E\left[Q_{\Delta^*}(x^{(m_0)},\theta_0)\right] < -4c_{\Delta^*} < 0.$$

Si esa esperanza no fuese finita, siempre se podrían definir las cantidades aleatorias $\{z_i\}$ como el supremo entre las anteriores z_i y una cierta constante de forma que las nuevas z_i tubieran esperanza finita y menor que $-4c_{\Delta^*}$. Por tanto se puede aplicar la ley débil de los grandes números a las cantidades aleatorias $\{z_i\}$. Es mas,

$$Q_{\Delta^*}(x^{(n)},\theta_0) \le m_0/n \sum_{i=1}^{r} z_i + n^{-1}\sum_{j=1}^{s} Q_{\Delta^*}(x_{rm_0+j},\theta_0) \le$$

$$\le m_0/n \sum_{i=1}^{r} z_i + n^{-1}\sum_{j=1}^{s} B_3(x_{rm_0+j},\theta_0) =$$

$$= (1-s/n)/r \sum_{i=1}^{r} z_i + n^{-1}\sum_{j=1}^{s} B_3(x_{rm_0+j},\theta_0)$$

y como las cantidades aleatorias que configuran la expresión anterior son independientes:

$$P\left[Q_{\Delta^*}(x^{(n)},\theta_0)<-c_{\Delta^*}\right] \ge$$

$$\ge P\left[(1-s/n)/r \sum_{i=1}^{r} z_i + 1/n \sum_{j=1}^{s} B_3(x_{rm_0+j},\theta_0) < -c_{\Delta^*}\right] \ge$$

$$\geq P\left\{(1-s/n)/r\sum_{i=1}^{r} z_i<-2c_{\Delta^*}\right\}\ P\left\{1/n\sum_{j=1}^{s} B_3(x_{rm_0+j},\theta_0)<c_{\Delta^*}\right\}\geq$$

$$\geq P\left\{(1-s/n)/r\sum_{i=1}^{r} z_i<-2c_{\Delta^*}\right\}\left\{P\left[n^{-1}B_3(x_{rm_0+1},\theta_0)<c_{\Delta^*}/s\right]\right\}^{s}\geq$$

$$\geq P\left\{(1-s/n)/r\sum_{i=1}^{r} z_i<-2c_{\Delta^*}\right\}\left\{P\left[n^{-1}B_3(x,\theta_0)<c_{\Delta^*}/m_0\right]\right\}^{m_0} \qquad (7)$$

Dado $\varepsilon>0$ sea $\varepsilon_1>0$ tal que $1-\varepsilon=(1-\varepsilon_1)^{m_0+1}$, esto es $\varepsilon_1=1-(1-\varepsilon)^{1/(m_0+1)}$.

$\forall\, n\geq 3m_0$, $1-s/n>1-m_0/n>1-1/3=2/3$, luego ,

$$-2(1-s/n)^{-1}c_{\Delta^*} > -3c_{\Delta^*}$$

por tanto:

$$P\left\{(1-s/n)/r\sum_{i=1}^{r} z_i<-2c_{\Delta^*}\right\}=P\left\{1/r\sum_{i=1}^{r} z_i<-2(1-s/n)^{-1}c_{\Delta^*}\right\}\geq$$

$$\geq P\left\{1/r\sum_{i=1}^{r} z_i< -3c_{\Delta^*}\right\}$$

y como $-3c_{\Delta^*}>-4c_{\Delta^*}>E(z_i)$, por la ley débil de los grandes números, dado $\varepsilon_1>0$, existe r_1 tal que para todo $r\geq r_1$,

$$P\left\{1/r\sum_{i=1}^{r} z_i<-3c_{\Delta^*}\right\} > 1-\varepsilon_1 .$$

Esto es, dado $\varepsilon_1>0$ existe $N_1=\text{Max}(r_1m_0,3m_0)$ tal que para todo $n=rm_0+s\geq N_1$,

$$P\left\{(1-s/n)/r\sum_{i=1}^{r} z_i<-2c_{\Delta^*}\right\} > 1-\varepsilon_1 \qquad (8)$$

Por otra parte,

$$P\left[n^{-1}B_3(x,\theta_0)<c_{\Delta^*}/m_0\right] = P\left[B_3(x,\theta_0)<nC\right] = P\left[S_n\right]$$

donde $C=c_{\Delta^*}/m_0$ es una constante estrictamente positiva, y donde $S_n \equiv \{x\epsilon\chi;\ B_3(x,\theta_0)<nC\}$.

La sucesión $\{S_n\}$ es una sucesión expansiva de conjuntos encajados y por tanto

$$\lim_n S_n = \bigcup_{n=1}^{\infty} S_n = S \equiv \{x\epsilon\chi;\ B_3(x,\theta_0)<+\infty\}$$

en consecuencia, (Bartle,1966, lema 3.4a),

$$\lim_n P(S_n) = P(\lim S_n) = P(\bigcup_{n=1}^{\infty} S_n) = P(S)$$

además, como $B_3(x,\theta_0)$ es una función positiva con medida finita, $\mu(B_3(x,\theta_0))=B_3(\theta_0)<+\infty$, (Bartle, 1966, ejercicio 4.R), $P(x\epsilon\chi;\ B_3(x,\theta_0)=+\infty)=0$, luego $P(S)=P(\chi)=1$.

$$\lim_n P(S_n) = P(S) = 1 \Rightarrow \text{Dado } \varepsilon_1 \quad \exists N_2;\ \forall n\geq N_2\ P(S_n)>1-\varepsilon_1$$

luego:

Dado $\varepsilon_1>0$ $\exists N_2$ tal que $\forall n\geq N_2$, $P(n^{-1}B_3(x,\theta_0)<c_{\Delta^*}/m_0)>1-\varepsilon_1$. Lo que, conjuntamente con (7) y (8) y la definición de ε_1, demuestra el teorema.

∎

TEOREMA 2

Si se cumplen las condiciones de regularidad CR.1 a CR.4 y CR.6 a CR.10 y la condición alternativa CA, entonces la distribución final $\pi(\theta|x^{(n)})$ *converge en distribución a una Normal con media el estimador máximo verosímil,* $\hat{\theta}$, *y matriz de precisión la matriz*

$$D^2_\theta(-Log\ p(x^{(n)}|\theta))$$

calculada en el punto $\hat{\theta}$.

La demostración de este teorema es totalmente similar a la dada por Walker(1969). La única diferencia estriba en que Walker obtiene la expresión (3) como consecuencia inmediata de su condición CR.5, mientras que aquí (3) se obtiene mediante el teorema 1.

APENDICE 2

BANCO DE DATOS 1

Tamaño del banco: 100
Número de clases: 2
Número de indicadores: 7 (Todos ellos dicotómicos)

BANCO DE DATOS 2

Tamaño del banco: 50
Número de clases: 2
Número de indicadores: 5 (2, dicotómicos;
3, continuos).

Estos datos están tomados de Press and Wilson(1978). Con ellos se estudia el cambio de población ocurrido en los cincuenta estados de USA entre los años 1960-1970.

Las clases consideradas son:

$\delta=1$, cambio de población por encima de la mediana de los cambios de todos los estados.

$\delta=0$, cambio de población por debajo de la mediana de los cambios de todos los estados.

Indicadores considerados:

y_1, renta per cápita (en miles de dólares).

y_2, natalidad (en tantos por cien).

y_3, presencia o ausencia de litoral en el estado (1, presencia; 0, ausencia).
y_4, urbanización (0, si menor del 70%; 1, en otro caso).
y_5, mortalidad (en tantos por cien).

BANCO DE DATOS 3

Tamaño del banco: 150
Número de clases: 3
Número de indicadores: 4 (Todos continuos).

Estos datos están tomados de Fisher (1936). Con ellos se estudia las diferencias entre las flores de los tres tipos de Iris, planta herbácea perenne de la familia de las irídeas.

Las clases consideradas son:
$\delta=1$, Iris setosa.
$\delta=2$, Iris versicolor.
$\delta=3$, Iris virgínica.

Indicadores considerados:
y_1, longitud de los sépalos.
y_2, anchura de los sépalos.
y_3, longitud de los pétalos.
y_4, anchura de los pétalos.

BANCO DE DATOS 4

Tamaño del banco: 500
Número de clases: 2
Número de indicadores: 1 (continuo)

Estos son los datos simulados utilizados en el apartado 3.3, estudio primero. Se han agrupado de cincuenta en cincuenta de igual forma como son utilizados en el apartado 3.3.

BANCO DE DATOS 5

Tamaño del banco: 500
Número de clases: 2
Número de indicadores: 1 (continuo)

Estos son los datos simulados utilizados en el apartado 3.3, estudio segundo. Se han agrupado de cincuenta en cincuenta de igual forma como son utilizados en el apartado 3.3.

BANCO DE DATOS 1

Numero	Indicadores							Clase
1	1	1	1	0	0	1	0	1
2	0	1	0	1	1	0	0	2
3	1	1	0	1	1	0	1	2
4	1	1	0	0	1	1	0	2
5	1	0	0	1	0	1	0	2
6	0	1	0	1	0	0	1	2
7	0	1	1	0	0	1	0	1
8	1	1	1	0	1	1	0	1
9	0	1	1	0	0	0	1	1
10	0	0	1	0	0	0	0	1
11	0	1	0	0	0	1	0	1
12	1	1	0	0	0	1	0	1
13	1	0	1	1	1	0	1	1
14	1	0	1	0	1	1	0	1
15	1	1	0	1	1	1	0	2
16	1	1	1	1	1	1	0	2
17	1	1	0	0	0	1	1	1
18	0	0	0	1	0	1	0	2
19	1	1	0	0	1	1	0	1
20	0	0	1	1	1	0	0	2
21	0	1	0	1	1	0	0	2
22	0	0	0	0	1	1	0	2
23	0	1	0	1	0	1	1	2
24	0	0	1	0	0	1	0	1
25	0	0	0	1	0	1	1	2
26	1	0	0	0	1	0	0	1
27	1	1	1	0	0	1	0	1
28	1	0	0	1	1	1	0	2
29	1	0	0	0	0	1	0	1
30	0	0	0	1	1	0	1	2
31	1	0	1	1	1	0	0	2
32	0	0	0	1	1	1	1	2
33	0	0	1	0	0	0	1	1
34	0	1	1	0	0	0	1	1
35	0	0	1	1	0	1	0	2
36	1	1	0	0	1	0	1	1
37	1	0	0	0	1	1	1	1
38	0	0	1	1	0	0	0	2
39	0	1	1	0	0	1	0	1
40	0	0	1	1	1	1	0	2
41	0	0	1	1	0	0	0	2
42	0	0	0	1	0	1	1	2
43	0	1	0	0	1	0	1	2
44	0	1	1	1	0	1	0	2
45	1	1	0	1	0	0	1	1
46	1	1	0	1	0	1	0	1
47	0	1	1	1	1	1	1	2
48	0	0	0	0	1	0	0	2
49	0	0	0	0	1	1	0	2
50	0	0	0	0	0	0	1	2

Numero	Indicadores						
51	1	0	1	0	1	0	1
52	0	0	0	0	1	1	1
53	1	1	0	0	1	0	1
54	1	0	0	1	1	1	1
55	1	0	1	0	1	1	1
56	1	0	0	1	1	0	0
57	1	0	0	0	1	0	1
58	0	1	0	1	1	0	1
59	0	1	1	1	1	1	0
60	1	1	1	0	1	1	0
61	1	0	0	0	0	0	1
62	0	0	0	1	1	1	0
63	1	1	0	1	0	1	1
64	1	1	0	0	0	1	0
65	1	0	1	0	1	0	1
66	0	0	1	0	0	0	1
67	0	0	0	1	0	0	0
68	1	1	1	1	1	0	1
69	1	0	1	0	0	1	0
70	0	1	1	0	1	1	0
71	1	1	0	1	1	0	0
72	1	0	1	1	0	0	0
73	0	0	1	1	0	0	0
74	0	1	1	1	1	0	1
75	1	1	1	0	0	1	0
76	0	1	1	1	0	1	0
77	1	0	0	0	1	1	0
78	1	1	1	1	0	1	1
79	1	1	0	1	0	1	1
80	0	1	0	1	0	0	1
81	0	0	0	0	1	0	1
82	0	0	1	1	1	0	0
83	0	0	0	0	1	1	1
84	0	1	0	1	0	0	0
85	0	0	1	0	0	0	1
86	1	0	0	0	1	1	0
87	0	0	0	1	0	0	0
88	0	1	1	1	1	0	1
89	1	1	0	1	0	1	1
90	1	0	1	1	0	0	1
91	1	1	1	0	0	0	1
92	0	1	0	0	0	0	0
93	1	1	0	0	0	1	0
94	1	0	1	0	0	1	0
95	1	1	0	1	1	0	0
96	0	1	0	1	0	1	0
97	0	1	1	0	0	1	0
98	0	0	0	1	0	0	0
99	0	0	0	1	1	0	0
100	0	1	1	0	0	1	0

BANCO DE DATOS 2

ESTADOS	Cambio de Población	INDICADORES				
Arkansas	0	2.878	1.8	0	0	1.1
Colorado	1	3.855	1.9	0	1	.8
Delaware	1	4.524	1.9	1	1	.9
Georgia	1	3.354	2.1	1	0	.9
Idaho	0	3.290	1.9	0	0	.8
Iowa	0	3.751	1.7	0	0	1.0
Mississippi	0	2.626	2.2	1	0	1.0
New Jersey	1	4.701	1.6	1	1	.9
Vermont	1	3.468	1.8	0	0	1.0
Washington	1	4.053	1.8	1	1	.9
Kentucky	0	3.112	1.9	0	0	1.0
Louisiana	1	3.090	2.7	1	0	1.3
Minnesota	1	3.859	1.8	0	0	.9
New Hampshire	1	3.737	1.7	1	0	1.0
North Dakota	0	3.086	1.9	0	0	.9
Ohio	0	4.020	1.9	0	1	1.0
Oklahoma	0	3.387	1.7	0	0	1.0
Rhode Island	0	3.959	1.7	1	1	1.0
South Carolina	0	2.990	2.0	1	0	.9
West Virginia	0	3.061	1.7	0	0	1.2
Connecticut	1	4.917	1.6	1	1	.8
Maine	0	3.302	1.8	1	0	1.1
Maryland	1	4.309	1.5	1	1	.8
Massachusetts	0	4.340	1.7	1	1	1.0
Michigan	1	4.180	1.9	0	1	.9
Missouri	0	3.781	1.8	0	1	1.1
Oregon	1	3.719	1.7	1	0	.9
Pennsylvania	0	3.971	1.6	1	1	1.1
Texas	1	3.606	2.0	1	1	.8
Utah	1	3.227	2.6	0	1	.7
Alabama	0	2.948	2.0	1	0	1.0
Alaska	1	4.644	2.5	1	0	1.0
Arizona	1	3.665	2.1	0	1	.9
California	1	4.493	1.8	1	1	.8
Florida	1	3.738	1.7	1	1	1.1
Nevada	1	4.563	1.8	0	1	.8
New York	0	4.712	1.7	1	1	1.0
South Dakota	0	3.123	1.7	0	0	2.4
Wisconsin	1	3.812	1.7	0	0	.9
Wyoming	0	3.815	1.9	0	0	.9
Hawaii	1	4.623	2.2	1	1	.5
Illinois	0	4.507	1.8	0	1	1.0
Indiana	1	3.772	1.9	0	0	.9
Kansas	0	3.853	1.6	0	0	1.0
Montana	0	3.500	1.8	0	0	.9
Nebraska	0	3.789	1.8	0	0	1.1
New Mexico	0	3.077	2.2	0	0	.7
North Carolina	1	3.252	1.9	1	0	.9
Tennessee	0	3.119	1.9	0	0	1.0
Virginia	1	3.712	1.8	1	0	.8

BANCO DE DATOS 3

Iris setosa				*Iris versicolor*				*Iris virginica*			
Indicadores				Indicadores				Indicadores			
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7	2·8	6·7	2·0
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3	2·7	4·9	1·8
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7	3·3	5·7	2·1
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2	6·0	1·8
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8	4·8	1·8
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0	4·9	1·8
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8	5·6	2·1
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

BANCO DE DATOS 4

Clase 1

1.35 1.47 -.39 .69 .86 .46 -.49
-.54 1.09 .17 .57 .14 .24 .66
-1.57 .76 .20 -1.29 1.79 .90 2.60
.35

1.80 -.44 .60 .14 .00 1.42 2.40
-.38 .40 -.57 -.12 .85 -2.51 .65
.81 .37 .64 -.01 1.21 .10 .71
-.24

-.55 .37 .40 .52 1.49 .15 .10
.16 -.29 .58 -1.47 -1.01 -.27 .13
1.42 3.87 -.47 -.92 -1.98 1.10 .15
.86 -.57 -.38 .69 1.16 -1.62 1.32
.74 .97

.77 1.01 .39 -1.50 -.86 -1.14 -1.35
.16 .68 -.62 .47 -.85 2.10 .60
.60 -1.05 2.11 1.26 -.63 .12 .89
-.98 .95 .25

.76 .73 .58 -.50 -.60 .41 .52
-.24 1.59 -.34 2.11 .31 -.18 -.74
.70 .21 .27 2.72 -.37 1.03 -1.47
-.58 .80 .31 -.90 1.19 -1.32 .47
.45 2.61 -.31

-1.04 -.85 1.87 .82 1.67 -.30 .35
.81 .18 -.41 -1.39 .61 1.11 .40
1.66 .74

1.93 -1.01 -.95 .39 -1.23 -.36 1.13
-.24 .68 .32 1.23 -.57 -1.14 -.73
1.13 2.01 .38 .24 -.56 .43 -1.23
-.69

.58 .67 .61 .01 .48 1.06 .24
.04 -2.09 .07 1.50 1.71 .70 -.47
.52 .94 .79 1.63 -.17 .43 -.26
.53 .59 -1.40 .21 1.91 1.06 1.64
.71 -.38 .06

-.24 -.25 1.57 .54 .58 2.17 -.62
.60 .10 -.70 .34 -.04 1.31 -.41
.42 -.32 -1.66 .85

2.04 1.33 1.29 .04 .13 .91 -.45
-1.07 .21 .63 .97 -.03 -.40 1.54
-.54 .95 1.45 -.03 1.38 -.09 -1.03
.33 .32 .26 -.06 1.52 -1.10 .47

Clase 2

-1.29 -.46 -.87 1.34 -.30 .08 -.15
-1.04 .00 -1.58 -.07 1.48 -1.30 -.20
-2.45 -1.16 -1.28 .17 .30 .10 -.93
-1.00 -.87 1.52 -.93 -1.42 -1.55 -.69

.41 .29 -.68 .81 1.53 .87 .77
-1.57 -1.35 .51 -1.37 -.77 -.57 -1.07
.17 -.37 2.10 -.32 .46 -1.11 -.49
-1.84 .74 -1.34 .81 .17 -1.11 .11

-1.79 -.28 .78 -1.57 -.76 .27 .10
.06 .06 -.22 2.01 1.77 -.63 .18
.18 .29 .48 .25 -.66 .38

-.35 -1.27 .83 -.40 -.40 -1.55 -.09
-1.62 .13 -1.37 .77 -.82 .45 -.43
-1.11 .58 .57 .48 1.12 .37 -.92
.28 -.04 -1.07 -.32 .52

-3.51 -2.34 .65 .68 -.77 -1.17 .72
-1.43 -1.32 -.56 -.39 -.92 -1.25 -.21
.47 -1.37 .71 -1.14 -.49

-.26 -.80 -1.06 -1.83 -1.52 .15 -2.38
-2.02 -.43 -.84 .79 1.36 .81 -.14
-.20 -1.48 .60 .17 -.76 -1.96 -1.85
-.14 -.66 .09 -.45 .35 -.98 1.65
.22 -1.18 -.20 .46 -.10 -.37

.59 -1.90 .12 -.67 1.41 .05 -.80
1.02 -1.85 .36 -.79 -.54 -.34 -.01
.10 -.55 -1.86 -.51 .34 .64 -.66
-.29 -2.00 -1.66 .03 .15 -1.74 -.35

1.95 .48 1.20 -1.28 .21 -.73 -1.21
-1.83 -.46 -.56 1.18 -1.09 .65 .76
.30 -.33 .72 -.79 -.47

-.49 .20 .65 .41 -.63 -.70 -1.06
-1.80 -2.09 1.03 -1.07 .62 .15 -1.34
-.91 1.48 -2.06 -.46 -.18 -.23 -2.93
-1.85 .49 -1.36 1.09 1.44 .09 -.91
-1.84 -1.14 -1.37 .41

-1.81 .72 .09 -.55 1.56 1.46 2.36
-.80 -1.02 -1.60 .12 -1.60 -.82 -1.39
-1.73 -.53 -1.30 -1.84 .99 1.03 -.14
2.55

BANCO DE DATOS 5

Clase 1

-.01 .47 .32 -.20 1.30 .72 .23 .82 .22 -.43
-.48 .69 .96 .92 -.03 -.39 -.62 -.60 1.65 1.31
.65 -1.01 .89 .63 -.01 .91 .15 .38 1.06 .32
.67 .87 -.29 -.70 1.09 .33 .89 1.16

.63 .19 .27 2.33 -.71 .20 -.31 2.68 .51 -.19
.20 .52 -.29 .71 -.03 .48 -.37 .83 .99 1.25
-.04 -.27 -.84 1.97 -.39 .67 1.82 1.06 -1.07 .12
.56 .10 .15 .11 -.27 .72 1.67 1.38 1.09

-.03 2.15 .28 -.37 .28 .04 .57 -.78 -.40 -.38
.59 -.50 -.28 .32 1.68 -.12 -.25 .03 .76 -1.21
-.34 -1.03 2.02 2.07 -.01 .97 1.08 .53 .31 -.29
-.13 .94 .96

-.62 .60 -.29 .36 .04 .39 .48 -.09 .63 1.02
1.46 -.32 -.32 -.28 -.44 -.53 .18 .57 .42 .47
-.23 .85 .18 -.25 .92 .93 3.08 1.47 1.79 1.18
.21 -.29 1.24 .45

.10 -.51 -.30 .80 .89 1.51 .15 -.62 -.81 .72
.17 -.00 .54 1.54 -.44 .53 .66 1.76 -.19 1.33
1.14 1.23 1.15 .51 .37 -.54 -.78 -.55 1.20 .40
1.47 .08 .74 .42 1.94 -.28 .74 1.04 .26 -.03
-1.42

-.27 -.22 .33 -.37 -.02 1.03 1.89 .76 2.59 .46
-.31 .10 2.37 .16 -.06 .07 .31 .29 1.83 -.56
1.13 2.68 .80 2.39 .66 -.50 .29 .51 -.21 .13
1.66 1.20 .53 2.39 -.47 -.39 .39 .18 -.23

.42 -.56 -.27 1.51 -.84 -.46 -.22 -.09 .48 .87
-.92 -.06 1.87 -.36 1.30 -.03 1.35 1.47 -.39 .69
.86 .46 -.15 .00 -.09 -.54 1.09 .17 1.48

.57 -.17 .30 .10 .14 .24 .61 .70 .20 1.79
1.52 -.93 .90 2.66 .35 1.80 .46 .26 .68 .61
.14 .00 .87 1.42 2.40 -.38 -.57 .40 1.71 -.52
1.94

-.62 .85 .12 -.39 .85 -.33 .46 .08 .35 .62
.74 1.21 .18 .81 .47 .37 .11 -.55 .37 .40
.52 1.49 -.15 .10 -.23 .78 .27 .16 .56 .06
.06 .13 1.42

3.07 -.22 2.01 1.79 -.63 1.10 .15 .17 .01 .10
.29 .46 -.38 .25 .47 1.16 1.30 .74 .99 .38
-.35 .77 1.01 .39 .83 -.40 -.09 .13 -.06 .77
-.82 .45 -.43 .68

Clase 2

-1.37 -.58 -1.42 -1.04 -1.29
-1.82 -.62 -1.64 -.92 -.51
-1.40 -1.51

-.36 -.91 -.84 -1.39 -1.57
-1.63 -.81 -1.59 -.36 -.67
-.81

-1.10 -1.32 -1.82 -.72 -1.15
-1.33 -1.22 -.94 -1.37 -.79
-.80 -1.65 -.27 -1.54 -1.12
-1.28 -1.47

-.87 -1.41 -1.41 -2.35 -.65
-.63 -1.26 -.82 -.44 -.19
-1.23 -.66 -2.08 -1.16 -1.27
-.83

-2.16 -1.02 -1.35 .03 -1.35
-1.68 -1.72 -1.32 -1.84

-.72 -.57 -1.02 -.80 -.93
-2.08 -1.08 -1.07 -1.42 -1.35
-.83

-1.35 -.94 -1.17 -1.85 -.41
-2.22 .38 -.63 -2.83 -.63
-1.29 -.46 -.87 -1.34 -.39
-.86 -1.01 -.49 -1.56 -1.31
-.28

-2.45 -1.16 -1.28 -1.57 -.93
-1.00 -1.29 -.87 -1.42 -1.55
-.67 -.44 -.64 -1.51 -.79
-1.17 -1.38 -.54 -1.36

-2.16 -2.51 -1.10 -.46 -1.69
-.66 -1.38 -1.18 -.24 -1.79
-1.57 -.76 -.10 -.29 -1.47
-1.01 -.27

-.47 -.93 -1.78 -.57 -.61
-1.62 -1.37 -.40 -1.55 -1.62
-1.50 -1.37 -1.14 -1.11 -1.35
.16

REFERENCIAS

AITCHISON, J. (1984). Practical Bayesian problems in simplex sample spaces. En *Bayesian Statistics II* (J.M. Bernardo *et al.* eds.). Amsterdan: North-Holland. (Con discusión). (En prensa).

AITCHISON, J. and DUNSMORE, I.R. (1975). *Statistical Prediction Analysis*. Cambridge: University Press.

AITCHISON, J., HABBEMA, J.D.F. and KAY, J.W. (1977). A critical comparison of two methods of statistical discrimination. *Appl. Statistics. 26*, pp. 15-25.

AITCHISON, J. and LAUDER, I.J. (1979). Statistical diagnosis for imprecise data. *Biometrika 66*, pp. 475-483.

ANDERSON, J.A. (1972). Separate sample logistic discrimination. *Biometrika 59*, pp. 19-35.

ANDERSON, T.W. (1958). *An introduction to multivariate statistical analysis* New-York: Wiley.

BARTLE, R.G. (1966). *The elements of integration*. New-York: John Wiley and Sons.

BERMUDEZ, J.D. (1979). *Modelos de decisión predictiva con aplicaciones médicas*. Tesis de licenciatura, Universidad de Valencia.

BERMUDEZ, J.D. (1981). La elección de tratamiento como problema de decisión predictiva. *Trabajos de Estadistica, 32, 3.* pp. 32-44.

BERMUDEZ, J.D. (1982). Clasificación Estadistica: Paradigma diagnóstico. En *Estudios dedicados a Juan Peset Alexandre*. Valencia: Imprenta universitaria. (Tomo I) pp. 299-309.

BERMUDEZ, J.D. (1984). Discusión al artículo de J. Aitchison, Practical bayesian problems in simplex sample spaces. En *Bayesian Statistics 2.*(J.M. Bernardo, D.V. Lindley, M.H. DeGroot, A.F.M. Smith, eds.). Amsterdan: North-Holland. (En prensa).

BERNARDO, J.M. (1975). *The use of information in design and analysis of scientific experimentation*.Ph.D. Thesis. University of London.

BERNARDO, J.M. (1978). Métodos bayesianos y diagnosis clínica. *Estadística Española 78-79*, pp. 39-56.

BERNARDO, J.M. (1979). Reference posterior distributions for bayesian inference. *J. Roy. Statist. soc. B 41.* pp. 113-147 (Con discusión).

BERNARDO, J.M. (1981). *Bioestadística, una prespectiva Bayesiana.* Barcelona: Vicens-Vives.

BERNARDO, J.M. (1983). Bayesian logistic diagnostic distributions. *Tech. Rep. nº 8/83.* Dept. Bioestadística, Univ. Valencia.

BERNARDO, J.M. and BERMUDEZ, J.D. (1984). The choice of variables in probabilistic classification. En *Bayesian Statistics 2.* (J.M. Bernardo, D.V. Lindley, M.H. DeGroot, A.F.M. Smith, eds.).Amsterdam: North-Holland. (Con discusión).(En prensa).

BRADLEY, R.A. and GART, J.J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika, 49.* pp. 205-214.

BRIER, G.W. (1950). Verification of forecasts expressed in terms of probability. *Month. Weather Rev. 78*, pp. 1-3.

CHANDA, K.C. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika, 41.* pp. 56-61.

CHAO, M.T. (1970). The asymptotic behavior of bayes' estimators. *The Annals of Math. Statists., 41.* pp. 601-608.

COX, D.R. (1970). *The analysis of binary data.* London: Methuen.

CRAMER, H. (1946). *Mathematical methods of statistics.* Princeton: University Press.

DAWID, A.P. (1970). On the limiting normality of posterior distributions. *Proc. Cambridge Philos. Soc. 67.* pp. 625-633.

DAWID, A.P. (1976). Properties of diagnostic data distributions. *Biometrics 32*, pp. 647-658.

DUNSMORE, I.R. (1966). A bayesian approach to classification. *J. Roy. Statist. B 28*, pp. 568-577.

FERRANDIZ, J. (1981). *Una alternativa bayesiana al contraste de hipótesis.* Tesis doctoral. Universidad de Valencia.

FINNEY, D.J. (1952/1978, 3ª ed.). *Statistical method in biological assay.* London: Charles Griffin.

FISHER, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics 7.* pp. 114-123.

GANTMACHER, F.R. (1959/1977). *Theory of matrices*.New-York: Macmillan Co.

GEISSER, S. (1964). Posterior odds for multivariate normal classification. *J. Roy. Statist. Soc. B. 26*, pp. 69-76.

GOOD, I.J. (1950). *Probability and the weighing of evidence*. London: Charles Griffin.

GRAYBILL, F.A. (1961). *An introduction to linear statistical models*. New-York: McGraw-Hill.

HUZURBAZAR, U.S. (1948). The likelihood equation, consistency and the maxima of the likelihood function. *Ann. Eugen. London, 14*. pp. 185-200.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A 186*, pp. 453-461.

JEFFREYS, H. (1939/1967 3ª ed.). *Theory of probability*. Oxford: Clarendon Press.

JOHN, P.W.M. (1971). *Statistical desing and analysis of experiments*. New-York: Macmillan Co.

JOHNSON, N.L. and KOTZ, S. (1970). *Continous univariate distributions-2: Distributions in statistics*. Boston: Houghton Mifflin Co.

JOHNSON, R.A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist. 41*. pp. 851-864.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist. 27*. pp. 887-906.

KUHN, T.S. (1962/1970 2ª ed. ampliada). *The structure of scientific revolutions*. Chicago: University Press.

LAUDER, I.J. (1980). Linear models for statistical diagnosis between three types. En *Proc. 42th ISI Session* (Manila, 1979). pp. 307-310. The Hague: ISI.

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist. 41*. pp. 802-828.

LINDLEY, D.V. (1965). *Introduction to probability and statistics. Part 1. Probability*. Cambridge: Cambridge University Press.

MANTEL, N. (1973). Synthetic retrospective studies and related topics. *Biometrics 29*, pp. 479-486.

PRESS, S.J. and WILSON, S. (1978).Choosing between logistic regression and discriminant analysis. *J. Amer. Statist. Assoc. 66.* pp. 783-801.

RALSTON, A. (1970). *Introducción al análisis numérico.* Mexico: Limusa.

RAO, C.R. (1965/1973). *Linear statistical inference and its applications.*New-York: Wiley.

RENYI, A. (1976). *Calculo de probabilidades.* Barcelona: Reverte

TALLIS,G.M. (1961). The moment generating function of the truncated multi-normal distribution.*J. Roy. Statist. Soc. B 23.*pp. 223-229.

TEATHER, D. (1974). Statistical techniques for diagnosis. *J. Roy. Statist. Soc. A 137.* pp. 231-244.

WALD, A. (1944). On a statistical problem arising in the classification of an individual in one of two groups. *Ann. Math. Statist. 13.* pp. 452.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist. 20.* pp. 595-600.

WALKER, A.M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. B 31.* pp 80-88.

WALKER, S.H. and DUNCAN, D.B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika 54.* pp. 167-179.

WELCH, B.L. (1939). Note on discriminant functions. *Biometrika 31.* pp. 218-220.

...AD DE VALENCIA FACULTAD DE CIENCIAS MATEMÁTICAS

Reunido el Tribunal que suscribe, en el día de la fecha, acordó otorgar, por unanimidad, a esta Tesis doctoral de

JOSE-DOMINGO BERMUDEZ EDO

la calificación de SOBRESALIENTE "CUM LAUDE"

Valencia, a 4 de MAYO de 1984

El Presidente

El Secretario,

UNIVERSIDAD DE VALENCIA - FACULTAD DE CIENCIAS MATEMÁTICAS - BURJASOT