Revista Colombiana de Estadística
No. 4 - 1981

CLASIFICACION CON VARIABLES DISCRETAS OBSERVADAS CON ERROR Y CON VARIABLES CONTINUAS

Víctor Hugo Prieto Bernal
Profesor Asociado
Universidad Nacional de Colombia

Abstract.

This article is concerned with a problem of classification with discretes variables observable with error and with continuous variables. A classification rule is given by the Bayesian method and the parameters are estimated by the method of Maximum likelihood described by N.D. Day (1969) for mixture of normal distributions.

Resumen.

Este artículo considera un problema de clasificación con variables discretas observables con error y con variables contínuas. Se da una regla de clasificación por el método Bayesiano y se estiman los parámetros por el método de máxima verosimilitud descrito por N.E. Day (1969) para mezcla de distribuciones normales.

1. INTRODUCCION.

El problema de discriminación entre dos grupos cuando la información disponible consta de variables discretas (binarias) y contínuas fue considerado por Krzanowski (1975).

En dicho trabajo se considera un vector de variables binarias $X=(x_1,x_2,\ldots,x_q)^1$ y uno de variables continuas $Y=(y_1, y_2,\ldots,y_p)^1$. Cada valor de las componentes de X define una celda, resultando así $M=2^q$ celdas. Se supone que la distribución de $W=(X, Y)$ en la población π_k $(k=1,2)$ es tal que, condicionalmente, $(Y|X=m)\sim N(\mu_k^{(m)},\Sigma)$ donde $X=m$ indica que X condujo a la celda m. Además, la probabilidad de obtener una observación de la celda m en la población π_k es p_{km}. Bajo estas hipótesis y suponiendo probabilidades a priori iguales y costos de mala clasificación iguales, se obtiene la siguiente regla de clasificación Bayesiana :

$$R_1: W\in\pi, \text{ si } (\mu_1^{(m)}-\mu_2^{(m)})'\,\Sigma^{-1}\left[Y-\frac{1}{2}(\mu_1^{(m)}+\mu_2^{(m)})\right]\geq \ell_n\,\frac{p_{2m}}{p_{1m}} \quad (1.1)$$

R_2: $W\in\pi_2$, en otro caso.

Además las probabilidades de mala clasificación están dadas por

$$P(1/2) = \sum_{m=1}^{M} p_{2m}\, P(D^{(m)} \geq 0 \mid \pi_2) \quad (1.2)$$

$$P(2/1) = \sum_{m=1}^{M} p_{1m}\, P(D^{(m)} < 0 \mid \pi_1)$$

donde

$$D^{(m)}=(\mu_1^{(m)}-\mu_2^{(m)})'\,\Sigma^{-1}\left[Y-\frac{1}{2}(\mu_1^{(m)}+\mu_2^{(m)})\right]-\ell_n\,\frac{p_{2m}}{p_{1m}}\,.$$

Cuando los parámetros en ambas poblaciones son desconocidos, como es el caso más usual en problemas prácticos,

se consideran dos muestras de tamaño n_1 y n_2 de π_1 y π_2, respectivamente. Estas muestras son utilizadas para construir una regla de clasificación estimada reemplazando los parámetros por las estimaciones de los parámetros en la regla dada anteriormente.

En el problema descrito hasta aquí, podría pensarse que la información proporcionada en las variables discretas es factible de contener error, lo cual nos lleva, en este trabajo, a considerar el problema teniendo en cuenta dos tipos de variables discretas, las observables con error y las no observables, además de las variables continuas. En esta situación, el propósito de este trabajo es determinar una regla de clasificación usando el método Bayediano, similar al analizado por Krzanowski, efectuando las estimaciones de los parámetros por el método de máxima verosimilitud descrito por N.E. Day (1969) para mezcla de distribuciones normales. Debe notarse que la situación en que las variables discretas se observan con error, se ha estudiado en problemas de análisis de datos categóricos, ver por ejemplo J. Hochberg (1977), quien presenta dos metodologías de estimación usando esquemas de doble muestreo.

2. DISCRIMINACION CON VARIABLES DISCRETAS OBSERVABLES CON ERROR Y CON VARIABLES CONTINUAS, ESTIMACION POR MAXIMA VEROSIMILITUD.

Sea $Z=(z_1,\ldots,z_q)'$ un vector de variables binarias observables con error y sea $X = (x_1,\ldots,x_q)'$ el vector

no observable que contiene los valores verdaderos de las variables Z_i $(i=1,2,\ldots,q)$.

Supongamos que si se observa un vector Z de variables binarias pertenecientes a la celda m-ésima, la probabilidad de que el verdadero valor de dichas variables sea dado por X perteneciente a la celda j-ésima, viene dada por

$$P(X|Z) = P(X = j \,|Z = m) = p_{jm},$$

y que la distribución conjunta de X y Z es dada por

$$P(X,Z) = q_{jn},$$

de modo que $p_{jm} = q_{jm} / (\sum_j q_{jm})$

Además, supongamos que la distribución condicional de Y dado Z,X, es $N(\mu_{Z,X}, \Sigma)$

Para Z fijo se tiene entonces que

$$\varphi(Y|Z) = \sum_X \varphi(Y|X, Z)\, p(X|Z)$$

$$= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \sum p_{jm}\left[\exp\{-\frac{1}{2}(Y-\mu_j^{(m)})' \Sigma^{-1}(Y-\mu_j^{(m)})\}\right], \quad (2.1)$$

es decir, que la distribución condicional de Y, dado Z, es una mexcla de normales multivariantes.

Ahora si $p_k(W)$ es la densidad de probabilidad de

$W=(Z,Y)$ en π_k $(k=1,2)$, podemos tomar como criterio de clasificación (suponiendo probabilidades a priori iguales y costos de mala clasificación iguales) la de asignar W a π_1 si $p_1(W)/p_2(W) \geq 1$ y a π_2 de otra manera. Pero $p_k(W) = p_k(Z,Y) = p_k(Z)\,\mathcal{f}_k(Y|Z)$ $(k=1,2)$ y por tanto asignamos la observación $W=(Z,Y)$ a la población π_1 si

$$\frac{p_1(W)}{p_2(W)} = \frac{\mathcal{f}_1(Y|Z)\; p_1(Z)}{\mathcal{f}_2(Y|Z)\; p_2(Z)} \geq 1$$

y a la población π_2 de otra manera.

Se tiene entonces que :

$$\frac{\mathcal{f}_1(Y|Z)p_1(Z)}{\mathcal{f}_2(Y|Z)p_2(Z)} = \frac{\left[\sum_{j=1}^{M} \exp\{-\frac{1}{2}(Y-\mu_{ij}^{(m)})'\,\Sigma^{-1}(Y-\mu_{ij}^{(m)}) + \ell_n p_{jn}\}\right] p_1(Z)}{\left[\sum_{j=1}^{M} \exp\{-\frac{1}{2}(Y-\mu_{2i}^{(m)})'\,\Sigma^{-1}(Y-\mu_{2i}^{(m)}) + \ell_n p_{1n}\}\right] p_2(Z)} \geq 1 \qquad (2.2)$$

es decir

$$\frac{p_2(Z)}{p_1(Z)} \leq \frac{\sum_{j=1}^{M} \exp\{-\frac{1}{2}(Y-\mu_{1j}^{(m)})'\,\Sigma^{-1}(Y-\mu_{1j}^{(m)}) + \ell_n\, p_{jn}\}}{\sum_{i=1}^{M} \exp\{-\frac{1}{2}(Y-\mu_{2i}^{(m)})'\,\Sigma^{-1}(Y-\mu_{2i}^{(m)}) + \ell_n\, p_{im}\}} \qquad (2.3)$$

entonces, transformando la expresión (2.3) se obtiene :

$$\frac{p_2(Z)}{p_1(Z)} \leq \sum_{j=1}^{M} \frac{1}{\sum_{i=1}^{M} \exp\{Y'\,A_{ij} + b_{ij}\}} \qquad (2.4)$$

donde $A_{ij} = \Sigma^{-1}\,(\mu_{2i}^{(m)} - \mu_{ij}^{(m)})$,

$$b_{ij} = (\mu_{1j}^{(m)'} \Sigma^{-1} \mu_{1j}^{(m)} - \mu_{2i}^{(m)'} \Sigma^{-1} \mu_{2i}^{(m)}) + \ell_n \frac{p_{jn}}{p_{im}} \quad .$$

De nuevo, en la práctica los parámetros poblacionales generalmente son desconocidos de modo que la regla de clasificación dada no puede usarse directamente y es necesario estimar tales parámetros. Con este fín podemos considerar muestras de tamaños, n_1 y n_2 de las poblaciones π_1 y π_2, respectivamente, y para cada una seguir el método de máxima verosimilitud descrito por N. E. Day (1969) para estimar los parámetros de una mezcla de distribuciones normales. En este caso, para una muestra de tamaño n, se obtienen las siguientes ecuaciones de máxima verosimilitud :

$$\hat{p}_{jm} = \frac{1}{n} \sum_{i=1}^{n} P(j|Y_i) \qquad (j = 1,2, \ldots, M) \qquad (2.5a)$$

$$\hat{\mu}_j^{(m)} = \{ \sum_{i=1}^{n} Y_i P(j|Y_i) \} / \{ \sum_{i=1}^{n} P(j|Y_i) \} \quad (j=1,2,\ldots,M) \quad (2.5b)$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} (Y_i - \mu_j^{(m)})(Y_i - \mu_j^{(m)})' P(j|Y_i) \qquad (2.5c)$$

donde

$$P(j|Y_i) = \exp(A_j' Y_i + b_j) \Big/ \sum_{k=1}^{M} \exp(A_k' Y_i + b_k)$$

si $\quad A_j' = \Sigma^{-1} \mu_j^{(m)} + t \ ,$

$$b_j = \mu_j^{(m)'} \Sigma^{-1} \mu_j^{(m)} + \ell_n p_{jm} + h,$$

siendo t un vector y h un. constante escalar.

A partir de estas ecuaciones y mediante un proceso de iteración se obtienen las estimaciones de los parámetros en consideración.

Esta estimación podría realizarse por otros métodos tales como el de mínimos cuadrados descrito por el mismo Day (1969), o por el método de la función generatriz de momentos descrito por Quandt y Ramsey (1978). En todos estos casos es factible usar el método iterado (algoritmo EM) descrito por Dempster, Laird y Rubin (1977) para obtener las estimaciones.

BIBLIOGRAFIA

Day, N.E. (1969), Estimating the components of a mixture of Normal Distributions. Biometrika, 53,3, p. 463-474.

Hochberg, Y. (1977), On the use of Double Sampling Schemes in Analyzing Categorical Data with Missclassification Errors , Journal of the American Statistical Association, 72, 914-921.

Krzanowski, W.J. (1975) Discrimination and Classification Using Both Binary and Continuous Variables , Journal of the American Statistical Association, 70,782-790.

Dempster, A.P., Laird, N.M. , and Rubin, D.B. (1977), Maximum likelihood from Incomplete Data Via the EM

Algorithm , Journal of the Royal Statistical Society, ser B. vol. 39, 1-38.

Quandt, R. and Ramsey, B.R. (1978), Estimating Mixtures of Normal Distributions and Switching Regressions with Comments, Journal of the American Statistical Association ; 73, 730-752.