Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.5, No.4, 2015

On Discrimination and Allocation with Continuous and Dichotomous Variables

Mbaeyi, G. C.* Anyanwu, P. E.
Department Mathematics and Statistics, The Federal Polytechnic, P. M. B. 55, Bida, Niger State, Nigeria

Abstract.
In discriminant analysis involving continuous and categorical variables, the simplest and conventional procedure is to assign an arbitrary numerical score to each possible state of the categorical variables and proceed as if all variables are continuous. A discrimination procedure is suggested for use in a situation where the discriminating variables are mixtures of more than one Continuous variable and one Dichotomous variable. The performance of the suggested procedure is compared alongside that of the conventional Fisher's Linear Discriminant and Logistic Discrimination procedures based on their error rates. The suggested procedure performed better when compared with the other procedures. Hence, the suggested procedure will be applicable for such situation.

Keywords: Dichotomous, Continuous, Discriminant Analysis, Error Rate

1. Introduction

Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more sample. The classical methodological approach for Discriminant analysis assumes a jointly normal distribution for the explanatory variables, with equal covariance matrices among the groups. Most bodies of data involves observations associated with various facets of a particular background, environment or experiment, thus in general sense, data are always multivariate in character (Gnanadesikan, 1997). Example can be found in medicine (where continuous laboratory measurements may be included with such categorical variables as site of tumor and presence/absence of a certain symptom for each patient). Inclusion of all variables in a Discriminant analysis may lead to complication. If the categorical variables are ordered, the simplest and conventional procedure, the Linear Discriminant Function (LDF), is to assign an arbitrary numerical score to each possible state of the variable and proceed as if all variables are continuous. The most widely used method of discriminatiom, introduced by Fisher (1938) assumes that **x** is distributed according to the multivariate normal distribution, the covariance matrix being the same whether **x** arises from, Π_1 or Π_2 but the vector of means being different. Another common method of discrimination is that described by Warner et al (1961), assumes that each variable may take only two values and that the separate variables are independent. Although both method of discrimination have been used successfully in many problems, they are not always suitable (Krzanowski, 1980).

Some approaches to the treatment of mixed dichotomous and continuous variables in discriminant analysis are evident in literature. Aitchison and Aitken (1967) suggest a method

based on the kernel approach to estimating densities (Parzen, 1962). Anderson (1972) advocates the use of logistic discrimination in which the probability of group membership is assumed to be a logistic function of the observed variates. Krzanowski (1975) proposes a likelihood ratio method derived from a probabilistic model for mixed categorical (including dichotomous) and continuous variables. Vlachonikolis (1990) derived the predictive allocation rule for classification of observations involving mixtures of binary and continuous variables. His approach was based on the usual frequency distributions of the location model and vague prior distributions for the unknown parameters.

Owing to the lack of techniques for dealing with mixed continuous and categorical (dichotomous/polychotomous) variables, such data are frequently analyzed by methods originally intended for continuous variables only thereby introducing some amount of distortion.

This work seeks to study the performance of some discriminant procedures mixed continuous variable and dichotomous variable. We shall adopt the method adopt the proposed by Chang and Afifi (1974) and compare its result with that of the classical and conventional Fisher's Linear Discriminant Function (FLDF) and the Logistic Discriminantion (LD) in a special case where we have one dichotomous and more than one continuous variables.

The suggested method, unlike the FLDF will not treat the dichotomous variable as if they are continuous and also unlike the LD, would make assumption about the distribution from which the observation comes from. The comparison will be made based on the error rate of each of these methods, that is, the proportion of the observation that was wrongly classified by each of the methods.

2. Methodology

2.1 The Model

Let **X** be a univariate Bernoulli variable with parameter θ (0,1) and **Y** be a k-variate random vector of continuous variables. We assume that the conditional distribution of **Y** given X=x is k-variate normal whose mean vector $u^{(x)} = \boldsymbol{\mu} + \boldsymbol{x\Delta}$ and covariance matrix is $\sum^{(x)} = \sum + x\boldsymbol{\Gamma}$, x=0 or 1, with density function denoted by $\phi\left(\mu^{(x)}, \sum^{(x)}\right)$. Then the joint density function of $w = \begin{pmatrix} x \\ y \end{pmatrix}$ is

$$f(w) = \emptyset\left(\mu^{(x)}, \sum^{(x)}\right)\theta^{x}(1-\theta)^{1-x} \qquad (1)$$

The marginal distribution of **Y** is then a two-component mixed normal. X and **Y** are independent if and only $\mu^{(0)} = \mu^{(1)}$ and $\sum^{(0)} = \sum^{(1)}$.

For the problem of classifying an observation **w** into one of two population, Π_1 and Π_2, it is assumed that, if **w** is from Π_i, then its density function is

$$f_i(\boldsymbol{w}) = \phi\left(\mu_i{}^{(x)}, \sum^{(x)}\right)\theta_i^x(1-\theta_i)^{1-x_i},\ \text{i=1,2}$$

The unconditional covariance matrices of **Y** given X are equal for the population.

The likelihood ratio derived from (1) is

$$Z^{(x)} = \mathbf{y}^1\left(\boldsymbol{\Sigma}^{(x)}\right)^{-1}\left(\boldsymbol{\mu}_1^{(x)} - \boldsymbol{\mu}_2^{(x)}\right) - \frac{1}{2}\left(\boldsymbol{\mu}_1^{(x)} + \boldsymbol{\mu}_2^{(x)}\right)^T\left(\boldsymbol{\Sigma}^{(x)}\right)^{-1}\left(\boldsymbol{\mu}_1^{(x)} - \boldsymbol{\mu}_2^{(x)}\right) + xIn\left(\frac{\theta_1}{\theta_2}\right) + (1-x)In\left(\frac{1-\theta_1}{1-\theta_2}\right) \quad (2)$$

If q_i is the prior probability of drawing an observation from Π_i, C(j/i) be the cost of misclassifying an observation from $\Pi_i\ to\ \Pi_j$ and $K = \left(\frac{q_2 C(1/2)}{q_1 C(2/1)}\right)$. The Bayes procedure is to classify **w** in Π_1 if $Z^{(x)} \geq K, x = 0,1$. The Bayes procedure is to classify an observation according to the values of two discriminant functions: $Z^{(0)}$ for an observation whose value of X is 0 and $Z^{(1)}$ for an observation whose value of X is 1, henceforth known as the Double Discriminant Function (DDF). All information needed and used for allocation comes from Y and the values assigned to the dichotomous variable X is not involved in the decision although allocations are made based on the state of the dichotomous variable. Thus, the conditional and marginal distribution of Y are identically normal.

2.2 Estimation of the Parameters

In practice, population parameters are generally unknown. Information available are those which comes in form of initial sample sizes n_1 & n_2 from Π_1 and Π_2 respectively drawn from a larger population. The population parameters are then replaced by estimates obtained from the initial samples available. Denote by $y_{ij}^{(x)}$ an observation from Π_i whose values of X is x, where x=0,1, j=1,…,$n_i^{(x)}$ and $n_i^{(0)} + n_i^{(1)} = n_i$.

$$\bar{y}_i^{(x)} = \sum_{j=1}^{n_i^{(x)}} \frac{y_{ij}^{(x)}}{n_i^{(x)}} \quad (3)$$

$$\boldsymbol{A}_i^{(x)} = \sum_{j=1}^{n_i^{(x)}}\left(y_{ij}^{(x)} - \bar{y}_i^{(x)}\right)\left(y_{ij}^{(x)} - \bar{y}_i^{(x)}\right)' \quad (4)$$

The maximum likelihood estimates of the parameters are as follows;

$$\hat{\theta}_i = \frac{n_i^{(1)}}{n_i}, \qquad \mu_i^{(x)} = \bar{y}_i^{(x)}, \quad \hat{\Sigma}^{(x)} = \frac{1}{n_1^{(x)} + n_2^{(x)}}\left[\boldsymbol{A}_1^{(x)} + \boldsymbol{A}_2^{(x)}\right]$$

Unbiased estimate of $\hat{\Sigma}^{(x)}$ then is $\boldsymbol{S}^{(x)} = \frac{1}{n_1^{(x)}+n_2^{(x)}}\left[\boldsymbol{A}_1^{(x)} + \boldsymbol{A}_2^{(x)}\right]$

Since we assumed that $\Sigma^{(0)} = \Sigma^{(1)} = \Sigma$, then the unbiased (pooled) estimate of Σ is given as

$$S_p = \frac{1}{n_1+n_2-4}\sum_{i=1}^{2}\sum_{x=0}^{1}\boldsymbol{A}_i^{(x)} \tag{5}$$

We can now substitute S_p for $\boldsymbol{S}^{(x)}$ in $\boldsymbol{Z}^{(x)}$ and denote the resulting expression by $Z_p^{(x)}$

Data obtained from 57 patients classified as being Non-Hypertensive (Π_1) and Hypertensive (Π_2). The continuous variables are Age (y_1), Weight (y_2) and Fasting Blood Sugar (y_3). The dichotomous variable is Non-Diabetic (x=0) and Diabetic (x=1).

3. Result of Analysis and Discussion

Let $y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} AGE \\ WGT \\ FBG \end{pmatrix}$, $n_1^{(0)} = 12,\ n_2^{(0)} = 15,\ n_1^{(1)} = 20,\ n_2^{(1)} = 10, n_1 = 32, n_2 = 25,$ $n = 57$

$$\bar{y}_1^{(0)} = \begin{pmatrix} 39.08 \\ 88.83 \\ 4.66 \end{pmatrix}, \quad \bar{y}_1^{(1)} = \begin{pmatrix} 40.35 \\ 89.7 \\ 6.23 \end{pmatrix}, \quad \bar{y}_2^{(0)} = \begin{pmatrix} 34.8 \\ 82.73 \\ 4.67 \end{pmatrix}, \quad \bar{y}_2^{(1)} = \begin{pmatrix} 38.2 \\ 94.9 \\ 5.9 \end{pmatrix}$$

$$S^{(0)} = \begin{pmatrix} 131.4127 & 25.6817 & -0.4135 \\ 25.6817 & 156.808 & 0.0213 \\ -0.4135 & 0.0213 & 0.2039 \end{pmatrix}, \quad S^{(1)} = \begin{pmatrix} 94.2235 & -16.729 & -1.1922 \\ -16.729 & 291.825 & 0.8268 \\ -1.1922 & 0.8268 & 0.5171 \end{pmatrix}$$

$$S_p = \begin{pmatrix} 111.7655 & 3.276 & -0.8249 \\ 3.276 & 228.1377 & 0.4468 \\ -0.8249 & 0.4468 & 0.3694 \end{pmatrix}$$

The sample discriminant functions are

$$Z_p^{(0)} = 0.0377y_1 + 0.0261y_2 + 0.0255y_3 - 4.2252$$

$$Z_p^{(1)} = 0.0272 - 0.02511 + 0.9864y_3 - 4.2774$$

Assuming that prior probabilities and the cost of misclassification are equal for the two population, then, an observation (patient) is classified/predicted to be in Π_1 (Non-Hypertensive) if his/her $Z^{(0)}$ or $Z^{(1)}$ is non-negative.

The summary of the classification using each of the discrimination procedures is presented in tables below

Table 1. Summary Classification Result from the Double Discriminant Function

	Predicted	
Original	Π_1	Π_2
Π_1	22	10
Π_2	7	18

Table2. Summary Classification Results from FLDF

			Predicted Group Membership		
		GROUP	1	2	Total
Original	Count	1	20	12	32
		2	10	15	25
	%	1	62.5	37.5	100.0
		2	40.0	60.0	100.0

a. 61.4% of original grouped cases correctly classified.

Table 3. Summary Classification Table from the Logistic Discrimination

			Predicted		
			GROUP		
	Observed		1	2	Total
Step 1	GROUP	1	23	9	32
		2	11	14	25
	Overall Percentage				64.9%

From table 1, the overall percentage of correct classification using our suggested method was 70.2%, i.e 7 patients was wrongly classified as being Non-Hypertensive while 10 was wrongly classified as being Hypertensive thereby producing an error rate of 0.298.

Table 2 is the result using FLDF, it had 61.4% correct classification with an error rate of 0.386

Table 3 is the result using LD, it had 64.9% correct classification with an error rate of 0.351

Table 4. Summary of the methods used and their Error Rates

Method	Error Rate
DDF	0.298
FLDF	0.386
LD	0.351

From the results above, it can be clearly seen that, with mixtures of continuous and Dichotomous variables, certain amount of distortion is being introduced by treating categorical variables as if they are continuous. This is justified by the least error rate produced by the suggested method in which no transformation was made to the dichotomous variable. The FLDF and LD made more wrong classifications, reasons could possibly be due to the mistreating of the categorical variables as being continuous as in the case of FLDF and because the Logistic Discrimination approach does not depend upon any strict assumption, it could commit less error of misclassification, however, that has not made it better than the DDF in situations as that considered in this work.

4. Conclusion

The efficiency and choice of any discriminating procedure is often based on the error of misclassification it commits expected to be minimal though optimal since that is the best it can give. For the three methods compared so far in this work, the DDF performed better than the other methods hence can be amenable in situation where our discriminating variables comprises of continuous and one dichotomous variables. The suggested procedure can be used by developing simple computer programs to enable speed and accuracy. When the number of dichotomous variables is more than one, the parameters to be estimated becomes increasingly much thereby making us resort to the FLDF approach consequent upon the fact that the probability of misclassification becomes somewhat not reliable.

References

Aitchson, J. and Aitken, C. G. G. (1976) Multivariate binary discrimination by the Kernel method. *Biometrika* 63(3): 413-420

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

Anderson, J. A. (1972) Separate Sample Logistic Discrimination. *Biometrika,* 59, (Mar.1972) 19-35.

Chang, P. C. and Afifi, A. A. (1974) Classification based on Dichotomous and Continuous variables. *Journal of the American Statistical Association,* Vol. 69, No. 346, (June, 1974) 336-339.

Fisher, R. A. (1938). The Statistical Utilization of Multiple Measurements. *Ann. Eng. Lond.* 7, 179-88.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & sons, New York.

Krzanowski, W. J. (1975) Discrimination and Classification Using Both Binary and Continuous Variables, *Journal of the American Statistical Association*, 70:352, 782-790

Parzen, E. (1962) On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist*. Volume 33, Number 3, (1962) 1065-1076

Vlachonikolis, I. G. (1990) Predictive Discrimination and Classification with Mixed Binary and Continuous Variables. *Biometrika* (1990) 77, 3, 657-62

Warner, H., Toronto, A., Veesey, L., and Stephenson, R. (1961). A Mathematical Approach to Medical Diagnosis. *J. Amer. Med. As.* 177-183

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar