International Journal of Mathematics and Statistics Studies
Vol.2, No.4,pp.78-97, September 2014
Published by European Centre for Research Training and Development UK (www.eajournals.org)

A TWO-GROUP CLASSIFICATION MODELS FOR BINARY VARIABLES

[1.] I. Egbo; [2.] S.I. Onyeagu & [3.] D.D. Ekezie

[1.] Department of Mathematics, Alvan Ikoku Federal College of Education, Owerri. Nigeria

[2.] Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria

[3.] Department of Statistics, Imo State University, Owerri, Nigeria

---

ABSTRACT: *This paper is a study of two-group classification models for binary variables. Eight classification procedures for binary variables are discussed and evaluated at each of 118 configurations of the sampling experiments. The results obtained ranked the procedures as follows: Optimal, Linear discriminant, Maximum likelihood, Predictive, Dillon Goldstein, Full multinomial, Likelihood and Nearest neighbour. Also the result of the study show that increase in the number of variables improve the accuracy of the models.*

KEYWORDS: Classification Models, Optimal, Linear Discriminant, Maximum Likelihood, Predictive, Dillon Goldstein, Binary Variables, Misclassification, And Multinomial.

INTRODUCTION

Estimation of error rates has received considerable attention in the literature. The task of classification is to classify unknown objects into predefined classes based on their observed attributes using a classification model learned from a set of training data. Many applications such as characters recognition, decision-making and disease diagnosis, can be viewed as extensions of the classification problem (Hen and Kamber 2001). A classification instrument can be modeled using different structures such as decision graphs, decision trees, neural networks and rules. Reducing the processing time and increasing the classification rate are the two main issues in the classification problem. We consider a classical problem of discriminant analysis: an individual is to be allocated to one k distinct classes $w_1,\ldots w_c$, whose members are described by an r-component vector of binary variables $X= (x_1,x_2\ldots x_r)$. These binary variables can be viewed equivalently as a single multinomial variable having $S = 2^r$ states. The problem of classification is that of assigning item(s) into one of k, $k \geq 2$ known populations assuming that the items actually belong to one of the populations. Suppose only

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

two populations are admitted with infinite number of individual objects. Let there be r characteristics of interest with corresponding measurement variables X_1, $X_2 \ldots X_r$, r ≥1. Let the response vector of individual objects in π_1 be $X_1 = (X_{11}, X_{12} \ldots X_{1r})^1$ and in π_2 be $X_2 = (X_{21}, X_{22} \ldots X_{2r})^1$. Suppose we find an object 0 with measurement vector $X_0 = (X_{01}, X_{02} \ldots X_{0r})$ outside π_1 and π_2. The problem is how to classify 0 into π_1 and π_2 in an optimum fashion. The measurement vector X can be discrete or continuous. It can also be a mixture of discrete and continuous variables. In this study, our interest is about X whose arguments are discrete. The problem is to classify 0 with measurement vector X_0 into π_1 and π_2. In this inferential setting, the researcher can commit one of the following errors. An object from π_1 may be misclassified into π_2. Also an object from π_2 may be misclassified into π_1. If misclassification occurs, a loss is incurred. Let c(i/j) be the cost of misclassifying an object from π_j into π_i. The objective of the study is to find the 'Best' classification rule. "Best" here means the rule that minimizes the expected cost of misclassification (ECM). Such a rule is referred to as the optimal classification rule (OCR) in this study we want to find the OCR where X is discrete and to be more precise, Bernoulli. Whereas classification rules with optimal properties for discriminant problems with multivariate normally distribute d attribute variables are well known (Wald 1944, 1949; Smith, 1947; Adebanji, Adeyemi and Iyaniwura, 2008; Oludare, 2011), alternative rules be more appropriate if some of the attributes are skewed. Most of the studies that compared non-normal classification methods with normality-based methods for various different data conditions have assumed equal misclassification costs across groups. Hence, it is not clear to what extent the conclusions in these studies can be generalized to typical problems with distributions that are skewed with unequal misclassification costs across groups. The purpose of the current study is to establish guidelines for choosing an appropriate classification method if the problem at hand is characterized by Bernoulli multivariate data. To achieve this objective, several Monte Carlo simulation experiments are conducted to compare the performance of some traditional classification methods designed specifically to handle problems with Bernoulli multivariate data. This study is limited to the two-group classification problem.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

CLASSIFICATION PROCEDURES

The Optimal Classification Rule

According to Onyeagu (2003), let π_1 and π_2 be any two multivariate Bernoulli populations. Let q_i be the prior probability of π_i $i=1,2$ with $q_1+q_2=1$ and probability mass Function $f_i(x)$. Suppose that we assign an item with response pattern X into π_1 if it is in some region R_1 and to π_2 if it in some region R_2 where $R_1 \cup R_2 = R$. The expected cost of misclassification is given by:

$$ECM = c(2/1)q_1 \sum_{R_2} f(x/\pi_1) + c(1/2)q_2 \sum_{R_1} f(x/\pi_2) \qquad 2.1.1$$

The optimal rule is one that determines R_x such that ECM is a minimum.

$$R_1 = \{X \in R / c(1/2)q_2 f(x/\pi_2) - c(2/1)q_1 f(x/\pi_1) \le 0\} \qquad 2.1.2$$

$$R_1 = \left\{ X \in R \left| \frac{f(x/\pi_1)}{f(x/\pi_2)} \ge \frac{c(1/2)q_2}{c(2/1)q_1}, f(x/\pi_2) \ne 0 \right. \right\} \qquad 2.1.3$$

Therefore the optimal classification rule is: classify an item with response pattern X_0 into π_1 if

$$\frac{f_1(x_0/\pi_1)}{f_2(x_0/\pi_2)} > \frac{q_2 c(1/2)}{q_1 c(2/1)} \qquad 2.1.4$$

Otherwise classify the item into π_2. Without loss of generality, we can assume that $c(1/2) = c(2/1)$. Then minimization of ECM becomes minimization of the probability of misclassification p(mc). The optimal rule reduces to classify an item with measurement X_0 into π_1 if $\frac{f_1(x_0/\pi_1)}{f_2(x_0/\pi_2)} \ge 1$ 2.1.5

otherwise classify into π_2. Since X is multivariate Bernoulli with $p_{ij} > 0,\ i=1,2$ $j=1,2,...r$, the optimal rule is: classify an item with response pattern X into π_1 if

$$\sum_{j=1}^{r} X_j In\left(\frac{p_{1j}q_{2j}}{q_{1j}p_{2j}}\right) > \sum_{j=1}^{r} In\frac{q_{2j}}{q_{1j}} \qquad 2.1.6$$

otherwise classify into π_2.

For the optimal classification rule, we considered two cases:

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Case I known parameters

(i) General case where $p_i = (p_{i1}, p_{i2} \ldots p_{ir})$

(ii) Special case where $p_i = (p_i, p_i \ldots p_i)$

(iii) Special case (1b) with additional assumption that $p_1 = \theta p_2, 0 < \theta < 1.$

Case 2 Unknown parameters

(i) General case $p_i = (p_{i1}, p_{i2} \ldots p_{ir})$ we estimate p_1 and p_2 by taking training samples of size n_1 and n_2 from π_1 and π_2 respectively.

(ii) Special case where $p_i = (p_i, p_i \ldots p_i)$. We also estimate p_1 and p_2.

(iii) Special case (2ii) with $p_1 = \theta p_2, 0 < \theta < 1$, we take training samples of size n_2 from π_2 and estimate p_2. For a fixed value of $\theta, p_1 = \theta p_2$ for case 1(ii) the classification rule reduces to: classify item with response pattern X into π_1 if:

$$\sum_{j=1}^{r} X_j \leq \frac{rIn\left(\frac{q_2}{q_1}\right)}{In\left(\frac{p_1 q_2}{q_1 p_2}\right)} \qquad 2.1.7$$

otherwise classify into π_2

The probability of misclassification is given by:

$$p(mc) = \frac{1}{2}\left[1 + B_{(r,p_2)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{q_1}\cdot\frac{q_2}{p_2}\right)}\right) - B_{(r,p_1)}\left(\frac{rIn(\frac{q_2}{q_1})}{In\left(\frac{p_1}{q_1}\cdot\frac{q_2}{p_2}\right)}\right)\right] \qquad 2.1.8$$

$$= \frac{1}{2}\left[1 + B(r, p_2, \lambda) - B(r, p_1, \lambda)\right] \qquad 2.1.9$$

where $\lambda = \dfrac{rIn\left(\frac{q_2}{p_1}\right)}{In\left(\frac{p_1}{q_1}\cdot\frac{q_2}{p_2}\right)}$

for case 1(iii) the probability of misclassification is

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$$p(mc)=\frac{1}{2}\left[1+B_{(r,p_2)}\left(\frac{rIn\left(\frac{1-q_2}{1-\theta p_2}\right)}{In\theta\left(\frac{1-p_1}{1-\theta p_2}\right)}\right)\right]-B_{(r,\theta p_1)}\left[\frac{rIn\left(\frac{1-p_2}{1-\theta p_1}\right)}{In\theta\left(\frac{1-p_2}{1-\theta p_2}\right)}\right] \quad 2.1.10$$

$$=\frac{1}{2}\left[1+B(r,\,p_2-\lambda)-B(r,\theta p_1-\lambda)\right] \quad 2.1.11$$

For cases 2(ii) and 2(iii) the formula remains the same except that the parameters are estimated by their MLE estimates.

Full Multinomial Rule

Suppose we have a d-dimensional random vector $x^1=(x_1,\ldots x_d)$ where each $x_j,\ j=1,\ldots d$ assumes one of the two distinct values: 0 or 1. The sample space then has a multinomial distribution consisting of the 2^d possible states. Given two disjoint populations, π_1 and π_2 with priori probabilities p_1 and p_2, the density is

$$f(x)=p_1 f_1(x)+p_2 f_2(x) \quad 2.2.1$$

The two group problem attempts to find an optimal classification rule that assigns a new observation x to π_1 if

$$f_1(x)/\,f_2(x)>p_2/p_1 \quad 2.2.2$$

When x has only two states, it will be a binomial random variable with $n_i(x)$ observation from π_i and expected value $np_i f_i(x), i=1,2.$ Estimates for prior probabilities can be obtained by $\hat{p}_i=\frac{n_i}{n}$, where $n=n_1+n_2$ represents the total number of sample observations. The full multinomial model estimates the class-conditional densities by

$$f_i(x)=\frac{n_i(x)}{n},\ i=1,2. \quad 2.2.3$$

where $n_i(x)$ is the number of individuals in a sample of size n_i from the population having response pattern X . The classification rule is: classify an item with response pattern X into π_i if

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$$q_1 \frac{n_1(x)}{n_1} > q_2 \frac{n_2(x)}{n_2} \qquad 2.2.4$$

and to π_2 if $q_1 \frac{n_1(x)}{n_1} < q_2 \frac{n_2(x)}{n_2}$ 2.2.5

and with probability $\frac{1}{2}$ if $q_1 \frac{n_1(x)}{n_1} = q_2 \frac{n_2(x)}{n_2}$ 2.2.6

The full multinomial rule is simple to apply and the computation of apparent error does not require rigorous computational formula. However, Pires and bronco (2004) noted as pointed out by Dillon and Goldstein (1978) that one of the undesirable properties of the full multinomial Rule is the way it treats zero frequencies. If $n_1(x) = 0$ and $n_2(x) \neq 0$, a new observation with vector X will be allocated to π_2, irrespective of the sample sizes n_1 and n_2.

The Predictive Rule

If the non-informative conjugate prior distribution for the parameter P_i of the multinomial model is chosen, that is the Dirichlet distribution with parameter $\alpha = 1$, then the posterior distribution will be a Dirichlet distribution with parameter z_i+1, where $z_i = (n_{i1},\ldots n_{is})^T$. (Note that $\sum_{j=1}^{s} n_{ij} = n_i$). The Dirichlet distribution with parameter α has a density function given by

$$f(p) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_s)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_s)} \prod_{i=1}^{s} p_i^{\alpha_i - 1}, 0 < P_i < 1, \sum_{i=1}^{s} P_i = 1, \alpha_i > 0 \qquad 2.3.1$$

Therefore the predictive density is simply

$$h_i(x / z_i) = \int_p p_{ij} \frac{\Gamma(s + n_i)}{\Gamma(1 + n_{i1})\ldots\Gamma(1 + n_{is})} \prod_{k=1}^{s} p_{ik}^{nik} dp \qquad 2.3.2$$

$$= \frac{n_{ij} + 1}{n_i + s} = \frac{n_i(x) + 1}{n_i + s} \qquad (0 < p_{ij} < 1, \sum_{j=1}^{s} p_{ij} = 1 \qquad 2.3.3$$

which leads to the predictive rule (or the P-rule)

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Classify in π_1 if: $\frac{n_1(x)+1}{n_1+s} > \frac{n_2(x)+1}{n_2+s}$ 2.3.4

Classify in π_2 if: $\frac{n_1(x)+1}{n_1+s} < \frac{n_2(x)+1}{n_2+s}$ 2.3.5

Classify randomly if: $\frac{n_1(x)+1}{n_1+s} = \frac{n_2(x)+1}{n_2+s}$ 2.3.6

Once again for $n_1 = n_2$, this rule is equivalent to the M-rule. The P-rule also avoids the zero frequency problems. For instance $n_1(x) = 0$ and $n_2(x) < (n_2+s)/ (n_1+s)-4$ leads to classification in π_1.

The Likelihood Rule

Consider the generalized ratio test for the hypothesis H0: X, $X_{11}...X_{1n} \sim f_1(x)$ and $X_{21}... X_{2n} \sim f_2(x)$ against H1: $X_{11}...X_{1n1} \sim f_1(x)$ and $X_{21}...X_{2n_2} \sim f_2(x)$. As was proposed by Anderson (1982), Pires & Bronco (2004) and Onyeagu et al (2013) found that the likelihood ratio criterion also handles the problem of zero frequency. For multinomial model, they proposed a test statistic that is a function of X and is given by:

$$L(x) = \frac{\left[1+\frac{1}{n_1(x)}\right]^{n_1(x)} \cdot \left(n_1(x)+1\right)}{\left[1+\frac{1}{n_2(x)}\right]^{n_2(x)} \cdot \left(n_2(x)+1\right)} x \frac{\left(1+\frac{1}{n_2}\right)^{n_2}\left(n_2+1\right)}{\left(1+\frac{1}{n_1}\right)^{n_1}\left(n_1+1\right)} \quad 2.4.1$$

This rule fails to take account of several factors that may be important in practice. These factors are the differential prior- probabilities of observing individuals from the two populations and differential cost incurred by misclassification and a-prior probabilities and if $n_1(x) = 0$ and $n_2(x) = 0$, the classification rule becomes: Classify item with response pattern into π_1 if $L(x) > 1$ and into π_2 $L(x) < 1$. For $n_1 = n_2$, this rule falls back to the Full Multinomial Rule. The L Rule also solves the zero frequency problem. A new observation X with $n_1(x) = 0$ will be classified in π_1 if and only if

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$$\left(1+\frac{1}{n_2(x)}\right)^{n_2(x)} \text{ x } n_2(x)+1<C \qquad 2.4.2$$

where c is the value of the second fraction in (2.4.1)

The Dillon-Goldstein Rule

Dillon and Goldstein (1978) proposed the following rule as a result of the problem *arising* from zero frequency. The rule called the D-rule is based on Matusita's distribution distance using the notation: $n_i(x) = n_{ij}$ if x belong to state j. The rule is classify item into π_1 if:

$$\frac{\left[n_{2j}(n_{1j}+1)\right]^{\frac{1}{2}}+\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}}}{\left[n_{1j}(n_{2j}+1\right]^{\frac{1}{2}}+\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}}}<\left[\frac{n_2(n_1+1)}{n_1(n_2+1)}\right]^{\frac{1}{2}} \qquad 2.5.1$$

and to π_2 if

$$\frac{\left[n_{2j}(n_{1j}+1)\right]^{\frac{1}{2}}+\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}}}{\left[n_{1j}(n_{2j}+1\right]^{\frac{1}{2}}+\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}}}>\left[\frac{n_2(n_1+1)}{n_1(n_2+1)}\right]^{\frac{1}{2}} \qquad 2.5.2$$

randomly classify if

$$\frac{\left[n_{2j}(n_{1j}+1)\right]^{\frac{1}{2}}+\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}}}{\left[n_{1j}(n_{2j}+1\right]^{\frac{1}{2}}+\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}}}=\left[\frac{n_2(n_1+1)}{n_1(n_2+1)}\right]^{\frac{1}{2}} \qquad 2.5.3$$

Note that if $n_1 = n_2$, the D-rule reduces to the Full Multinomial Rule. For $n_1 < n_2$ and $n_{1j} = 0$ and $n_{2j} = 0$ the rule becomes: classify x into π_1 if:

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$$\sqrt{n_{2j}} < \left[\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}} - 1\right]\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}} \qquad 2.5.4$$

and to π_2 if $\sqrt{n_{2j}} > \left[\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}} - 1\right]\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}}$ 2.5.5

and randomly classify if

$$\sqrt{n_{2j}} = \left[\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}} - 1\right]\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}} \qquad 2.5.6$$

However, suppose $n_2 > n_1$ and $n_{2j} = 0$ but $n_{1j} > 0$ we shall classify item with response pattern x into π_1 if:

$$\left[1-\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}}\right]\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}} < \left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}}\sqrt{n_{2j}} \qquad 2.5.7$$

and to π_2 if:

$$\left[1-\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}}\right]\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}} > \left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}}\sqrt{n_{2j}} \qquad 2.5.8$$

randomly classify if:

$$\left[1-\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}}\right]\sum_{k\neq j}(n_{1k}n_{2k})^{\frac{1}{2}} = \left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{\frac{1}{2}}\sqrt{n_{2j}} \qquad 2.5.9$$

Nearest Neighbour procedure

Hills (1967) introduced perhaps the simplest nearest neighbour estimator for binary data, which classifies a particular response vector x based on the number of cells in response

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

vectors y that differ from x. Specifically, let k be the number of cells in which x and y differ. Then define $R_j = \{y_j / (x - y_j)^1 (x - y_j) \leq k\}$ to be a rule which classifies x if each of its cells differs by no more than k components. That is, classify x into π_1 if:

$$\sum_{R_j} \frac{n_1(y_j)}{n_1} > \sum_{R_j} \frac{n_2(y_j)}{n_2} \qquad 2.6.1$$

and into π_2 otherwise.

For example, with d =3 and x =(111), the neighbours of order k =1 are R_{111} = 110, 101, 011. Note that k =0 reduces to the full multinomial model. In practice, one simply needs to construct the table of frequencies for all possible pattern of x and use a counting procedure over the set Rj to form the sample-based likelihood ratio for classification purpose. If the cell count for the jth cell is nij, then the nearest neighbour procedure assigns the observation to

π_1 if $$\frac{\left[n_{ij} + \sum_A n_{ij}\right] / n_1}{\left[n2j + \sum_A n_{2j}\right] / n_2} > \frac{p_2}{p_1} \qquad 2.6.2$$

where A is the set of neighbour of state j. Hills comments that the estimate of the likelihood ratio has less sampling variability than the simple method using cell frequencies.

The Linear Discriminant Rule

The linear discriminant function for discrete variables is given by

$$\hat{L}(x) = \sum_j \sum_k (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} x_k - \tfrac{1}{2} \sum \sum (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} (\hat{p}_{2k} + \hat{p}_{1k}) \qquad 2.7.1$$

where s^{kj} are the elements of the inverse of the pooled sample covariance matrix, $\hat{p}_{1j}$ and $\hat{p}_{2j}$ are the elements of the sample means in π_1 and π_2 respectively. The classification rule obtained using this estimation is: classify an item with response pattern X into v if

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$$\sum_j \sum_k (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} X_k - \tfrac{1}{2} \sum_j \sum_k (\hat{p}_{2j} - \hat{p}_{1j}) s^{kj} (\hat{p}_{2k} + \hat{p}_{1k}) > 0 \qquad 2.7.2$$

and to π_2 or otherwise.

Maximum Likelihood Rule (ML Rule)

The maximum likelihood discriminant rule for allocating an observation x to one of the population $\pi_1 ... \pi_n$ is to allocate x to the population which gives the largest likelihood to x. that is the maximum likelihood rule says one should allocate x to π_j when

$$L_i = \max L_i(x) \qquad \text{Anderson (1984)}$$

Theorem: if π_i is the $N_p(\mu_i, \sum)$ population, $i = 1...g$ and $\sum > 0$, then the maximum likelihood discriminant rule allocate x to π_j where $j \in \{1,...n\}$ is that value of i which minimized the Mahalanobis distance $(x-\mu)^1 \sum^{-1} (x-\mu)$ where $g = 2$ the rule allocate x to π_1 if $\alpha^1 (x-\mu) > 0$ and $\alpha^1 \{x - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_2)\} > 0$, where $\alpha = \sum^{-1} (\mu_1 - \mu_2)$ and $\mu = (\mu_1 + \mu_2)$ and to π_2 or otherwise.

Application with life data

The data used in this example was collected at the University of Nigeria Teaching Hospital, Enugu. The data is made up of the following categories of heart disease patients. (i) Heart failure, (ii) Hypertensive heart failure, (iii) Hypertensive heart failure with stroke in evolution (iv) congestive heart failure (v) cardiovascular accident. Here, there are two populations (i) those who survived the attack, (ii) those that died. Three variables were used (i) systolic blood pressure (ii) Diastolic blood pressure (iii) Heart rate. The following measurements were obtained from 264 patients, 174 in the first population (π_1) and 90 in the second population (π_2) respectively. The systolic blood pressure is normal if it is less than 140mmHg i.e 90-140. The diastolic blood pressure is normal if it is less than 90mmHg i.e 60-90. The heart rate is normal if it is less than 100 beats per minute i.e (60-100). With the above information we dichotomized the measurements. The variables are:

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Variable 1: systolic blood pressure =1 if it is less than 140mmHg = 0 otherwise.

Variable 2: diastolic blood pressure = 1 if it is less than 90mmHg = 0 otherwise.

Variable 3: heart rate = 1 if it is less than 100 beats per minute = 0 otherwise.

Let $x=(x_1,x_2,x_3)$ denote the total response to the measurements and this leads to the following 2^8 response patterns: 000,100,010,110,001,101,011,111.The frequency of each of these response patterns in π_1 and π_2 are recorded in the following table.

State (x_1,x_2,x_3)	**Survival group Frequency**	**Non survival group Frequency**
000	23	9
100	8	3
010	5	6
110	18	20
001	58	25
101	11	3
011	8	5
111	43	19
Total	**174**	**90**

We have used the whole data to compute as follows:

$$\hat{p}_{ij} = \frac{\sum_{k=1}^{n_i} x_{ijk}}{n_i}$$

$$\hat{p}_1 = \left(\frac{80}{174}, \frac{74}{174}, \frac{120}{174}\right) = (0.4598, 0.4253, 0.6897)$$

$$\hat{p}_2 = \left(\frac{45}{90}, \frac{50}{90}, \frac{52}{90}\right) = (0.5000, 0.5556, 0.5778)$$

Using these estimates, we obtained the classification rule as classify the item with response pattern into π_1 if

$$R_{B3} : In\left(\frac{p_{11}}{q_{11}} \cdot \frac{q_{21}}{p_{21}}\right)x_1 + In\left(\frac{p_{12}}{q_{12}} \cdot \frac{q_{22}}{p_{22}}\right)x_2 + In\left(\frac{p_{13}}{q_{13}} \cdot \frac{q_{23}}{p_{23}}\right)x_3 > In\left(\frac{q_{21}q_{22}q_{23}}{q_{11}q_{12}q_{13}}\right) \qquad 3.1.183$$

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Otherwise classify the item into π_2.

Substituting the values of p_{1j} and p_{2j} above we have

$$In\left(\frac{0.4598}{0.5402}\cdot\frac{0.5000}{0.5000}\right)x_1 + In\left(\frac{0.4253}{0.5747}\cdot\frac{0.4444}{0.5556}\right)x_2 + In\left(\frac{0.6897}{0.3103}\cdot\frac{0.4222}{0.5778}\right)x_3$$
$$> In\left(\frac{0.5000}{0.5402}\cdot\frac{0.4444}{0.5747}\cdot\frac{0.4222}{0.3103}\right)$$

$$-0.1611x_1 - 0.5244x_2 + 0.4850x_3 > -0.0265$$

Response Pattern	**Classification**
000	π_1
100	π_2
010	π_2
110	π_2
001	π_1
101	π_1
011	π_2
111	π_2

$$\hat{p}(1/2) = p\left\{x = (000)(001),(101)/\hat{p}_2\right\} = (0.5000,0.5556,0.5778)$$

$$= \hat{q}_{21},\hat{q}_{22}\,\hat{q}_{23} + \hat{q}_{21}\,\hat{q}_{22}\,\hat{p}_{23} + \hat{p}_{21}\,\hat{q}_{22}\,\hat{p}_{23}$$

$$(0.5000x0.4444x0.4222x0.5000x0.4444x0.5778)$$
$$= 0.128387 + 0.128387 + 0.5000x0.4444x0.5778)$$
$$= 0.380157$$

$$\hat{p}(2/1) = p\left[x = (1,0,0),(0,1,0),(1,1,0),(0,1,1),(1,1,0)/\hat{p}_1\right] = (0.4598,0.4253,0.6897)$$

$$= \hat{p}_{11}\,\hat{q}_{12}\,\hat{q}_{13} + \hat{q}_{11},\hat{p}_{12}\,\hat{q}_{13} + \hat{p}_{11}\,\hat{p}_{12}\,\hat{q}_{13} + \hat{q}_{11}\,\hat{p}_{12}\,\hat{p}_{13} + \hat{p}_{11}\,\hat{p}_{12}\,p_{13}$$

$$= 0.4598x0.5747x0.3103 + 0.5402x0.4253x0.3103 + 0.4598x0.4253$$
$$x0.3103 + 0.5402x0.4253x0.6897 + 0.4598x0.4253x0.6897$$

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$$0.0819958 + 0.0712905 + 0.0606800 + 0.1584565 + 0.13487286 = 0.50729566$$
$$= 0.50730$$

$$\hat{p}(mc) = \frac{174}{264}(0.50730) + \frac{90}{264}(0.38016)$$

$$= 0.33436 + 0.12960$$
$$= 0.46396$$

Suppose we took the first 50 patients from each group and computed a classification rule. The frequency distribution is shown below.

State (x_1,x_2,x_3)	**Survival group Frequency**	**Non survival group Frequency**
000	4	3
100	1	2
010	1	3
110	3	12
001	22	16
101	2	1
011	1	2
111	16	11
Total	**50**	**50**

The population parameters are not known, so they are estimated by their maximum likelihood estimators

$$\hat{p}_{ij} = \frac{\sum_{k=1}^{n_i} x_{ijk}}{n_i} = \frac{n_i(x_j)}{n_i} \qquad i - 1,2 \quad j = 1,2,...r$$

$$\hat{p}_1 = \left(\frac{22}{50}, \frac{21}{50}, \frac{41}{50}\right) = (0.44, 0.42, 0.82)$$

$$\hat{p}_2 = \left(\frac{26}{50}, \frac{28}{50}, \frac{30}{50}\right) = (0.52, 0.56, 0.60)$$

Using the estimates above we obtained the classification rule R_{B3}: classify the item with response pattern x into π_1 if

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$-0.3211979x_1 - 0.6592278x_2 + 1.11094213x_3 > 0.368026$

Otherwise classify to π_2.

An item with any of the response patterns are classified as follows:

Response Pattern	**Classification**
000	π_2
100	π_2
010	π_2
110	π_2
001	π_1
101	π_1
011	π_1
111	π_2

$$\hat{p}(1/2) = p[x = (001),(101),(011) / \hat{p}_2 = (0.52,0.56,0.60)]$$

$$= \hat{q}_{11}\hat{q}_{22}\hat{p}_{23} + \hat{p}_{21},\hat{q}_{22}\hat{p}_{23} + \hat{q}_{21}\hat{p}_{22}\hat{p}_{23}$$

$$= (0.48x0.44x0.60 + 0.52x0.44x0.60 + 0.48x0.56x0.60)$$
$$= 0.12672 + 0.13728 + 0.16128$$
$$= 0.42528$$

$$\hat{p}(2/1) = p[x = (000),(100),(010),(110),(111) / \hat{p}_1 = (0.44,0.42,0.82)]$$

$$= \hat{q}_{11}\hat{q}_{12}\hat{q}_{13} + \hat{p}_{11}\hat{q}_{12}\hat{q}_{13} + \hat{q}_{11}\hat{p}_{12}\hat{q}_{13} + \hat{p}_{11}\hat{p}_{12}\hat{q}_{13} + \hat{p}_{11}\hat{p}_{12}p_{13}$$

$$= 0.56x0.58x0.18 + 0.44x0.58x0.18 + 0.56x0.42x0.18$$
$$+ 0.44x0.42x0.18 + 0.44x0.42x0.82$$
$$= 0.058464 + 0.045936 + 0.042336 + 0.033264 + 0.151536$$
$$= 0.33154$$

The plug-in estimate of the error rate is given by

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$$T\left(\hat{R}, f\right) = q_1 \sum_{R_2} \hat{f}_1(x) + q_2 \sum_{R_1} \hat{f}_2(x)$$
$$= \frac{1}{2}(0.33154 + 0.42528)$$
$$= 0.37841$$

With $n_1 = n_2 = 70$ being the seventy patients in the sample, the frequency distribution is shown below:

Simulation Experiments and Results

The eight classification procedures are evaluated at each of the 118 configurations of n, r and d. The 118 configurations of n, r and d are all possible combinations of n =20, 40, 60, 80, 100, 200, 300, 400, 600, 700, 800, 900, 1000, r =3, 4, 5 and d = 0.1, 0.2, 0.3, and 0.4. A simulation experiment which generates the data and evaluates the procedures is now described.

(i) A training data set of size n is generated via R-program where $n_1 = n/2$ observations are sampled from π_1, which has multivariate Bernoulli distribution with input parameter p_1 and $n_2 = n/2$ observations sampled from π_2 which is multivariate Bernoulli with input parameter $p_2, j = 1...r$. These samples are used to construct the rule for each procedure and estimate the probability of misclassification for each procedure is obtained by the plug-in rule or the confusion matrix in the sense of the full multinomial.

(ii) The likelihood ratios are used to define classification rules. The plug-in estimates of error rates are determined for each of the classification rules.

(iii) Step (i) and (ii) are repeated 1000 times and the mean plug-in error and variances for the 1000 trials are recorded. The method of estimation used here is called the resubstitution method.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

The following table contains a display of one of the results obtained

Table 4.1(a) Effect of input parameters P_1 and P_2 on classification rules at various values of sample size and Replications (mean apparent error rates)

P_1 = (.3, .3, .3) P_2 = (.5, .5, .5)

Sample size	Optimal	Full M.	PR	LIK	DG	NN	LD	ML
40	0.33634	0.31571	0.31557	0.31846	0.31889	0.45150	0.33953	0.33616
60	0.34554	0.33076	0.33145	0.33088	0.33043	0.42851	0.34734	0.34488
100	0.35036	0.33915	0.34231	0.34068	0.34099	0.406445	0.351545	0.34985
140	0.35211	0.34689	0.34657	0.34619	0.34659	0.39735	0.35295	0.35166
200	0.35490	0.34958	0.35148	0.35063	0.35044	0.38378	0.35578	0.35491
300	0.35621	0.35425	0.35400	0.35339	0.35334	0.37477	0.35648	0.35604
400	0.35671	0.35392	0.35371	0.35519	0.35384	0.36763	0.35694	0.35669
600	0.35740	0.35627	0.35582	0.35562	0.35597	0.36208	0.35768	0.35731
700	0.35727	0.35666	0.35611	0.35619	0.35653	0.36024	0.35738	0.35724
800	0.35772	0.35641	0.35664	0.35669	0.35655	0.35984	0.35782	0.35767
900	0.35794	0.35735	0.35668	0.35709	0.35741	0.35938	0.35811	0.35789
1000	0.35749	0.35697	0.35676	0.35623	0.35678	0.35810	0.35748	0.35744

p(mc) = 0.358

Table 4.1(b) Effect of input parameters P_1 and P_2 on classification rules at various values of sample size and Replications (actual error rates)

P_1 = (.3, .3, .3) P_2 = (.5, .5, .5) $\left| p(mc) - \hat{p}(mc) \right|$

Sample size	Optimal	Full M.	PR	LIK	DG	NN	LD	ML
40	0.04817	0.05434	0.05273	0.05440	0.05485	0.09480	0.04811	0.04735
60	0.03955	0.04613	0.04508	0.04463	0.04598	0.08479	0.03950	0.039434
100	0.03060	0.03560	0.03680	0.03631	0.03714	0.07205	0.030149	0.03063
140	0.02656	0.03346	0.03158	0.03318	0.03379	0.06370	0.02655	0.02652
200	0.02225	0.02702	0.02821	0.02780	0.02871	0.05086	0.02241	0.02243
300	0.01826	0.02429	0.02352	0.02491	0.02443	0.04129	0.01825	0.01825

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

400	0.01655	0.01955	0.01956	0.02064	0.01999	0.03257	0.01661	0
600	0.012709	0.01763	0.01688	0.01759	0.01613	0.02266	0.01283	0.01660
700	0.01178	0.01470	0.01424	0.01398	0.01426	0.01824	0.01182	0.01277
800	0.01136	0.01511	0.01486	0.01575	0.01464	0.01649	0.011396	0.01174
900	0.01058	0.01346	0.01307	0.01405	0.01381	0.01385	0.01062	0.01135
1000	0.01106	0.014018	0.01419	0.01294	0.01330	0.01243	0.01102	0.01059
								0.01103

From table 4.1(a) and (b), optimal classification rule ranks best followed by linear discriminant analysis, maximum likelihood (ML), Nearest neighbour (NN), likelihood ratio (LIK), Dillon-Goldstein (DG), Full multinomial and predictive rule. The apparent error rate for the Nearest neighbour (NN) is larger than the other rules. The order of performance is as follows:

Classification Rule	**Performance**
Optimal (OP)	1
Linear Discriminant Analysis (LDA)	2
Maximum Likelihood (ML)	3
Nearest Neighbour (NN)	4
Likelihood Ratio (LIK)	5
Dillon-Goldstein (DG)	6
Full Multinomial (FM)	7
Predictive Rule (PR)	8

DISCUSSION OF RESULTS

This study attempted to gain insight into the performance of some discrete classification techniques applied to binary data pertaining to Bernoulli multivariate distribution. The data in question was of binary nature and had a good number of sparse states. The eight classification procedures assuming multinomial structures are expected to capture more of the available information than the classical procedures. The results of Glick (1972, 1973) who considered the general problem of sample based classification induced through density estimates showed that the error rates for multinomial rules converged exponentially to the optimal Baye's error rate, favours the use of multinomial procedures over other methods. The result of this work is

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

in line with the above submissions. This favoured multinomial approach, also has the advantages of being easy to understand and implement, asymptotically optimal and yield the unique minimum variance unbiased estimates of state probabilities. In addition, in line with previous results, increased sample size, effect size and discrepancy in the ratio of sample sizes all lead to increases in overall classification accuracy. In comparing accuracy among the classification rules, optimal, LDA and maximum likelihood displayed higher classification accuracy. The results from the present studies provide strong implications for the practical use of classification rules. Research into misclassification analysis is very important for understanding when to use which classification method, and the implications of using one method over another. A better understanding of these concepts will eventually lead to better accuracy in classification and better accuracy for classification should be a goal for all areas of research that use classification methods. At each of the configurations, the classification rule that has minimum variance is declared the "best"

CONCLUSION

We have observed several marginal trends. We have observed the good performance of the optimal classification rule, the linear discriminant analysis and maximum likelihood criterion rule. The full multinomial, the likelihood and nearest neighbour rule were the worst when one considers their plug-in estimates of error rates from the exact error rates. From the analysis so far carried out, the procedures can be ranked as follows: optimal, linear discriminant, maximum likelihood rule, predictive rule, Dillon Goldstein rule, full multinomial, likelihood and nearest neighbour rule. Secondly, we concluded that it is better to increase the number of variables because accuracy increases with increasing number of variables.

REFERENCES

Adebanji, A.O., Adeyemi, S. & Iyaniwura, J.O., (2008). Effect of unequal sample size Ratio on the performance of the Linear Dischminant Function: *International Journal of Modern Mathematics 2(1),* 97-108.

Anderson, T.W. (1981). An Introduction to Multivariate Statistical Analysis, 2nd Edition, Wiley, New York.

Anderson, T.W. (1984). An Introduction to Multivariate statistical methods 2 edition. John Willey, New York.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Dillon, W.R., & Goldstain, M. (1978). On the performance of some multinomial classification rules, *Journal of American Statistical Association*, 73, No 362, pp 305-320.

Hen,J.E. & Kamber, M. (2001). Data Mining: Concepts and Techniques. Academic press, San Diego, CA.

Hills, M.(1967). "Discrimination and allocation with discrete data", *Applied Statistics*, 16, 237-250.

Oludare, S. (2011). Robust Linear classifier for equal Cost Ratios of misclassification, *CBN Journal of Applied Statistics,* (2)(1).

Onyeagu, S.1. (2003). Derivation of an optimal classification rule for discrete variables. *Journal of Nigerian Statistical Association,*

Onyeagu, S.1., Osuji, G.A., Ekezie, D.D., & Ogbonna C.J. (2013). A Review of Classification Models Using Discrete Variables. *Research Journal of Mathematical and Statistical Sciences*, vol 1(8), 28-38.

Pires, A.M. & Bronco, J.A. (2004). Comparison of vol 4, 79-80multinomial classification rules. *Journal of the American Statistics Association,* 73, 305-313.

Smith, C.A. (1947). The robust estimation of classification error rates; some examples of discrimination. *Annals of Eugenics*, 18, 272-282.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Annals of Mathematical Statistics,* 15, 145-162.

Wald, A. (1949). Statistical decision functions. *Annals of Mathematical Statistics,* 20, 165-205

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)