A COMPARISON OF MULTIVARIATE DISCRIMINATION OF BINARY DATA

[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: *The use of classification rules for binary variables are discussed and evaluated. R-software procedures for discriminant analysis are introduced and analyzed for their use with discrete data. Methods based on the full multinomial, optimal, maximum likelihood rule and nearest neighbour procedures are treated. The results obtained ranked the procedures as follows: optimal, maximum likelihood, full multinomial and nearest neighbour rule.*

Keywords: classification rules, optimal, maximum likelihood, full multinomial, nearest neighbour, binary data.

INTRODUCTION

Data often arise in the real world involving many objects with a number of measurements taken from them. These measurements may be quantitative (continuous or discrete) or qualitative (ordered or unordered categories). The latter may, in some cases be defined by only two categories and are then binary variables. Many important outcomes are binary such as program receipt, labour market status and educational attainment. These outcomes are frequently misclassified in data sets for reasons such as misreporting in surveys, the need to use a proxy variable or imperfectly linked data. A binary variable suffers from misclassification if some zeros are incorrectly recorded as ones and vice versa, which can arise from various causes. Binary classification is the task of classifying the elements of a given set into two groups on the basis of a classification rule. Some typical binary classification tasks are;

(i) Medical testing to determine if a patient has certain disease or not (the classification property is the presence of the disease).

(ii) Quality control in factories i.e.. deciding if a new product is good enough to be sold, or if it should be discarded (the classification property is being good enough).

The problem of discriminating between two populations characterized by multinomial distribution is receiving extensive coverage in the statistical literature. One reason for the rebirth of interest in the area is the frequent use of discriminant analysis in the social and behavioural sciences where data are often not of interval or ratio scale. In studies involving questionnaire data, demographic variables (more often than not measured by a two, three or four point scale) are utilized to discriminate between two or more groups. In such cases, it is more natural to assume underlying multinomial structures and proceed with classification procedures based on such characterizations, than to elect as in most frequently done, some variant of Fisher's linear discriminant function.

Several authors have studied multinomial classification in varying degrees of generality and with varying orientations. Included in this list are Cochran and Hopkins (1961), Hills (1967), Gilbert (1968), Glick (1972), Moore (1973), Goldstein and Rabiowitz (1975), Krzanowski (1975), Ott and Kronmal (1976), Goldstein and Wolf (1977) and Onyeagu and Osuji (2013). The present study is in line with the work of Gilbert and Moore in that it attempts to assess the performance of various procedures through Monte Carlo sampling experiments under different population structures. 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.

2. The Optimal Classification Rule

Independent Random Variables:

Let π_1 and π_2 be any two multivariate Bernoulli populations. According to Onyeagu (2003), Let $c(i/j)$ be the cost of misclassifying an item with measurement $\underline{x}$ from π_j into π_i and let q_j be the prior probability on π_i, where $i = 1,2$ with $q_1 + q_2 = 1$ and probability mass Function $f_i(x)$ in π_i where $i = 1,2$. Suppose that we assign an item with measurement vector x to π_1 if it is in some region $R_1 \subseteq R^r$ and to π_2 if $\underline{x}$ is in some region $R_2 \subseteq R^r$ where $R^r = R_1 \cup R_2$ and $R_1 \cap R_2 = 0$. The expected cost of misclassification is given by:

$$ECM = c(2/1)q_1 \sum_{R_2} f(x/\pi_1) + c(1/2)q_2 \sum_{R_1} f(x/\pi_2) \qquad 2.1$$

where $\sum_{R_2} f(x/\pi_1)$ =p(classifying into π_2/π_1) =p(2/1).

The optimal rule is the one that partitions R^r such that

$ECM = \sum_{R_1} f(x/\pi_2) =$ p(classifying into π_1/π_2) =p(1/2) is a minimum.

$$ECM = c(2/1)q_1 \left[1 - \sum_{R_2} f(x/\pi_1)\right] + c(1/2)q_2 \sum_{R_1} f(x/\pi_2) \qquad 2.2$$

$$= c(2/1)q_1 + \sum_{R_1}\left[c(1/2)q_2 f(x/\pi_2) - c(2/1)q_1 f(x/\pi_1)\right] \qquad 2.3$$

ECM is minimized if the second term is minimized. ECM is minimized if R_1 is chosen such that

$$c(1/2)q_2 f(x/\pi_2) - c(2/1)q_1 f(x/\pi_1) \leq 0 \qquad 2.4$$

$$c(2/1)q_1 f(x/\pi_1) \geq c(2/1)q_2 f(x/\pi_2) \qquad 2.5$$

$$R_1 = \left[x / \frac{f(x/\pi_1)}{f(x/\pi_2)} \geq \frac{c(1/2)q_2}{c(2/1)q_1} \right] \qquad 2.6$$

Therefore the optimal classification rule with respect to minimization of the expected cost of misclassification (ECM) is given by classify object with measurement x_0 into π_1 if

$$\frac{f_1}{f_2} \geq \frac{q_2 c(1/2)}{q_1 c(2/1)} \qquad 2.7$$

Otherwise classify into π_2.

Without loss of generality, we assume that $q_1 = q_2 = 1/2$ and c(1/2) = c(2/1). Then the minimization of the ECM becomes the minimization of the probability of misclassification, p(mc) under these assumptions, the optimal rule reduces to classifying an item with measurement x_0 into π_1 if

$$R_{opt}:\frac{f_1(x_0/\pi_1)}{f_2(x_0/\pi_2)}\geq 1 \qquad 2.8$$

Otherwise classify the item 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 $\underline{x}$ into π_1 if

$$\frac{\underset{j=1}{\overset{r}{\pi}}\left[p_{1j}^{x_j}(1-p_{1j})^{1-x_j}\right]}{\underset{j=1}{\overset{r}{\pi}}\left[p_{2j}^{x_j}(1-p_{2j})^{1-x_j}\right]}>1 \qquad 2.9$$

Otherwise, classify the item into π_2. This rule simplifies to:

Classify an item with response pattern $\underline{x}$ into π_1 if

$$\sum x_j In\left(\frac{p_{ij}}{q_{ij}}\cdot\frac{q_{2j}}{p_{2j}}\right)>\sum_{j=1}^{r} In\frac{q_{2j}}{q_{1j}} \qquad 2.10$$

Otherwise, classify into π_2.

2.1 The Optimal Rule for a case of two variables

Suppose we have only two independent Bernoulli variables, x_1,x_2. Then the rule becomes: classify an item with response pattern $\underline{x}$ into π_1 if:

$$R_{B_2}:In\left[\frac{p_{11}q_{21}}{q_{11}p_{21}}\right]x_1+In\left[\frac{p_{12}q_{22}}{q_{12}p_{22}}\right]x_2>In\frac{q_{21}}{q_{11}}+In\frac{q_{22}}{q_{12}} \qquad 2.1.1$$

Otherwise, classify the item into π_2. Written in another form the rule simplifies to: classify an item with response pattern $\underline{x}$ into π_1 if:

$$R_{B_2}:w_1x_1+w_2x_2>c \qquad 2.1.2$$

Otherwise, classify the item into π_2 where

$$w_1 = In\left[\frac{p_{11}}{1-p_{11}} - \frac{1-p_{21}}{p_{21}}\right] = In\frac{p_{11}}{1-p_{11}} - In\frac{p_{21}}{1-p_{21}} \qquad 2.1.3$$

$$w_2 = In\frac{p_{12}}{1-p_{12}} - In\frac{p_{22}}{1-p_{22}} \qquad 2.1.4$$

$$c = In\left[(1-p_{21})(1-p_{22})\right] - In\left[(1-p_{11})(1-p_{12})\right] \qquad 2.1.5$$

To find the distribution of z we note that

$$p[x_j = x_j / \pi_i] = \begin{cases} p_{ij}^{x_j}(1-p_{ij}) \\ 0, otherwise, i=1,2, j=1,2 \end{cases} \qquad 2.1.6$$

Since $z = \sum_{j=1}^{2} w_j x_j = w_1x_1 + w_2x_2$ 2.1.7

2.2 Optimal rule for a case of three variables.

Suppose we have three independent variables according to Onyeagu (2003), the rule is: classify an item with response pattern $\underline{x}$ into π_1 if:

$$R_{B_3} : In\left(\frac{p_{11}q_{21}}{q_{11}p_{21}}\right)x_1 + In\left(\frac{p_{12}}{q_{12}}\cdot\frac{q_{22}}{p_{22}}\right)x_2 + In\left(\frac{p_{13}}{q_{13}}\cdot\frac{q_{23}}{p_{23}}\right)x_3 > In\left(\frac{q_{21}q_{22}q_{23}}{q_{11}q_{12}q_{13}}\right) \qquad 2.2.1$$

otherwise, classify the item into π_2. Written in another form the rule simplifies to: classify an item with response pattern $\underline{x}$ into π_1 if: $R_{B_3} : w_1x_1 + w_2x_2 + w_3x_3 > c$ 2.2.2

otherwise classify the item into π_2.

$$w_1 = In\left(\frac{p_{11}}{q_{11}}\cdot\frac{q_{21}}{p_{21}}\right), w_2 = In\left(\frac{p_{12}}{q_{12}}\cdot\frac{q_{22}}{p_{22}}\right) \qquad 2.2.3$$

$$w_3 = In\left(\frac{p_{13}}{q_{13}}\cdot\frac{q_{23}}{p_{23}}\right), c = In\left(\frac{q_{21}q_{22}q_{23}}{q_{11}q_{12}q_{13}}\right) \qquad 2.2.4$$

2.3 Optimal rules for a case of four variables

Suppose we have four independent Bernoulli variables, the rule is classify an item with response pattern $\underline{x}$ into π_1 if

$$R_{B_4}: In\left(\frac{p_{11}}{q_{11}}\cdot\frac{q_{21}}{p_{21}}\right)x_1 + In\left(\frac{p_{12}}{q_{12}}\cdot\frac{q_{22}}{p_{22}}\right)x_2 + In\left(\frac{p_{13}}{q_{13}}\cdot\frac{q_{23}}{p_{23}}\right)x_3$$
$$+ In\left(\frac{p_{14}}{q_{14}}\cdot\frac{q_{24}}{p_{24}}\right)x_4 > In\frac{q_{21}}{q_{11}} + In\frac{q_{22}}{q_{12}} + In\frac{q_{23}}{q_{13}} + In\frac{q_{24}}{q_{14}} \qquad 2.3.1$$

Otherwise, classify the item into π_2. Written in another form, the rule simplifies to: classify an item with response pattern $\underline{x}$ into π_1 if:

$R_{B4} = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 > c_1 + c_2 + c_3 + c_4$ otherwise, classify the item into π_2.

2.3.2

For the case of four variables, let

$$w_1 = In\left(\frac{p_{11}}{q_{11}}\cdot\frac{q_{21}}{p_{21}}\right), w_2 = In\left(\frac{p_{12}}{q_{12}}\cdot\frac{q_{22}}{p_{22}}\right), w_3 = In\left(\frac{p_{13}}{q_{13}}\cdot\frac{q_{23}}{p_{23}}\right),$$
$$w_4 = In\left(\frac{p_{14}}{q_{14}}\cdot\frac{q_{24}}{p_{24}}\right) \qquad 2.3.3$$

$$c_1 = In\frac{q_{21}}{q_{11}}, c_2 = In\frac{q_{22}}{q_{12}}, c_3 = In\frac{q_{23}}{q_{13}}, c_4 = In\frac{q_{24}}{q_{14}} \qquad 2.3.4$$

Then $z = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 = \sum_{j=1}^{4} w_j x_j$ 2.3.5

2.4 Probability of misclassification

In constructing a procedure of classification, it is desired to minimize on the average the bad effects of misclassification (Onyeagu 2003, Richard and Dean, 1988, Oludare 2011). Suppose we have an item with response pattern x from either π_1 or π_2. We think of an item as a point in a r-dimensional space. We partition the space R into two regions R₁ and R₂ which are mutually exclusive. If the item falls in R₁, we classify it as coming from π_1 and if it falls in R₂ we classify it as coming from π_2. In following a given classification procedure, the researcher can make two kinds of errors in classification. If the item is actually from π_1, the researcher can classify it as coming from π_2. Also the researcher can classify an item from π_2 as coming from π_1. We need to know the relative undesirability of these two kinds of errors in classification. Let the prior probability that an observation comes from π_j be q_1, and from π_2 be q_2. Let the probability mass function of π_1 be $f_1(x)$ and that of π_2 be $f_2(x)$. Let the regions of classifying into π_1 be R1 and into π_2 be R2. Then the probability of correctly classifying an observation that is actually from π_1 into π_1 is

$p(1/1) = \sum_{R_1} f_1(x)$ and the probability of misclassifying such an observation into π_2 is $p(2/1) = \sum_{R_2} f_1(x)$ 2.4.1

Similarly, the probability of correctly classifying an observation from π_2 into π_2 is

$p(2/2) = \sum_{R_2} f_2(x)$ and the probability of misclassifying an item from π_1 into π_2 is

$$p(1/2) = \sum_{R_1} f_2(x) \qquad 2.4.2$$

The total probability of misclassification using the rule is

$$TPMC(R) = q_1 \sum_{R_2} f_1(x) + q_2 \sum_{R_1} f_2(x) \qquad 2.4.3$$

In order to determine the performance of a classification rule R in the classification of future items, we compute the total probability of misclassification known as the error rate. Lachenbruch (1975) defined the following types of error rates.

(i). Error rate for the optimum classification rule, Ropt. When the parameters of the distributions are known, the error rate is

$TPMC(R) = q_1 \sum_{R_2} f_1(x) + q_2 \sum_{R_1} f_2(x)$ which is optimum for this distribution.

(ii) Actual error rate: The error rate for the classification rule as it will perform in future samples.

(iii) Expected actual error rate: The expected error rates for classification rules based on samples of size n_1 from π_1 and n_2 from π_2

(iv) The plug-in estimate of error rate obtained by using the estimated parameters for π_1 and π_2 .

(v) The apparent error rate: This is defined as the fraction of items in the initial sample which is misclassified by the classification rule.

	π_1	π_2	
π_1	n_{11}	n_{12}	n_1
π_2	n_{21}	n_{22}	n_2 n

The table above is called the confusion matrix and the apparent error rate is given by

$$\hat{P}(mc) = \frac{n_{12} + n_{21}}{n} \qquad 2.4.4$$

Hills (1967) called the second error rate the actual error rate and the third the expected actual error rate. Hills showed that the actual error rate is greater than the optimum error rate and it in turn, is greater than the expectation of the plug-in estimate of the error rate. Martin and Bradley (1972) proved a similar inequality. An algebraic expression for the exact bias of the apparent error rate of the sample multinomial discriminant rule was obtained by Goldstein and Wolf (1977), who tabulated it under various combinations of the sample sizes n₁ and n₂, the number of multinomial cells and the cell probabilities. Their results demonstrated that the bound described above is generally loose.

2.5 Evaluating the probability of misclassification for the optimal rule R$_{\text{opt}}$

The optimal classification rule R$_{\text{opt}}$ for $\underline{x} = (x_1, x_2 ... x_r)$ which is distributed multivariate Bernoulli is: classify an item with response pattern $\underline{x}$ into π_1 if

$$R_{opt} : \sum_{j=1}^{r} x_j In\left(\frac{p_{1j}}{q_{1j}} \cdot \frac{q_{2j}}{p_{2j}}\right) > \sum_{j=1}^{r} In \frac{q_{2j}}{q_{1j}} \qquad 2.5.1$$

Otherwise classify into π_2

We can obtain the probability of misclassification for two cases

Case I Known parameters

(a) General case where $p_1 = (p_{i1}, p_{i2} \ldots p_{ir})$ 2.5.2

(b) Special case where $p_i = (p_i, p_i \ldots p_i)$ with the assumption $p_1 < p_2$ 2.5.3

(c) Special case (b) with additional assumption that $p_1 = \theta p_2, 0 < \theta < 1$ 2.5.4

For case (1a) the optimal classification rule R_{opt} for $\underline{x} = (x_1, x_2 \ldots x_r)$ which is distributed multivariate Bernoulli is:

Classify an item with response pattern x if

$$R_{opt} : \sum x_j In\left(\frac{p_{1j}}{q_{1j}} \cdot \frac{q_{2j}}{p_{2j}}\right) > \sum_{j=1}^{r} In\frac{q_{2j}}{q_{1j}} \qquad 2.5.5$$

Otherwise classify into π_2

Case 1b: Special case where $p_i = p(p_i, \ldots p_i)$ with the assumption that $p_1 < p_2$, the optimal classification rule R_{opt} for the r-variate Bernoulli models becomes: classify an item with response pattern $\underline{x}$ into π_1 if otherwise classify into π_2. The probability of misclassification using the special case of R_{opt} is

$$R_{opt} : \sum_{j=1}^{r} x_j \le \frac{rIn\left(\frac{q_2}{q_1}\right)}{In\left(\frac{p_1}{q_1} \cdot \frac{q_2}{p_2}\right)} \qquad 2.5.6$$

$$p(2/1) = p\Bigg[\sum_{j=1}^{r} x_j > \frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2} \cdot \frac{q_2}{q_1}\right)}\Bigg|\pi_1\Bigg| = 1 - B_{(r,p_1)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2} \cdot \frac{q_2}{q_1}\right)}\right) \qquad 2.5.7$$

$$B_{r,p}(x) = \sum_{y=0}^{r} \binom{x}{y} p^{y} (1-p)^{r-y} \qquad 2.5.8$$

$$p(1/2) = p\left[\sum_{j=1}^{r} x_j < \frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2}\cdot\frac{q_2}{q_1}\right)}\middle|\pi_2\right| = B_{(r,p_2)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2}\cdot\frac{q_2}{q_1}\right)}\right) \qquad 2.5.9$$

$$p(mc) = \frac{1}{2}\left[1 + B_{(r,p_2)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2}\cdot\frac{q_2}{q_1}\right)}\right) - B_{(r,p_2)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2}\cdot\frac{q_2}{q_1}\right)}\right)\right] \qquad 2.5.10$$

Case 1c: Special case (1b) with additional assumption that $p_1 = \theta p_2$ and $q_1 = 1 - p_1 = 1 - \theta p_2$ and $q_2 = 1 - p_2$. The optimal classification rule R_{opt} for $\underline{x} = (x_1, x_2 ... x_r)$ distributed multivariate Bernoulli is: classify the item with response pattern x into π_1 if

$$R_{opt} : \sum_{j=1}^{r} x_j > \left[\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\theta\left(\frac{1-p_2}{1-\theta P_2}\right)}\right] \qquad 2.5.11$$

and to π_2 otherwise.

The probability of misclassification using the special case of R_{opt} when $p_1 = \theta p_2$ is

$$p(2/1) = 1 - B_{(r,\theta p_2)}\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\theta\left(\frac{1-p_2}{1-\theta p_2}\right)} \qquad 2.5.12$$

$$p(1/2)B_{(r,p_2)}\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\theta\left(\frac{1-p_2}{1-\theta p_2}\right)}$$

$$p(mc) = \frac{1}{2}\left[1 + B_{(r,p_2)}\left(\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\theta\left(\frac{1-p_2}{1-\theta p_2}\right)}\right)\right] - B_{r,\theta p_2}\left[\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\theta\left(\frac{1-p_2}{1-\theta p_2}\right)}\right] \qquad 2.5.13$$

For the fixed values of r and different values of p_1 and p_2

Case 2:Unknown parameters

(a) General case $p_i = (p_{i1}, p_{i2} ... p_{ik})$

In order to estimate p_1 and p_2 we take training samples of size n_1 and n_2 from π_1 and π_2 respectively. In π_1 we have the sample

$$\begin{aligned}
x_{11} &= (x_{111}, x_{121}, x_{131}, \ldots x_{1k1}, \ldots x_{1r1}) \\
x_{12} &= (x_{112}, x_{122}, x_{132}, \ldots x_{1k2}, \ldots x_{1r2}) \\
&. \\
&. \qquad\qquad 2.5.14 \\
&. \\
x_{1n1} &= (x_{11n1}, x_{12n1}, x_{13n1}, \ldots x_{1kn1}, \ldots x_{1rn1})
\end{aligned}$$

The maximum likelihood estimate of p_1 is

$$\hat{p}_{1k} = \sum_{j=1}^{n_1} \frac{x_{1kj}}{n_1} \qquad 2.5.15$$

Similarly the maximum likelihood of estimate of p_2 is

$$\hat{p}_{2k} = \sum_{j=1}^{n_2} \frac{x_{2kj}}{n_2} \qquad 2.5.16$$

We plug in this estimate into the rule for the general case in 1(a) to have the following classification rule: classify an item with response pattern x into π_1 if

$$R_{Br} : \sum_{j=1}^{r} x_j In\left(\frac{\hat{p}_{ij}}{\hat{q}_{ij}} \cdot \frac{\hat{q}_{2j}}{\hat{p}_{2j}} \right) > rIn \frac{\hat{q}_{2j}}{\hat{q}_{ij}} \qquad 2.5.17$$

otherwise classify into π_2

(b) Special case of 1b where $p_i = (p_i, p_i \ldots p_i)$ with the assumption that $p_{1i} < p_{2i}$

In this special case

$$\hat{p}_1 = \sum_{j=1}^{n_1} \frac{x_{11j}}{n_1} = \sum_{j=1}^{n_1} \frac{x_{12j}}{n_1} \ldots \sum_{j=1}^{n_1} \frac{x_{1kj}}{n_1} = \ldots = \sum_{j=1}^{n_1} \frac{x_{1rj}}{n_1} \qquad 2.5.18$$

$\sum_{j=1}^{r} x_{ij}$ is distributed $B(r, p_i)$

$\sum_{k=1}^{n_1}\sum_{j=1}^{r} x_{1jk}$ is distributed $B(rn_1, p_1)$

The maximum likelihood estimate of p_1 is

$$\hat{p}_1 = \frac{\sum_{k=1}^{r}\sum_{j=1}^{n_1} x_{1jk}}{rn_1} \qquad 2.5.19$$

Likewise, the maximum likelihood estimate of p_2 is

$$\hat{p}_2 = \frac{\sum_{k=1}^{r}\sum_{j=1}^{n_2} x_{2jk}}{rn_2} \qquad 2.5.20$$

We plug in these two estimates into the equation for the special case (1b) to have the following classification rule: classify the item with response pattern x into π_1 if

$$\sum_{j=1}^{r} x_j \le \frac{rIn\left(\frac{\hat{q}_2}{\hat{q}_1}\right)}{In\left(\frac{\hat{p}_1}{\hat{q}_1}\cdot\frac{\hat{q}_2}{\hat{p}_2}\right)} \qquad 2.5.21$$

Otherwise classify into π_2

The probability of misclassification is given by

$$\hat{p}(mc) = \frac{1}{2}\left[1 + B_{(r,p_2)}\left(\frac{rIn\frac{\hat{q}_2}{q_1}}{In\left(\frac{\hat{p}_1}{\hat{p}_2}\cdot\frac{\hat{q}_2}{\hat{q}_1}\right)}\right) - B_{(r,\hat{p}_1)}\frac{rIn\left(\frac{\hat{q}_2}{\hat{q}_1}\right)}{In\left(\frac{\hat{p}_1}{\hat{p}_2}\cdot\frac{\hat{q}_2}{\hat{q}_1}\right)}\right] \qquad 2.5.22$$

$$\hat{p}(mc)=\frac{1}{2}\left[1+B(r,\hat{p}_{2},\lambda)-B(r,\hat{p}_{1},\lambda)\right]$$

Where $$\lambda=\frac{rIn\left(\frac{\hat{q}_2}{\hat{q}_1}\right)}{In\left(\frac{\hat{p}_1}{\hat{p}_2}\cdot\frac{\hat{q}_2}{\hat{q}_1}\right)}$$

$$B(k,\alpha,x)=\sum_{y=0}^{k}\binom{k}{y}\alpha^{y}(1-\alpha)^{k-y} \qquad 2.5.23$$

(c) Special case of 2b with $p_1=\theta p_2,\ p_1<p_2, 0<\theta<1$ we take training samples of size n_2 from π_2 and estimate p_2 by

$$\hat{p}_{2}=\sum_{k=1}^{n}\sum_{j=1}^{r}\frac{x_{2jk}}{rn_2} \qquad 2.5.24$$

For a fixed value of $\theta, \hat{p}_1=\theta\,\hat{p}_2$

The classification rule is: classify the item with response pattern x into π_1 if

$$R_{B_r}:\sum_{j=1}^{r}x_j>\frac{rIn\left[\frac{1-\hat{p}_2}{1-\theta\,\hat{p}_2}\right]}{In\theta\left[\frac{1-\hat{p}_2}{1-\theta\,\hat{p}_2}\right]} \qquad 2.5.25$$

otherwise classify into π_2 .

The probability of misclassification is given by

$$\hat{p}(mc)=\frac{1}{2}\Bigg[1+B_{(r,p_2)}\frac{rIn\left[\frac{(1-\hat{p}_2)}{(1-\theta\,\hat{p}_2)}\right]}{In\theta\left[\frac{(1-\hat{p}_2)}{(1-\theta\,\hat{p}_2)}\right]}\Bigg]-B_{r,\theta p_2}\Bigg[\frac{rIn\left[\frac{(1-\hat{p}_2)}{(1-\theta\,\hat{p}_2)}\right]}{In\theta\left[\frac{(1-\hat{p}_2)}{(1-\theta\,\hat{p}_2)}\right]}\Bigg]\Bigg] \qquad 2.5.26$$

$$\hat{p}(mc) = \frac{1}{2}\left[1 + B(r, p_2, \lambda) - B(r, \theta \hat{p}_2, \lambda)\right]$$

$$\lambda = \frac{rIn\left(\frac{1-\hat{p}_2)}{1-\theta \hat{p}_2}\right)}{In\theta\left(\frac{1-\hat{p}_2}{1-\theta \hat{p}_2}\right)} \qquad 2.5.27$$

3. Maximum Likelihood Rule (ML-Rule)

The maximum likelihood discriminant rule for allocating an observation x to one of the populations $\pi_1,..\pi_n$ is to allocate x to the population which gives the largest likelihood to x.

Classify in π_1 if $p(w_1/x) > p(w_2/x)$ or to π_2 if $p(w_1/x) < p(w_2/x)$ 3.1.

where $p(w_1/x)$ is the posterior probability which can be found by the Bayes Rule. But this is the same as: classify to π_1 if $\frac{p(x/w_1)\cdot p(w_1)}{p(x)} > \frac{p(x/w_2)\cdot p(w_2)}{p(x)}$ 3.2

where $p(x/w_i)$ is the class conditional probability density function and $p(w_i)$ is the prior probability. By denoting the classes as π_1, $\pi_2 \ldots \pi_n$, the maximum likelihood classifier is based on the assumed multivariate normal probability density function for each class given by

$$f(x/\pi_i) = \frac{1}{(2\pi)^{p/2}\left|\hat{\Sigma}_i\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left((x-\hat{\mu}_i)\right)^T \hat{\Sigma}_i^{-1}\left((x-\hat{\mu}_i)\right)} \qquad 3.3$$

where $\hat{\mu}_i$ is the estimated mean vector for class i and $\hat{\Sigma}_i$ is the estimated variance covariance matrix for class π_i and p is the number of characteristics measured (ie the length of each vector x into one of the classes, recall that the density function $f(x/\pi_i)$ is evaluated for each of the k classes and the x is assigned to π_i if (assuming equal costs of misclassification and equal a prior probabilities) one has

$f(x/\pi_i) > f(x/\pi_j)$ for all $j \neq i$ 3.4

We assumed that the data can be modeled adequately by a multi-normal distribution. If the class-conditional probability density function $p(x/w_i)$ is estimated by using the frequency of occurrence of the measurement vectors in the training data, the resulting classifier is non-parametric. An important advantage of the non-parametric classifier is that any pattern, however irregular it may be, can be characterized exactly. This advantage is generally outweighed by two difficulties with the non-parametric approach.

(i) It is difficult to obtain a large enough training sample to adequately characterize the probability distribution of a multi-band data set.

(ii) Specification of a meaningful n-dimensional probability density function requires a massive amount of memory or very clever programming.

4. The Full Multinomial Rule

Suppose we have discrete random variables x_1, $x_2 \ldots x_r$, each assuming values 0 or 1. The joint probability mass function pmf according to Hand (1983) is given by:

$$p(x_1, x_2 \ldots x_r) = \frac{n!}{x_1!x_2!\ldots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r} \qquad 4.1$$

for $x_i = 0,1\ldots n$ for each i but with the x_i subject to the restriction $\sum_{i=1}^{r} x_i = n$

The range space R_x of x consists of vectors $\underline{x} = (x_1, x_2, \ldots x_r)$, where $x_i \in R_{x_i}, i = 1,\ldots r$ and for our purposes is assumed that it is generated by an r-random vector whose argument is 0 or 1. Suppose that two groups, π_1 and π_2, are large populations having prior probabilities q_1 and q_2, where $q_1 + q_2 = 1$ under the full multinomial, the probability mass function denoted by $f_i(x)$ with minimum variance unbiased estimators is

$$f_i(x) = \frac{n_i(x)}{n_i}, i = 1,2 \qquad 4.2$$

where $n_i(x)$ is the number of the individuals in a sample of size n_i from the ith population having response pattern x. The classification rule is: classify an item with response pattern x into π_1 if :

$$q_1 \frac{n_1(x)}{n_1} > q_2 \frac{n_2(x)}{n_2} \qquad 4.3$$

and to π_2 if $q_1 \frac{n_1(x)}{n_1} < q_2 \frac{n_2(x)}{n_2}$ 4.4

and with probability $\frac{1}{2}$ if : $q_1 \frac{n_1(x)}{n_1} = q_2 \frac{n_2(x)}{n_2}$ 4.5

An intuitive estimate based on D of the optimal error is the apparent error simply defined as the proportion of errors made by the rule. The apparent error of the sample based full multinomial classification rule assumes the form:

$$\hat{D} = \frac{\sum_x \min(n_1(x), n_2(x))}{n_1 + n_2} \qquad 4.6$$

where $n_1(x), n_2(x)$ are the number of sample values from population π_1 and π_2 respectively.

The advantages of using the full multinomial rule according to Onyeagu (2003) are as follows:

It is extremely simple to apply. Secondly, the computation of apparent error does not require rigorous computational formula.

The disadvantages are as follows:

There are however, certain observations that are apparent and point to potential difficulties in applying the so-called full-multinomial rule. Perhaps, the most prominent is the problem of state proliferation made especially troublesome in practice by the availability of relatively small sample sizes, r variables each assuming only k distinct values, generate k^r states. Obviously, a large number of observations related to the number of variables is required if sufficient data in each state are to be available for estimation of state probabilities.

A part from the problem of zeros in states or the potential of far observations on which to base the estimation of state probabilities in the issue that for a given state a zero from π_1 might mean something entirely different from a zero coming from π_2 if the optimal procedure is used. Moreover, especially if samples of disproportionate sizes are available, error rates generated by the somewhat forced allocation caused by a zero in, say, state j from π_2 can be potentially misleading. It is because of these difficulties that some researchers have been reluctant to apply the full multinomial procedures in situations where the data differs from severe sparseness.

5. The Nearest Neighbour Procedure

The kth nearest neighbour method (K-NN) is another tool that is used whenever the class density functions, $f_i(x)$ are known. In fact, this was the first non-parametric method for classification and was introduced by Fix and Hodges (1951). The idea behind the method is relatively simple. Clark (1978) define a random observation X_m, $x_m \in \{x_1,...x_n\}$ as the nearest neighbour to x if: Min $d(x_j,x)=d(x_m,x)$, j=1,2 …n 5.1 where $d(x_j,x)$ is a distance function. The nearest neighbour rule decides that x belongs to the class of its neighbour X_m. The above is the single nearest neighbour rule, that is k = I, and only applies to the single nearest neighbour to x. All other observations are ignored. The idea is extended naturally to the k-nearest neighbours of x. Lachenbruch (1975) describes the general K-NN rules as follows: Suppose there are n_1 and n_2 sample observations from π_1 and π_2 respectively. Suppose that the objective is to classify an observation x to one of π_1 or π_2. Using a distance function, $d(x_{ij},x)$, order the values, x_{ij}. Let k_i be the number of observations from π_1 among the k closet observations to x. The rule is to assign x to π_1 if:

$$\frac{k_1}{n_1} > \frac{k_2}{n_2} \qquad 5.2$$

otherwise to π_2. In other words, the procedure involves the relatively simple concept of assigning a random observation x to the class having the greater proportion of observations closet to x. As $n_i \to \infty$ it has been found that (3.1.79) tends to the maximum likelihood rule. There are several variations of the discrete analogue to the above estimator, each with their own operational difficulties. (See Hand (1993) for details. 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 vectors y that differ from x. Specifically, let k be the number of cells in which x and y differ. Then define

$R_j = \left\{ y_j \middle| (x-y)^1 (x-y) \le k \middle| \right\}$ 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: $\sum_R \frac{n_1(y_j)}{n_1} > \sum_R \frac{n_2(y_j)}{n_2}$ 5.3

and into π_2 otherwise.

Advantages and disadvantages

However, these methods are not without their limitations and are based on some assumptions. Although nearest neighbour is distribution free and the classifier then has no explicit functional form, it is very difficult to check the assumption that the distribution is locally constant near x. Also the choice of the distance function must be taken into consideration. It must be appropriate and meaningful. For example, Euclidean distance is usually the default choice but may not be appropriate as in such cases where the variables are of very different magnitudes and must be standardized first. Also, distances in high dimensions becomes complicated and assigning one object to be nearer than other gets blurred because as p gets increasingly larger the ratio of nearest to furthest neighbours approaches 1.

6. Simulation Experiments and Results

The four 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.

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

Table 6(a) Apparent error rates for classification rules under different parameter values, sample sizes and Replications

P_1 = (.3, .3, .3, .3, .3) P_2 = (.7, .7, .7, .7, .7)

Sample sizes	Optimal	Full M.	NN	ML
40	0.157125	0.110074	0.504725	0.151137
60	0.161900	0.127855	0.504016	0.160683
100	0.163290	0.143526	0.498380	0.162775
140	0.162967	0.149837	0.501275	0.162137
200	0.162565	0.156384	0.500097	0.163167
300	0.162783	0.159788	0.498055	0.163158
400	0.404243	0.384500	0.500762	0.404750
600	0.163018	0.161992	0.499159	0.162877
700	0.163075	0.162454	0.499249	0.162631
800	0.163463	0.163084	0.496118	0.163254
900	0.163354	0.163508	0.497806	0.163643
1000	0.163273	0.162916	0.502523	0.163039

p(mc) = 0.16308

Table 6(b) Actual Error rate for the classification rules under different parameter values, sample sizes and replications.

$P_1 = (.3, .3, .3, .3, .3)$ $P_2 = (.7, .7, .7, .7, .7)$ $\left| p(mc) - \hat{p}(mc) \right|$

Sample size	Optimal	Full M.	NN	ML
40	0.040271	0.052706	0.094217	0.054898
60	0.032751	0.042691	0.089946	0.045567
100	0.027786	0.037015	0.087373	0.039006
140	0.022462	0.031623	0.083515	0.031104
200	0.017981	0.026657	0.083256	0.025252
300	0.0150903	0.020882	0.081724	0.020167
400	0.012793	0.018476	0.082592	0.017482
600	0.010874	0.014643	0.081390	0.014753
700	0.009666	0.013574	0.082388	0.013359
800	0.009308	0.012778	0.080319	0.012764
900	0.008725	0.012243	0.081073	0.012117
1000	0.010713	0.022517	0.030042	0.017158

Tables 6(a) and (b) present the mean apparent error rates and standard deviation (actual error rates) for classification rules under different parameter values. The mean apparent error rates increases with the increase in sample sizes and standard deviation decreases with the increase in sample sizes. Predictive and Dillon-Goldstein improved with the increase in the number of variables while maximum decrease in performances. From the analysis, optimal is ranked first, followed by linear discriminant analysis, predictive rule, Dillon-Goldstein, likelihood ratio, maximum likelihood, full multinomial and nearest neighbour occupied the last position as shown below.

Classification Rule	**Performance**
Optimal (OP)	1
Maximum Likelihood (ML)	2
Full Multinomial (FM)	3
Nearest Neighbour (NN)	4

Conclusion

We obtained two major results from this study. Firstly, using simulation experiments we ranked the procedures as follows: Optimal, Maximum likelihood, Full multinomial and Nearest Neighbour. The best method was the optimal classification rule. Secondly, we concluded that it is better to increase the number of variables because accuracy increases with increasing number of variables.

References

Clark, G.M. (1978).Predicting the Survival in Burned Patients Using Discriminant Analysis, *BURNS*, 4, 81-89

Cochran, W.G., & Hopkins, C.E. (1961). Some classification problems with multivariate Qualitative Data, *Biometrics 17*, 10-32.

Gilbert, S.E. (1968). “On Discrimination using Qualitative Variables” *Journal of the American Statistical Association 1399-1418.*

Glick, N. (1972). Sample based classification procedures derived from density estimators. *Journal of American Statistical Association*, 67, 116-120.

Goldstein, M. & Rabinowitz (1975). Selection of variables for the two group multinomial classification problem. *JASA 70*, 776-781

Goldstein, M. & Wolf (1977). On the problem of Bias in multinomial classification. *Biometrics 33*, 325-331.

Hand, D.J. (1993). “New instruments for identifying good and bad credit risks: a feasibility study”, *Report,* Trustee Savings Bank, London

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

Krzanowski, W.J. (1976). Principles of Multivariate Analysis: Users perspective. John Willey and sons Inc. new York..

Lachenbruch, P.A. (1975). “Discriminant Analysis, Hafrier press New York.

Martin, D.C. & Bradley, R.A. (1972). Probability Models, Estimation and Classification for Multivariate Dichotomous Populations, *Biometrics,* 23, 203-221

Moore, D.H. (1973). Evaluation of five discrimination procedures for binary variables. *Journal of the American Statistical Association 68, 399-404.*

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 vol 4,* 79-80

Onyeagu, S.I. & Osuji, G.A. (2013). Evaluation of seven classification procedures for binary variables. *Journal of the Nigerian Statistical Association vol 20. ISSN 0331-9504.*

Otti, J. & Kronmal, R.A. (1976). Some classification procedures for multivariate binary data using orthogonal functions. *Journal of American Statistical Association,* 71, 391-399.

Richard, A.J., & Dean, W.W. (1998). Applied Multivariate Statistical Analysis. 4th edition, Prentice Hall, Inc. New Jessey.