Journal of Mathematics Research; Vol. 6, No. 4; 2014
ISSN 1916-9795 E-ISSN 1916-9809
Published by Canadian Center of Science and Education

A Comparison of the Optimal Classification Rule and Maximum Likelihood Rule for Binary Variables

I. Egbo[1], S. I. Onyeagu[2], D. D. Ekezie[3] & Uzoma Peter O.[4]

[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.

[4] Department of Computer Science, Alvan Ikoku Federal College of Education, Owerri, Nigeria

Correspondence: I. Egbo, Department of Mathematics, Alvan Ikoku Federal College of Education, Owerri, Nigeria. E-mail: egboike@gmail.com

Received: September 26, 2014 Accepted: October 29, 2014 Online Published: November 12, 2014

doi:10.5539/jmr.v6n4p124 URL: http://dx.doi.org/10.5539/jmr.v6n4p124

Abstract

Optimal classification rule and maximum likelihood rules have the largest possible posterior probability of correct allocation with respect to the prior. They have a 'nice' optimal property and appropriate for the development of linear classification models. In this paper we consider the problem of choosing between the two methods and set some guidelines for proper choice. The comparison between the methods is based on several measures of predictive accuracy. The performance of the methods is studied by simulations.

Keywords: optimal classification rule, maximum likelihood rule, Binary variables.

1. Introduction

Optimal classification rules and maximum likelihood rule are widely used multivariate statistical methods for analysis of data with categorical outcome variables. Both of them are appropriate for the development of linear classification models, i.e. models associated with linear boundaries between the groups. Binary classification is the task of classifying the elements of a given set into two groups on the basis of a classification rule.

Classification is of broad interest in science because it permeates many scientific studies and also arises in the contexts of many applications (Panel on Discriminant Analysis, Classification and Clustering, 1989). Examples in the educational, social and behavioural sciences include identifying children in kindergarten at risk for future reading difficulties (Catts, Fey, Zhang and Tomblin (2001), identifying individuals at risk for addiction (Robinson, 2002) and predicting the crimes that male juvenile offenders may commit according to their personality characteristics (Glaser, Calhoun and Petrocelli, 2002). In the biological and medical sciences, application of classification procedures include identifying patients with chronic heart failure (Udris, 2010) detecting lung cancer (Philips, 2003) and determining whether certain breast masses are malignant or being (Sahiner, 2004). In the management sciences, methods for classification have been used for such purposes as predicting bankruptcy (Jo, Han and Lee, 1997); Dichotomous classification of Foreign Assisted Project implementation status (Nworuh and Anyiam, 2010).

In this paper, we shall be concerned with k=2 population classification problems. Our interest is in deriving a rule that can be used to optimally assign an item to one of the populations. The optimality criterion is to minimize the risk associated with the rule (Onyeagu & Osuji 2010). The goal of this paper is to set some guidelines as to when the choice of either one of the methods is still appropriate. While optimal is much more general and has a number of theoretical properties, maximum likelihood must be the better choice if we know the population is normally distributed. However, in practice, the assumptions are nearly always violated and we have therefore tried to check the performance of both methods with simulations. This kind of research demands a careful control, so we have decided to study just a few chosen situations, trying to find a logic in the behaviour and then think about the expansion onto more general cases. We have confined ourselves to compare only the predictive power of the methods.

Section 2 and 3 briefly describes the algorithms; section 4 describes the process of the simulations. The results

obtained are presented and discussed in section 5 and conclusions and recommendations are given in section 6.

2. The Optimal Classification Rule

Independent Random Variables:

Let π_1 and π_2 be any two multivariate Bernoulli populations. 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:

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

where $\sum_{R_2} f(x/\pi_1)$ =p(classifying into π_2/π_1) =p(2/1), where $p(2/1)=$ when π_1 observation is incorrectly classified as π_2.

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.

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

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

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

$$c(1/2)q_2 f(x/\pi_2) - c(2/1)q_1 f(x/\pi_1) \le 0 \tag{2.4}$$

$$c(2/1)q_1 f(x/\pi_1) \ge c(2/1)q_2 f(x/\pi_2) \tag{2.5}$$

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

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

$$R_1 : \frac{f_1(x)}{f_2(x)} \ge \frac{q_2 c(1/2)}{q_1 c(2/1)}\left(\frac{p_2}{p_1}\right) \qquad R_2 : \frac{f_1(x)}{f_2(x)} < \frac{q_2 c(1/2)}{q_1 c(2/1)}\left(\frac{p_2}{p_1}\right) \tag{2.7}$$

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

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

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

$$\frac{\prod_{i=1}^{r}\left[p_{1j}^{x_j}(1-p_{1j})^{1-x_j}\right]}{\prod_{i=1}^{r}\left[p_{2j}^{x_j}(1-p_{2j})^{1-x_j}\right]}>1 \tag{2.9}$$

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

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

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

Otherwise, classify into π_2.

For any rule, the average or expected cost of misclassification (ECM) is provided by the product of the off-entries by their probabilities of occurrence. A good classification rule should have an ECM as small as possible. The regions R_1 and R_2 that minimize the ECM are defined by the values x for which the inequalities are defined.

If the parameters are unknown, then they are estimated by their maximum likelihood estimators given by

$$\hat{p}_{ij}=\sum_{k=1}^{n}\frac{x_{ijk}}{n_i}=\frac{n_i(x_j)}{n_i}=\bar{x}_{ij} \tag{2.11}$$

where $n_i(x_j)=\sum_{k=1}^{n_1}x_{ijk}$ is equal to the number of observation from π_i with jth variable. The rule for unknown parameters is: classify an item with response pattern $\underline{x}$ into π_1 if

$$\sum_{j=1}^{r} In\left(\frac{\hat{p}_{1j}}{\hat{q}_{1j}}\cdot\frac{\hat{q}_{2j}}{\hat{p}_{2j}}\right)x_j>\sum_{j=1}^{r} In\frac{\hat{q}_{2j}}{\hat{q}_{1j}} \tag{2.12}$$

otherwise classify the item into π_2

2.1 The Optimal Rule for a Case of Two Variables in Two Group Classifications

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:

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

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:

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

Otherwise, classify the item into π_2 where

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

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

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

To find the distribution of z we note that

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

Since

$$z = \sum_{j=1}^{2} w_j x_j = w_1 x_1 + w_2 x_2 \tag{2.1.7}$$

The range of z is

$$R_2 = \{0, w_1, w_2, w_1 + w_2\}$$

$$p[z = 0/\pi_i] = p(x_1 = 0, x_2 = 0/\pi_i) = q_{i1}q_{i2} \tag{2.1.8}$$

$$p(z = w_1/\pi_i) = p(x_1 = 1, x_2 = 0/\pi_i) = p_{i1}q_{i2} \tag{2.1.9}$$

$$p(z = w_1 + w_2/\pi_i] = p(x_1 = 1, x_2 = 1/\pi_i) = p_{i1}p_{12}q_{i1}q_{i2} \tag{2.1.10}$$

$$\text{if } z = 0 \; p(z/\pi_i) = p_{i1}q_{i2} \text{ if } z_1 = w_1 \tag{2.1.11}$$

$$q_{i1}p_{i2} \text{ if } z = w_2 \tag{2.1.12}$$

$$p_{i1}p_{i2} \text{ if } z = w_1 + w_2, i = 1,2 \tag{2.1.13}$$

If $w_1 < w_2$ the distribution function of z is given by 0 if z =0

$$q_{i1}q_{i2} \text{ if } 0 \le z < w_1 \tag{2.1.14}$$

$$p(z/\pi_i) = q_{i1}q_{i2} + p_{i1}q_{i2} \text{ if } w_1 \le z < w_2 = q_{i1}q_{i2} + p_{i1}q_{i2} + p_{i1}q_{i2} \text{ if } w_2 \le z < w_1 + w_2 \tag{2.1.15}$$

$$1 \text{ if } w_1 + w_2 \le z \tag{2.1.16}$$

2.2 Optimal Rule for a Case of Three Variables in Two Group Classifications.

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:

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

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:

$$R_{B_3} : w_1x_1 + w_2x_2 + w_3x_3 > c \tag{2.2.2}$$

otherwise classify the item into π_2.

$$w_1 = In\left(\frac{p_{11}}{q_{11}} \cdot \frac{q_{21}}{p_{21}}\right), w_2 = In\left(\frac{p_{12}}{q_{12}} \cdot \frac{q_{22}}{p_{22}}\right) \quad w_3 = In\left(\frac{p_{13}}{q_{13}} \cdot \frac{q_{23}}{p_{23}}\right), c = In\left(\frac{q_{21}q_{22}q_{23}}{q_{11}q_{12}q_{13}}\right) \tag{2.2.3}$$

2.3 Optimal Rules for a Case of Four Variables in Two Group Classifications

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

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

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: $R_{B_4} = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 > c_1 + c_2 + c_3 + c_4$ otherwise, classify the item into π_2. For the case of four variables, let

$$w_1 = In\left(\frac{p_{11}}{q_{11}} \cdot \frac{q_{21}}{p_{21}}\right), w_2 = In\left(\frac{p_{12}}{q_{12}} \cdot \frac{q_{22}}{p_{22}}\right), w_3 = In\left(\frac{p_{13}}{q_{13}} \cdot \frac{q_{23}}{p_{23}}\right),$$
$$w_4 = In\left(\frac{p_{14}}{q_{14}} \cdot \frac{q_{24}}{p_{24}}\right) \tag{2.3.2}$$

The distribution function is derived just the same way as the case of three variables. Using the same method the probability mass function of z and the distribution function for the case of five variables could be derived.

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 R_1 and R_2 which are mutually exclusive. If the item falls in R_1, we classify it as coming from π_1 and if it falls in R_2 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_1 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 R_1 and into π_2 be R_2. 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

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

The total probability of misclassification using the rule is

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

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, R_{opt}. 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 these 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

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

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. Fukunaga and Kessel (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 n_1 and n_2, 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_{opt}

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 $\underline{x}$ into π_1 if

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

Otherwise classify into π_2

We can obtain the probability of misclassification for two cases

Case I Known parameters

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

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

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

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

$$R_{opt} : \sum x_j In\left(\frac{p_{1j}}{q_{1j}} \cdot \frac{q_{2j}}{p_{2j}}\right) > \sum_{j=1}^{r} In\frac{q_{2j}}{q_{1j}} \tag{2.5.2}$$

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

$$R_{opt} : \sum_{j=1}^{r} x_j \le \frac{rIn\left(\frac{q_2}{q_1}\right)}{In\left(\frac{p_1}{q_1} \cdot \frac{q_2}{p_2}\right)} \tag{2.5.3}$$

$$p(2/1) = p\left[\sum_{j=1}^{r} x_j > \frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2} \cdot \frac{q_2}{q_1}\right)} \middle| \pi_1\right] = 1 - B_{(r,p_1)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2} \cdot \frac{q_2}{q_1}\right)}\right) \tag{2.5.4}$$

$B_{r,p}(x) = \sum_{y=0}^{r} (_y^x) p^y (1-p)^{r-y}$, where $B_{(r,p)}$ have binomial distribution with parameters r and p (2.5.5)

$$p(1/2) = p\left[\sum_{j=1}^{r} x_j < \frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2} \cdot \frac{q_2}{q_1}\right)} \middle| \pi_2\right] = B_{(r,p_2)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2} \cdot \frac{q_2}{q_1}\right)}\right) \tag{2.5.6}$$

$$p(mc)=\frac{1}{2}\left[1+B_{(r,p_2)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2}\cdot\frac{q_2}{q_1}\right)}\right)-B_{(r,p_2)}\left(\frac{rIn\frac{q_2}{q_1}}{In\left(\frac{p_1}{p_2}\cdot\frac{q_2}{q_1}\right)}\right)\right] \tag{2.5.7}$$

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

$$R_{opt}:\sum_{j=1}^{r}x_j>\left[\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\,\theta\left(\frac{1-p_2}{1-\theta P_2}\right)}\right] \tag{2.5.8}$$

and to π_2 otherwise.

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

$$p(2/1)=1-B_{(r,\theta p_2)}\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\,\theta\left(\frac{1-p_2}{1-\theta p_2}\right)} \tag{2.5.9}$$

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

$$p(mc)=\frac{1}{2}\left[1+B_{(r,p_2)}\left(\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\,\theta\left(\frac{1-p_2}{1-\theta p_2}\right)}\right)\right]-B_{r,\theta p_2}\left[\frac{rIn\left(\frac{1-p_2}{1-\theta p_2}\right)}{In\,\theta\left(\frac{1-p_2}{1-\theta p_2}\right)}\right] \tag{2.5.10}$$

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

$$\begin{aligned}
&x_{11}=(x_{111},x_{121},x_{131},...x_{1k1},...x_{1r1})\\
&x_{12}=(x_{112},x_{122},x_{132},...x_{1k2},...x_{1r2})\\
&.\\
&.\\
&.\\
&x_{1n1}=(x_{11n1},x_{12n1},x_{13n1},...x_{1kn1},...x_{1rn1})
\end{aligned} \tag{2.5.11}$$

The maximum likelihood estimate of p_1 is

$$\hat{p}_{1k}=\sum_{j=1}^{n_1}\frac{x_{1kj}}{n_1} \tag{2.5.12}$$

Similarly the maximum likelihood of estimate of p_2 is

$$\hat{p}_{2k}=\sum_{j=1}^{n_2}\frac{x_{2kj}}{n_2} \tag{2.5.13}$$

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

$$R_{Br}:\sum_{j=1}^{r}x_j In\left(\frac{\hat{p}_{ij}}{\hat{q}_{ij}}\cdot\frac{\hat{q}_{2j}}{\hat{p}_{2j}}\right)>rIn\frac{\hat{q}_{2j}}{\hat{q}_{ij}} \tag{2.5.14}$$

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

$$\hat{p}_1 = \sum_{j=1}^{n_1} \frac{x_{11j}}{n_1} = \sum_{j=1}^{n_1} \frac{x_{12j}}{n_1} \ldots \sum_{j=1}^{n_1} \frac{x_{1kj}}{n_1} = \ldots = \sum_{j=1}^{n_1} \frac{x_{1rj}}{n_1} \tag{2.5.15}$$

$\sum_{j=1}^{r} x_{ij}$ 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

$$\hat{p}_1 = \frac{\sum_{k=1}^{r} \sum_{j=1}^{n_1} x_{1jk}}{rn_1} \tag{2.5.16}$$

Likewise, the maximum likelihood estimate of p_2 is

$$\hat{p}_2 = \frac{\sum_{k=1}^{r} \sum_{j=1}^{n_2} x_{2jk}}{rn_2} \tag{2.5.17}$$

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

$$\sum_{j=1}^{r} x_j \leq \frac{rIn\left(\frac{\hat{q}_2}{\hat{q}_1}\right)}{In\left(\frac{\hat{p}_1}{\hat{q}_1} \cdot \frac{\hat{q}_2}{\hat{p}_2}\right)} \tag{2.5.18}$$

Otherwise classify into π_2

The probability of misclassification is given by

$$\hat{p}(mc) = \frac{1}{2}\left[1 + B_{(r,p_2)}\left(\frac{rIn\frac{\hat{q}_2}{\hat{q}_1}}{In\left(\frac{\hat{p}_1}{\hat{p}_2} \cdot \frac{\hat{q}_2}{\hat{q}_1}\right)}\right) - B_{(r,\hat{p}_1)} \frac{rIn\left(\frac{\hat{q}_2}{\hat{q}_1}\right)}{In\left(\frac{\hat{p}_1}{\hat{p}_2} \cdot \frac{\hat{q}_2}{\hat{q}_1}\right)}\right] \tag{2.5.19}$$

$$\hat{p}(mc) = \frac{1}{2}\left[1 + B(r, \hat{p}_2, \lambda) - B(r, \hat{p}_1, \lambda)\right], \quad \hat{p}(mc) = \text{Estimate of Binomial in terms of difference.} \tag{2.5.20}$$

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

$$B(k, \alpha, x) = \sum_{y=0}^{k} \binom{k}{y} \alpha^y (1-\alpha)^{k-y} \text{ where } \quad B(k, \alpha, x) \text{ is Binomial function} \tag{2.5.21}$$

(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

$$\hat{p}_2 = \sum_{k=1}^{n}\sum_{j=1}^{r}\frac{x_{2jk}}{rn_2} \tag{2.5.22}$$

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. That is the maximum likelihood rule says one should allocate x to π_{ij} when

$$L_i = \max L_i(x) \text{ (Anderson, 1984)} \tag{3.1}$$

if π_i is the $Np(\mu_i, \Sigma)$ population, $i = 1,...g$ and $\sum > 0$, then the maximum likelihood discriminant rule allocate x to π_{ij} where $j \in \{1,...n\}$ is that value of i which minimized the Mahalanobis distance $(x-\mu)^1 \sum^{-1}(x-\mu_1)$ where $g = 2$ the rule allocate x to π_1. If $a^1(x-\mu) > 0$ and $a^1\{x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\} > 0$, where $a = \sum^{-1}(\mu_1 - \mu_2)$ and $\mu = (\mu_1 + \mu_2)$ and to π_2 otherwise. Alternatively classify in π_1 if $p(w_1/x) > p(w_2/x)$ or to π_2 if $p(w_1/x) < p(w_2/x)$ (3.2)

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 $\frac{p(x/w_1)p(w_1)}{p(x)} > \frac{p(x/w_2)p(w_2)}{p(x)}$ (3.3)

4. Simulation Experiments and Results

The two 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 where r = number of variables, d = effect size, n = sample size. 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 some 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 = (.4, .4, .4)		P_2 = (.7, .7, .7)
Sample sizes	Optimal	ML
40	0.277450	0.277400
60	0.281258	0.281041
100	0.282180	0.282255
140	0.284160	0.284328
200	0.283407	0.283390
300	0.284403	0.284405
400	0.283510	0.283498
600	0.284085	0.284099
700	0.284371	0.284371
800	0.283587	0.283587
900	0.283666	0.283666
1000	0.283992	0.283992

p(mc) = 0.284

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 = (.4, .4, .4) P_2 = (.7, .7, .7) $\left|p(mc) - \hat{p}(mc)\right|$

Sample sizes	Optimal	ML
40	0.047562	0.046876
60	0.038324	0.038201
100	0.032243	0.032223
140	0.026643	0.026636
200	0.022393	0.022361
300	0.018255	0.018272
400	0.016095	0.016082
600	0.013370	0.013380
700	0.011884	0.0118846
800	0.010636	0.010636
900	0.010394	0.010394
1000	0.009664	0.009664

Tables 4.1(a) and (b) present the mean apparent error rate and standard deviation (actual error rates) of two classification rules. The apparent error rates increases with the sample size. From the table 4.1(b) the error rates decreases with the sample size. With n =1000, two classification rules have the same error rate. On the average, maximum likelihood ranks first, followed by optimal.

Classification Rule	**Performance**
Maximum Likelihood (ML)	1
Optimal (OP)	2

Table 4.2(a) Apparent error rates for classification rules under different parameter values, sample sizes and Replications

P_1 = (.3, .3, .3, .3) P_2 = (.6, .6, .6, .6)

Sample sizes	Optimal	ML
40	0.246987	0.244475
60	0.254608	0.252350
100	0.257285	0.256100
140	0.260228	0.259317
200	0.261217	0.260507
300	0.262273	0.262145
400	0.264232	0.264286
600	0.263918	0.263860
700	0.263235	0.263257
800	0.263443	0.263575
900	0.263276	0.263302
1000	0.264275	0.264263

p(mc) = 0.2637

Table 4.2(b) Actual Error rate for the classification rules under different parameter values, sample sizes and replications.

$P_1 = (.3, .3, .3, .3)$ $P_2 = (.6, .6, .6, .6)$ $\left| p(mc) - \hat{p}(mc) \right|$

Sample size	Optimal	ML
40	0.045504	0.044464
60	0.038768	0.038508
100	0.030567	0.030057
140	0.026331	0.026123
200	0.021757	0.021927
300	0.018459	0.018234
400	0.015636	0.015766
600	0.012377	0.012332
700	0.011465	0.011620
800	0.010715	0.010750
900	0.010140	0.010172
1000	0.009687	0.009657

Tables 4.2(a) and (b) present the mean apparent error rates and standard deviation for the classification rules under different parameter values. The apparent error rates increases with the increase in the sample sizes.

Classification Rule	**Performance**
Maximum Likelihood (ML)	1
Optimal (OP)	2

Table 4.3(a) Apparent error rates for classification rules under different parameter values, sample sizes and Replications

$P_1 = (.5, .5, .5, .5, .5)$ $P_2 = (.6, .6, .6, .6, .6)$

Sample sizes	Optimal	ML
40	0.365212	0.362220
60	0.376908	0.375385
100	0.389975	0.384240
140	0.393925	0.396101
200	0.4007250	0.398143
300	0.402866	0.402204
400	0.404201	0.402156
600	0.405495	0.403902
700	0.406001	0.403770
800	0.406843	0.405535
900	0.406832	0.404521
1000	0.407625	0.405044

p(mc) = 0.40872

Table 4.3(b) Actual error rate for the classification rules under different parameter values, sample sizes and replications.

P_1 = (.5, .5, .5, .5, .5) P_2 = (.6, .6, .6, .6, .6) $\left| p(mc) - \hat{p}(mc) \right|$

Sample size	Optimal	ML
40	0.047146	0.074752
60	0.040174	0.060813
100	0.031479	0.047585
140	0.026298	0.040519
200	0.023616	0.035990
300	0.019186	0.028217
400	0.016343	0.023954
600	0.013147	0.019303
700	0.012653	0.019036
800	0.012157	0.017060
900	0.010951	0.016578
1000	0.010528	0.015555

Table 4.3(a) and (b) show the mean apparent error rates and standard deviation (actual error rates) for the classification rules under different parameter values. It is clear to see that the mean apparent error rate increases with the increase in the sample sizes. The standard deviation decreases with the increase in sample sizes. As the number of variables increases, the performance of the maximum likelihood decreases. From the analysis optimal rule is ranked first, followed by maximum likelihood.

Classification Rule	**Performance**
Optimal (OP)	1
Maximum Likelihood (ML)	2

5. Conclusion

Maximum likelihood procedure performed well for small and moderate number of variables irrespective of the sample size while optimal classification rule appears to be more consistent for small, moderate and large number of variables. Therefore, optimal is more effective classifier than maximum likelihood.

References

Anderson, T. W. (1982). Classification by Multivariate Analysis. *Psychometrika*, *16*, 31-50. http://dx.doi.org/10.1007/BF02313425

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, & Hearing services in schools, 32*, 38-50. http://dx.doi.org/10.1044/0161-1461(2001/004)

Fukunaga, K., & Kessel, D. L. (1972). Application of optimum error-reject function. *IEEE Trans. Information* IT-814-617.

Glaser, B. A., Calhoun, G. B., & Petrocelli, J. V. (2002). Personality characteristics of male juvenile offenders by adjudicated offenses as indicated by the MMH-A. *Criminal Justice and Behaviour, 29*, 184-201. http://dx.doi.org/10.1177/0093854802029002004

Goldstein, M., & Wolf. (1977). On the problem of Bias in multinomial classification. *Biometrics, 33*, 325-331. http://dx.doi.org/10.2307/2529782

Hills, M. (1967). Discrimination and allocation with discrete data. *Applied Statistics, 16*, 237-250. http://dx.doi.org/10.2307/2985920

Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications, 13,* 97-108. http://dx.doi.org/10.1016/S0957-4174(97)00011-0

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

Nworuh, G. E., & Anyiam, K. E. (2010). Applications of Discriminant Analysis on the classification of Implemented Foreign Assisted project in Nigeria. *Journal of the Nigerian Statistical Association, 22*.

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

Onyeagu, S. I. (2003). Derivation of an optimal classification rule for discrete variables. *Journal of Nigerian Statistical Association, 4,* 79-80.

Onyeagu, S. I., & Osuji, G. A. (2010). Evaluation of seven classification procedures for binary variables. *Journal of the Nigerian Statistical Association, 20.*

Panel on Discriminant Analysis, Classification and Clustering. *Statistical Science, 4*, 34-69.

Phillips, M. (2003). Detection of lung cancer with volatile markers in the breath. *Chest, 123,* 2115-2123. http://dx.doi.org/10.1378/chest.123.6.2115

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

Robinson, B. (2002). A structural and discriminant analysis of the Work Addiction Risk Test. *Educational and Psychological Measurement, 62,* 517-526.

Sahiner, B. (2004). Computerized characterization of breast masses of three-dimensional ultrasound volumes. *Medical Physics, 31*, 744-754. http://dx.doi.org/10.1118/1.1649531

Udris, E. M. (2001). Comparing methods to identify general internal medicine clinic patients with chronic heart failure. *American Heart Journal, 142,* 1003-1009. http://dx.doi.org/10.1067/mhj.2001.119130