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

I. Egbo1, S. I. Onyeagu2, D. D. Ekezie3 & 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 \( \pi_1 \) and \( \pi_2 \) be any two multivariate Bernoulli populations. Let \( c(i / j) \) be the cost of misclassifying an item with measurement \( X \) from \( \pi_j \) into \( \pi_i \) and let \( q_j \) be the prior probability on \( \pi_j \), where \( i, j = 1, 2 \) with \( q_i + q_j = 1 \) and probability mass function \( f_i(x) \) in \( \pi_i \) where \( i = 1, 2 \). Suppose that we assign an item with measurement vector \( x \) to \( \pi_1 \) if it is in some region \( R_1 \subseteq R' \) and to \( \pi_2 \) if \( x \) is in some region \( R_2 \subseteq R' \) where \( R' = R_1 \cup R_2 \) and \( R_1 \cap R_2 = 0 \). The expected cost of misclassification is given by:

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

(2.1)

where \( \sum_{R_i} f(x / \pi_i) = p(\text{classifying into } \pi_i / \pi_1) = p(2/1) \), where \( p(2/1) \) when \( \pi_1 \) observation is incorrectly classified as \( \pi_2 \).

The optimal rule is the one that partitions \( R' \) such that

\[
ECM = \sum_{R_i} f(x / \pi_i) = p(\text{classifying into } \pi_i / \pi_2) = p(1/2)
\]

is a minimum.

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

(2.2)

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

(2.3)

ECM is minimized if the second term is minimized. ECM is minimized if \( R_i \) is chosen such that

\[
c(1/2)q_2 f(x / \pi_2) - c(2/1)q_1 f(x / \pi_1) \leq 0
\]

(2.4)

\[
c(2/1)q_1 f(x / \pi_1) \geq c(2/1)q_2 f(x / \pi_2)
\]

(2.5)

\[
R_i = \left\{ x \mid \frac{f(x / \pi_1)}{f(x / \pi_2)} \geq \frac{c(1/2)q_2}{c(2/1)q_1}\right\}
\]

(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 \( \pi_1 \) if

\[
R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{q_2 c(1/2)}{q_1 c(2/1)} \left( \frac{p_2}{p_1} \right)
\]

(2.7)

\[
R_2 : \frac{f_1(x)}{f_2(x)} < \frac{q_2 c(1/2)}{q_1 c(2/1)} \left( \frac{p_2}{p_1} \right)
\]

Otherwise classify into \( \pi_2 \).

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, \( p(\text{mc}) \) under these assumptions, the optimal rule reduces to classifying an item with measurement \( x_0 \) into \( \pi_1 \) if

\[
R_{\text{opt}} : \frac{f_1(x_0 / \pi_1)}{f_2(x_0 / \pi_2)} \geq 1
\]

(2.8)

Otherwise classify the item into \( \pi_2 \). Since \( x \) is multivariate Bernoulli with \( P_{q_i > 0}, i=1,2, j=1,2...r \) the optimal rule is:

classify an item with response pattern \( x \) into \( \pi_1 \) if
\[
\prod_{i} \left[ p_{ij} (1-p_{ij})^{1-x_i} \right] / \prod_{i} \left[ p_{2j} (1-p_{2j})^{1-x_i} \right] > 1
\]

(2.9)

Otherwise, classify the item into \( \pi_1 \). This rule simplifies to:

Classify an item with response pattern \( x \) into \( \pi_1 \) if

\[
\sum_{j=1}^{r} x_j \ln \left( \frac{p_{ij}}{q_{ij}} \right) + \sum_{j=r+1}^{r} \ln \left( \frac{q_{ij}}{p_{ij}} \right) > \sum_{j=1}^{r} \ln \left( \frac{q_{2j}}{q_{1j}} \right)
\]

(2.10)

Otherwise, classify into \( \pi_2 \).

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

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

(2.11)

where \( n_i(x_j) = \sum_{k=1}^{n} x_{ijk} \) is equal to the number of observation from \( \pi_i \) with jth variable. The rule for unknown parameters is: classify an item with response pattern \( x \) into \( \pi_1 \) if

\[
\sum_{j=1}^{r} \ln \left( \frac{\hat{p}_{ij}}{\hat{q}_{ij}} \right) x_j > \sum_{j=1}^{r} \ln \left( \frac{\hat{q}_{2j}}{\hat{q}_{1j}} \right)
\]

(2.12)

otherwise classify the item into \( \pi_2 \).

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 \( x \) into \( \pi_1 \) if:

\[
R_1: \ln \left( \frac{p_{11}q_{21}}{q_{11}p_{21}} \right) x_1 + \ln \left( \frac{p_{12}q_{22}}{q_{12}p_{22}} \right) x_2 > \ln \frac{q_{21}}{q_{11}} + \ln \frac{q_{22}}{q_{12}}
\]

(2.1.1)

Otherwise, classify the item into \( \pi_2 \). Written in another form the rule simplifies to: classify an item with response pattern \( x \) into \( \pi_1 \) if:

\[
R_2: w_1 x_1 + w_2 x_2 > c
\]

(2.1.2)

Otherwise, classify the item into \( \pi_2 \) where

\[
w_1 = \ln \left( \frac{p_{11}}{1-p_{11}} - \frac{1-p_{21}}{p_{21}} \right) = \ln \frac{p_{11}}{1-p_{11}} - \ln \frac{p_{21}}{1-p_{21}}
\]

(2.1.3)

\[
w_2 = \ln \left( \frac{p_{12}}{1-p_{12}} - \frac{1-p_{22}}{p_{22}} \right) = \ln \frac{p_{12}}{1-p_{12}} - \ln \frac{p_{22}}{1-p_{22}}
\]

(2.1.4)

\[
c = \ln \left[ \frac{(1-p_{11})(1-p_{12})}{(1-p_{21})(1-p_{22})} \right]
\]

(2.1.5)

To find the distribution of \( z \) we note that

\[
p[x_j = x_j | \pi_i] = \left\{ \begin{array}{ll} p_{ij} (1-p_{ij}) & j = 1, 2 \\ 0 & \text{otherwise} \end{array} \right.
\]

(2.1.6)
Since

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

(2.1.7)

The range of $z$ is

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

(2.1.8)

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

$$q_1 q_{i2} \text{ if } 0 \leq z < w_1$$

(2.1.14)

$$p(z/\pi_i) = q_1 q_{i2} + p_2 q_{i2} \text{ if } w_1 \leq z < w_2 = q_1 q_{i2} + p_2 q_{i2} + p_3 q_{i3} \text{ if } w_2 \leq z < w_1 + w_2$$

(2.1.15)

$$1 \text{ if } w_1 + w_2 \leq z$$

(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 $x$ into $\pi_1$ if:

$$R_{R_1} : \ln \left( \frac{P_{11} q_{21}}{q_{11} P_{21}} \right) x_1 + \ln \left( \frac{P_{12} q_{22}}{q_{12} P_{22}} \right) x_2 + \ln \left( \frac{P_{13} q_{23}}{q_{13} P_{23}} \right) x_3 > \ln \left( \frac{q_{22} q_{22} q_{23}}{q_{12} q_{12} q_{13}} \right)$$

(2.2.1)

otherwise, classify the item into $\pi_2$. Written in another form the rule simplifies to: classify an item with response pattern $x$ into $\pi_1$ if:

$$R_{R_2} : w_1 x_1 + w_2 x_2 + w_3 x_3 > c$$

(2.2.2)

otherwise classify the item into $\pi_2$.

$$w_1 = \ln \left( \frac{P_{11} q_{21}}{q_{11} P_{21}} \right), w_2 = \ln \left( \frac{P_{12} q_{22}}{q_{12} P_{22}} \right), w_3 = \ln \left( \frac{P_{13} q_{23}}{q_{13} P_{23}} \right)$$

$$c = \ln \left( \frac{q_{22} q_{22} q_{23}}{q_{12} q_{12} q_{13}} \right)$$

(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 $x$ into $\pi_1$ if

$$R_{R_1} : \ln \left( \frac{P_{11} q_{21}}{q_{11} P_{21}} \right) x_1 + \ln \left( \frac{P_{12} q_{22}}{q_{12} P_{22}} \right) x_2 + \ln \left( \frac{P_{13} q_{23}}{q_{13} P_{23}} \right) x_3 + \ln \left( \frac{P_{14} q_{24}}{q_{14} P_{24}} \right) x_4 > \ln \left( \frac{q_{21}}{q_{11}} + \ln \frac{q_{22}}{q_{12}} + \ln \frac{q_{23}}{q_{13}} + \ln \frac{q_{24}}{q_{14}} \right)$$

(2.3.1)
otherwise, classify the item into $\pi_2$. Written in another form, the rule simplifies to: classify an item with response pattern $x$ into $\pi_1$ if: $R_{\theta_0} = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 > c_1 + c_2 + c_3 + c_4$ otherwise, classify the item into $\pi_2$. For the case of four variables, let

$$w_1 = \ln \left( \frac{p_{11}}{q_{11}} \cdot \frac{q_{21}}{p_{21}} \right), w_2 = \ln \left( \frac{p_{12}}{q_{12}} \cdot \frac{q_{22}}{p_{22}} \right), w_3 = \ln \left( \frac{p_{13}}{q_{13}} \cdot \frac{q_{23}}{p_{23}} \right),$$

$$w_4 = \ln \left( \frac{p_{14}}{q_{14}} \cdot \frac{q_{24}}{p_{24}} \right) \tag{2.3.2}$$

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 $\pi_1$ or $\pi_2$. 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 $\pi_1$ and if it falls in $R_2$ we classify it as coming from $\pi_2$. In following a given classification procedure, the researcher can make two kinds of errors in classification. If the item is actually from $\pi_1$, the researcher can classify it as coming from $\pi_2$. Also the researcher can classify an item from $\pi_2$ as coming from $\pi_1$. We need to know the relative undesirability of these two kinds of errors in classification. Let the prior probability that an observation comes from $\pi_1$ be $q_1$, and from $\pi_2$ be $q_2$. Let the probability mass function of $\pi_1$ be $f_1(x)$ and that of $\pi_2$ be $f_2(x)$. Let the regions of classifying into $\pi_1$ be $R_1$ and into $\pi_2$ be $R_2$. Then the probability of correctly classifying an observation that is actually from $\pi_1$ into $\pi_1$ is

$$p(1/1) = \sum_{x} f_1(x)$$

and the probability of misclassifying such an observation into $\pi_2$ is

$$p(2/1) = \sum_{x} f_1(x) \tag{2.4.1}$$

Similarly, the probability of correctly classifying an observation from $\pi_2$ into $\pi_2$ is

$$p(2/2) = \sum_{x} f_2(x)$$

and the probability of misclassifying an item from $\pi_1$ into $\pi_2$ is

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

The total probability of misclassification using the rule is

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

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, $R_{\text{opt}}$. When the parameters of the distributions are known, the error rate is

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

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 $\pi_1$ and $n_2$ from $\pi_2$.

(iv) The plug-in estimate of error rate obtained by using the estimated parameters for $\pi_1$ and $\pi_2$.

(v) The apparent error rate: This is defined as the fraction of items in the initial sample which is misclassified by the classification rule.
The table above is called the confusion matrix and the apparent error rate is given by

\[
\hat{P}(mc) = \frac{n_{12} + n_{21}}{n}
\]

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 \(x = (x_1, x_2, ..., x_r)\) which is distributed multivariate Bernoulli is: classify an item with response pattern \(x\) into \(\pi_1\) if

\[
R_{opt}^{(1)}: \sum_{j=1}^{r} x_j \ln \left( \frac{p_{ij}}{q_{ij}} \right) > \sum_{j=1}^{r} \ln \frac{q_{ij}}{q_{1j}}
\]

Otherwise classify into \(\pi_2\)

We can obtain the probability of misclassification for two cases

Case I: Known parameters

(a) General case where \(p_i = (p_{1i}, p_{2i}, ..., p_{ri})\)

(b) Special case where \(p_i = (p_1, p_2, ..., p_i)\) 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 \(x = (x_1, x_2, ..., x_r)\) which is distributed multivariate Bernoulli is: classify an item with response pattern \(x\) into \(\pi_1\) if

\[
R_{opt}^{(1)}: \sum_{j=1}^{r} x_j \ln \left( \frac{p_{ij}}{q_{ij}} \right) > \sum_{j=1}^{r} \ln \frac{q_{ij}}{q_{1j}}
\]

Otherwise classify into \(\pi_2\)

Case 1b: Special case where \(p_i = p(p_1, p_2, ..., p_i)\) 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 \(x\) into \(\pi_1\) if otherwise classify into \(\pi_2\). The probability of misclassification using the special case of \(R_{opt}\) is

\[
R_{opt}^{(2)}: \sum_{j=1}^{r} x_j \leq \frac{r \ln \left( \frac{q_{1j}}{q_{2j}} \right)}{\ln \left( \frac{p_{1j}}{p_{2j}} \right)}
\]

\[
p(2/1) = p \left[ \sum_{j=1}^{r} x_j > \frac{r \ln \left( \frac{q_{1j}}{q_{2j}} \right)}{\ln \left( \frac{p_{1j}}{p_{2j}} \right)} \right] \pi_1 = 1 - B_{(r, p_1)} \left( \frac{r \ln \left( \frac{q_{1j}}{q_{2j}} \right)}{\ln \left( \frac{p_{1j}}{p_{2j}} \right)} \right)
\]

\[
B_{r, p}(x) = \sum_{y=0}^{r} \binom{r}{y} p^y (1 - p)^{r-y}, \text{ where } B_{r, p} \text{ have binomial distribution with parameters } r \text{ and } p
\]

\[
p(1/2) = p \left[ \sum_{j=1}^{r} x_j < \frac{r \ln \left( \frac{q_{1j}}{q_{2j}} \right)}{\ln \left( \frac{p_{1j}}{p_{2j}} \right)} \right] \pi_2 = B_{(r, p_2)} \left( \frac{r \ln \left( \frac{q_{1j}}{q_{2j}} \right)}{\ln \left( \frac{p_{1j}}{p_{2j}} \right)} \right)
\]
\[ p(mc) = \frac{1}{2} \left[ 1 + B_{(r,p_2)} \left( \frac{r \ln \frac{q_i}{q_j}}{\ln \left( \frac{p_i}{p_j} \cdot \frac{q_i}{q_j} \right)} \right) - B_{(r,p_2)} \left( \frac{r \ln \frac{q_i}{q_j}}{\ln \left( \frac{p_i}{p_j} \cdot \frac{q_i}{q_j} \right)} \right) \right] \]  

(2.5.7)

Case 1c: Special case (1b) with additional assumption that \( p_i = \phi_{r_2} \) and \( q_i = 1 - p_i = 1 - \phi_{r_2} \) and \( q_i = 1 - p_i \). The optimal classification rule \( R_{opt} \) for \( \pi = (\pi_1, \pi_2, ..., \pi_r) \) distributed multivariate Bernoulli is: classify the item with response pattern \( x \) into \( \pi_1 \) if

\[ R_{opt}: \sum_{j=1}^{r} x_j > \left[ \frac{r \ln \frac{1-p_1}{1-\phi_{r_2}}}{\ln \left( \frac{1-p_1}{1-\phi_{r_2}} \right)} \right] \]

(2.5.8)

and to \( \pi_2 \) otherwise.

The probability of misclassification using the special case of \( R_{opt} \) when \( p_i = \phi_{r_2} \) is

\[ p(2/1) = 1 - B_{(r,\phi_{r_2})} \frac{r \ln \frac{1-p_2}{1-\phi_{r_2}}}{\ln \left( \frac{1-p_2}{1-\phi_{r_2}} \right)} \]

(2.5.9)

\[ p(1/2)B_{(r,p_2)} \frac{r \ln \frac{1-p_2}{1-\phi_{r_2}}}{\ln \left( \frac{1-p_2}{1-\phi_{r_2}} \right)} \]

(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 \( \pi_1 \) and \( \pi_2 \) respectively. In \( \pi_1 \) we have the sample

\[
\begin{align*}
  x_{1i} &= (x_{111}, x_{121}, x_{131}, ..., x_{1r1}) \\
  x_{12} &= (x_{112}, x_{122}, x_{132}, ..., x_{1r2}) \\
  &\vdots \\
  x_{1n1} &= (x_{11n1}, x_{12n1}, x_{13n1}, ..., x_{1rn1})
\end{align*}
\]

The maximum likelihood estimate of \( p_1 \) is

\[ \hat{p}_{ik} = \frac{\sum_{j=1}^{r} x_{ijk}}{n_1} \]

(2.5.12)

Similarly the maximum likelihood estimate of \( p_2 \) is

\[ \hat{p}_{2k} = \frac{\sum_{j=1}^{r} x_{2kj}}{n_2} \]

(2.5.13)

We plug 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 \( \pi_1 \) if

\[ R_{Br}: \sum_{j=1}^{r} x_j \ln \left( \frac{\hat{p}_{ij} \hat{q}_{2j}}{\hat{q}_{ij} \hat{p}_{2j}} \right) > \frac{r \ln \frac{\hat{q}_{2j}^2}{\hat{p}_{2j}^2}}{\hat{q}_{ij}^2} \]

(2.5.14)
otherwise classify into \( \pi_2 \).

(b) Special case of 1b where \( p_i = (p_1, p_2, \ldots, p_n) \) with the assumption that \( p_{1i} < p_{2i} \).

In this special case

\[ \hat{p}_1 = \frac{\sum_{j=1}^{m_1} x_{1ij}}{n_1} = \frac{\sum_{j=1}^{m_2} x_{2ij}}{n_2} = \cdots = \frac{\sum_{j=1}^{m_n} x_{nij}}{n_n} \]  

(2.5.15)

\[ \sum_{j=1}^{m} x_{ij} \text{ is distributed } B(r, p_i) \]

\[ \sum_{j=1}^{m} \sum_{j=1}^{m} x_{ij} \text{ is distributed } B(r m_1, p_i) \]

The maximum likelihood estimate of \( p_i \) is

\[ \hat{p}_1 = \frac{\sum_{j=1}^{m} \sum_{j=1}^{m} x_{1ij}}{rm_1} \]  

(2.5.16)

Likewise, the maximum likelihood estimate of \( p_2 \) is

\[ \hat{p}_2 = \frac{\sum_{j=1}^{m} \sum_{j=1}^{m} x_{2ij}}{rm_2} \]  

(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 \( \pi_i \) if

\[ \sum_{j=1}^{m} x_j \leq \frac{r \ln \left( \frac{\hat{p}_i}{\hat{p}_j} \right)}{\ln \left( \frac{\hat{p}_i}{\hat{p}_j} \right)} \]  

(2.5.18)

Otherwise classify into \( \pi_2 \)

The probability of misclassification is given by

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

(2.5.19)

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

\[ \hat{p}(mc) = \text{Estimate of Binomial in terms of difference.} \]  

(2.5.20)

Where \( \lambda = \frac{r \ln \left( \frac{\hat{p}_i}{\hat{p}_j} \right)}{\ln \left( \frac{\hat{p}_i}{\hat{p}_j} \right)} \)

\[ B(k, \alpha, x) = \sum_{\gamma=0}^{k} \binom{k}{\gamma} \alpha^\gamma (1-\alpha)^{k-\gamma} \text{ where } B(k, \alpha, x) \text{ is Binomial function} \]  

(2.5.21)

(c) Special case of 2b with \( p_i = \theta p_{2i}, p_{2i} < p_{1i}, 0 < \theta < 1 \) we take training samples of size \( n_2 \) from \( \pi_2 \) and estimate \( p_2 \) by
3. Maximum Likelihood Rule (ML-Rule)

The maximum likelihood discriminant rule for allocating an observation $x$ to one of the populations $\pi_1, \ldots, \pi_n$ 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 $\pi_j$ when

$$L_j = \max L_i(x) \quad \text{(Anderson, 1984)}$$

(3.1)

if $\pi_i$ is the $\mathcal{N}(\mu_i, \Sigma)$ population, $i = 1, \ldots, g$ and $\Sigma > 0$, then the maximum likelihood discriminant rule allocate $x$ to $\pi_{ij}$ where $j \in \{1, \ldots, n\}$ is that value of $i$ which minimized the Mahalanobis distance $(x-\mu)^T \Sigma^{-1} (x-\mu)$ where $g = 2$ the rule allocate $x$ to $\pi_1$. If $a^T (x-\mu) > 0$ and $a^T (x-\mu) > 0$, where $a = \sum^{-1} (\mu_1 - \mu_2)$ and $\mu = (\mu_1 + \mu_2)$ and to $\pi_2$ otherwise. Alternatively classify in $\pi_1$ if $p(w_1 / x) > p(w_2 / x)$ or to $\pi_2$ 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 $\pi_1$ 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 $n_1 = \frac{n}{2}$ observations are sampled from $\pi_1$, which has multivariate Bernoulli distribution with input parameter $p_1$, and $n_2 = \frac{n}{2}$ observations sampled from $\pi_2$, which is multivariate Bernoulli with input parameter $p_2$, $j = 1, \ldots, r$. 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>
<thead>
<tr>
<th>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)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1 = (0.4, 0.4, 0.4)$</td>
</tr>
<tr>
<td>Sample sizes</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>140</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>300</td>
</tr>
<tr>
<td>400</td>
</tr>
<tr>
<td>600</td>
</tr>
<tr>
<td>700</td>
</tr>
<tr>
<td>800</td>
</tr>
<tr>
<td>900</td>
</tr>
<tr>
<td>1000</td>
</tr>
</tbody>
</table>

$p(mc) = 0.284$
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)

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

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 increase 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.

<table>
<thead>
<tr>
<th>Classification Rule</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Likelihood (ML)</td>
<td>1</td>
</tr>
<tr>
<td>Optimal (OP)</td>
<td>2</td>
</tr>
</tbody>
</table>

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

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

$p(mc) = 0.2637$
Table 4.2(b) Actual Error rate for the classification rules under different parameter values, sample sizes and replications.

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

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**

<table>
<thead>
<tr>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Likelihood (ML)</td>
</tr>
<tr>
<td>Optimal (OP)</td>
</tr>
</tbody>
</table>

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

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

p(mc) = 0.40872
Table 4.3(b) Actual error rate for the classification rules under different parameter values, sample sizes and replications.

\[ P_1 = (.5, .5, .5, .5, .5) \]
\[ P_2 = (.6, .6, .6, .6, .6) \]

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

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

<table>
<thead>
<tr>
<th>Classification Rule</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal (OP)</td>
<td>1</td>
</tr>
<tr>
<td>Maximum Likelihood (ML)</td>
<td>2</td>
</tr>
</tbody>
</table>

### 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


**Copyrights**

Copyright for this article is retained by the author(s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).