Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.11, No.3, 2021

EVALUATION OF THREE CLASSIFICATION RULES FOR MIXTURE OF DISCRETE AND CONTINOUS VARIABLES

Iwuagwu Chukwuma E[1]. Onyeagu Sidney I.[2], Chrisogonus K. Onyekwere[3]

[1]Department of Statistics, Abia State Polytechnic Aba, Nigeria

[2,3]Department of Statistics, Nnamdi Azikiwe University Awka, Nigeria

Corresponding author: E-mail: chrisogonusjohnson@gmail.com

ABSTRACT

The best classification rule is the one that leads to the smallest probability of misclassification which is called the error rate. This work focused on three classification rules for mixture of discrete and continuous variables with the aim to evaluate the performance of these rules to in classification of individuals into several categories. Applications were done using simulated data and real life data. The result obtained revealed that the location model achieved better result than the other two rules in minimizing the average error rate in both datasets.

Keyword: Location Model, Linear Discriminant Models, Quadratic, Discriminant Model, Error Rate.

1 Introduction

Discriminant analysis is a statistical method that enable one in understanding the differences of objects between two or more groups with respect to several variables simultaneously Hamid (2010). We use discriminant analysis when we have a categorical outcome. The assumption of normality must be satisfied for one to use this statistical technique. When the assumption fails, we resort to logistic regression, which is assumption free. In discriminant analysis, researchers are often faced with misclassification problems when assigning an unknown observation to a group with low error rate, therefore in order to avoid misclassification problem which could lead to losses incurred by the estimation procedure when converting discrete to continuous or continuous to discrete, Krzanowski (1975a) developed the location model which can handle mixed variables on classification simultaneously. He showed that, the model yielded a better result than the linear discriminant function did. Based on his findings, this paper will centre on determining the average error rate for the Location Model (LM), Linear Discrimination Function (LDF) and Quadratic Discrimination Function (QDF). One of the significant of this research is that it will help to determine the strength and weakness of the models in terms of the losses in the estimation procedures.

In this paper we will also look at the problem of discriminant analysis when an individual is to be allocated to one or the other when the data comprises a mixture of discrete and continuous variables without incurring many losses in the estimation procedure when converting variables from discrete to continuous or from continuous to discrete. According to Krzanowski (1975b), the binary variables (x) expressed as multinomial variable having $k = 2^q$ states and continuous variables (y) has a multivariate normal distribution with mean (μ) and common

dispersion matrix Σ in all cells. The optimal decision is to construct a rule that minimizes the average of the two probabilities of misclassification that is $(P_1 + P_2)/2$ is minimized, where $P_1(P_2)$ represents the probability that an observation from $\pi_1(\pi_2)$ is classified into$\pi_2(\pi_1)$.

2 Methodology

The data used for this research work was based on the simulated and real data. In the simulated data, a data set was generated using R-Programming language and the average error rate were computed for $2 \leq q \leq 10$ and $0.1 \leq P_1 P_2 \leq 0.9$ for two situations, that is situation (a) a case with no interaction between discrete and continuous and situation (b) a case involving interactions between discrete and continuous variables, also q, are the components of the discrete variables x and p are the components of the continuous variable, y. The real data came from primary and secondary source.

The primary data consist of 12 variables obtained from project implementation which involved ten continuous variables and (2) two were discrete. The secondary data consist of 15 variables obtain from UNDP report on Human Development Index. Thirteen variables are continuous and two are discrete.

2.1 Location model

The location model was introduced by Krzanowski (1975c) provided a method for discriminating between two groups and allocating individuals to one or the other, when the available data consists of both binary and continuous variables

This model is a predictive discriminant rule that can be used to assign new observations into one of the two predefined groups Mahat *et al* (2009).

In the development of this model Krzanowski made use of both continuous and categorical variables.

Let x, denote the vector of discrete variable with q, components and y, represent the vector of the continuous variables with P, components. If the discrete variables have been allocated to an individual cell j, the continuous variables, y, have a multivariate normal distribution with mean μ (i) and dispersion matrix Σ, in population π_i. π_i(i = i, 2, j = 1.2). Then the conditional probability density of j is

$$\frac{1}{(2\pi)^{c/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}\left(y-\mu_i^{(j)}\right)'\Sigma^{-1}\left(y-\mu_i^{(j)}\right)\right\} \text{ in } \pi_i, (i = 1,2). \quad (1)$$

Thus, the joint probability density of obtaining the individual in cell j and observing the continuous variable value, y, is

$$y, \text{ is } \frac{p_{ij}}{(2\pi)^{c/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}\left(y-\mu_i^{(j)}\right)'\Sigma^{-1}\left(y-\mu_i^{(j)}\right)\right\} \text{ in } \pi_i, (i = 1,2) \quad (2)$$

2.2 Estimation of Error Rates

If the parameter is known in the location model, the error rates are given in (3) and (4) below

$$p(2/1) = \sum_{m=1}^{k} p_{1_m} \Phi\left(log\left(P_{2_m}/p_{1_m}\right) - {}^{1}/_{2}\, D_m^2/D_m\right) \quad (3)$$

$$p(1/2) = \sum_{m=1}^{k} p_{2_m} \Phi\left(log\left(p_{1_m}/p_{2_m}\right) - {}^{1}/_{2}\, D_m^2/D_m\right) \quad (4)$$

where, Φ is the cumulative standard normal distribution function and

$$D_m^2 = \left(\mu_1^{(m)} - \mu_2^{(m)}\right) \sum^{-1} \left(\mu_1^{(m)} - \mu_2^{(m)}\right) \quad (5)$$

is the Mahalanobi's squared distance between π_1 and π_2 in cell j of the multinomial table.

2.3 Fisher's linear discriminant function

Fisher suggested using a linear combination of the observations and choosing the coefficients so that the ratio of the differences of the means of linear combination in the two of groups to its variance is maximized.

In Fisher's approach, the linear combination is denoted by $Y = \lambda^{\mathrm{I}} x$, the mean of y is $\lambda^{\mathrm{I}}\mu_1$, in π_1 and $\lambda^{\mathrm{I}}\mu_2$ in π_2. Its variance $\lambda^{\mathrm{I}}\sum\lambda$ in either population and covariance matrices $\sum_1 = \sum_2 = \sum$. In practical term the general form of linear discriminant function is

$$Y = L_1 X_1 + L_2 X_2 + \cdots + L_p X_p \quad (6)$$

The coefficients are estimated by solving the simultaneous equations.

$$Y = L_1 S_{i1} + L_2 S_{i2} + \cdots + L_p S_{ip} = d_i \quad (i = 1, 2, \cdots, p) \quad (7)$$

Where S_{ij} are the elements of the pooled dispersion matrix and $d = \bar{x}_{i2} - \bar{x}_{i1}$ and the equation above can be written in matrix form as

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_p \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{bmatrix}$$

2.4 Quadratic discriminant analysis

Quadratic Discriminant Analysis (QDA) is a standard probabilistic classification method in statistics and machine learning and it assumes class-conditional distribution to be normal and then classifies given point, by the posterior distributions Wenbo (2015). The densities of class conditional probabilities can be written as

$$P(X = x/Y = K; \mu_K \Sigma_K) = \frac{1}{\sqrt{(2\pi)^D \, det\, \Sigma_k}} \ell^{1/2(X-\mu_K)^T} \sum_k^{-1} (X - \mu_K) \quad (8)$$

where $\mu_k$and $\Sigma_k$are the mean and covariance matrix of the class conditional probability for class k.

The posterior probabilities can be derived via Bayes theorem. That is,

$$P\left(Y = k/X = X_i \mu_1 \ldots \mu_{k,} \Sigma_1 \ldots_k, \Pi\right) = \frac{\Pi_k P(X/K; \mu_K \Sigma_K)}{\sum_{K-1}^{K} \Pi_K \; P\left(X_K/K; \mu_{K,} \Sigma_K\right)} \quad (9)$$

where $\mu_1 \ldots \mu_{k,}\, \Sigma_1 \ldots \Sigma_k and \Pi = (\Pi_1\, \Pi_2, \ldots, \Pi_k)$ are unknown parameters.

In classical QDA, $\Pi'^{S}_{k}, \mu'^{S}_{k}$ $and \Sigma'^{S}_{k}$ are estimated by maximizing the joint likelihood of observations and their labels which can be formally written as follows:

$$Max\ \Pi_{k-1}^{k}\, \Pi_{yn} - kP(X = x_n/Y = k; \Sigma_k; \mu_k)\Pi_k \tag{10}$$

$$\Sigma_k > 0 \quad k = 1,\ 2, \cdots, k,$$

$$\Pi_k > 0 \quad k = 1,\ 2, \cdots, k;$$

s.t. $\sum_{k=1}^{k} \Pi_k = 1.$

If $\Sigma_k > 0$, the $\Sigma_k$is a positive definite matrix. The estimations can be obtained using the following

$$\widehat{\Pi}_k = \frac{N_k}{N} \tag{11}$$

$$\mu_k^2 = \frac{1}{N_K} \sum_{y_{n=k}} x_n \tag{12}$$

$$\sum_k^2 = \frac{1}{\mu_k} \sum_{y_n=k} (X_n - \mu_k^2)\,(X_n - \mu_k^2)^T \tag{13}$$

and $\sum_0^2 = \sum_{k=1}^{k} \frac{\mu_K}{N}$

N_k is the number of class k observations in the given training set**.**

3 Estimations

The result of the simulated data is shown in the tables below. The simulated data was generated with R-Programming Language and average error rates were computed as stated in methodology.

Table 3.1: Average Error Rate for $2 \leq q \leq 5$ & $0.1 \leq P_1, P_2 \leq 0.9$

			q = 2			q = 3			q = 4			q = 5		
P1	P2	Situation	LM	LDF	QDF	LM	LDF	QDF	LM	LDF	QDF	LM	LDF	QDF
	0.1	A	0.2323	0.2453	0.2490	0.2490	0.2608	0.2453	0.2520	0.2435	0.2369	0.2790	0.2580	0.3237
		B	0.2560	0.2620	0.2533	0.2478	0.2515	0.2563	0.2585	0.2500	0.2356	0.2510	0.2343	0.3225
	0.3	A	0.2605	0.2708	0.2493	0.2660	0.2503	0.2463	0.2315	0.2423	0.2339	0.2540	0.2538	0.2761
		B	0.2765	0.2448	0.2515	0.2583	0.2485	0.2588	0.2463	0.2513	0.2280	0.2433	0.2450	0.2783
0.1	0.5	A	0.2665	0.2575	0.2545	0.2405	0.2400	0.2465	0.2568	0.2420	0.2369	0.2560	0.2515	0.2592
		B	0.2388	0.2555	0.2418	0.2455	0.2658	0.2413	0.2485	0.2510	0.2796	0.2435	0.2628	0.2630
	0.7	A	0.2425	0.2503	0.2528	0.2668	0.2465	0.2463	0.2665	0.2468	0.2406	0.2420	0.2613	0.2611
		B	0.2415	0.2680	0.2368	0.2595	0.2580	0.2455	0.2555	0.2423	0.2187	0.2586	0.2783	0.2640
	0.9	A	0.2265	0.2420	0.2523	0.2443	0.2343	0.2520	0.2383	0.2543	0.2678	0.2570	0.2345	0.2600
		B	0.2475	0.2618	0.2525	0.2538	0.2538	0.2515	0.2535	0.2690	0.2538	0.2660	0.2438	0.2640
	0.3	A	0.2490	0.2530	0.2410	0.2520	0.2450	0.2490	0.2518	0.2350	0.2378	0.2403	0.2428	0.3224
		B	0.2295	0.2743	0.2410	0.2615	0.2515	0.2480	0.2483	0.2435	0.2650	0.2518	0.2558	0.3186
0.3	0.5	A	0.2463	0.2540	0.2463	0.2505	0.2628	0.2380	0.2550	0.2608	0.2448	0.2570	0.2475	0.3064
		B	0.2570	0.2633	0.2520	0.2473	0.2613	0.2585	0.2560	0.2510	0.2630	0.2528	0.2470	0.3032
	0.7	A	0.2385	0.2478	0.2448	0.2515	0.2783	0.2508	0.2590	0.2450	0.2600	0.2405	0.2463	0.2882
		B	0.2248	0.2583	0.2488	0.2523	0.2535	0.2445	0.2513	0.2473	0.2378	0.2323	0.2350	0.2875
	0.9	A	0.2565	0.2328	0.2520	0.2378	0.2438	0.2598	0.2558	0.2448	0.2383	0.2363	0.2488	0.2730
		B	0.2278	0.2483	0.2463	0.2293	0.2328	0.2345	0.2533	0.2625	0.2625	0.2478	0.2625	0.2800
	0.5	A	0.2235	0.2748	0.2493	0.2513	0.2558	0.2510	0.2520	0.2440	0.2473	0.2438	0.2803	0.3164
0.5		B	0.2428	0.2450	0.2455	0.2533	0.2475	0.2488	0.2540	0.2468	0.2600	0.2430	0.2580	0.3222
	0.7	A	0.2595	0.2475	0.2455	0.2503	0.2470	0.2510	0.2510	0.2485	0.2490	0.2350	0.2580	0.3403
		B	0.2458	0.2705	0.2523	0.2508	0.2463	0.2470	0.2298	0.2518	0.2605	0.2436	0.2343	0.3391
	0.9	A	0.2503	0.2330	0.2498	0.2575	0.2350	0.2475	0.2533	0.2540	0.2508	0.2598	0.2538	0.3328
		B	0.2438	0.2703	0.2450	0.2288	0.2488	0.2468	0.2505	0.2523	0.2500	0.2498	0.2450	0.3297
	0.7	A	0.2388	0.2410	0.2513	0.2635	0.2625	0.2458	0.2595	0.2680	0.2490	0.2503	0.2515	0.3460
0.7		B	0.2715	0.2705	0.2448	0.2403	0.2803	0.2510	0.2505	0.2568	0.2523	0.2703	0.2623	0.3492
	0.9	A	0.2363	0.2383	0.2453	0.2473	0.2580	0.2643	0.2420	0.2538	0.2513	0.2438	0.2580	0.3237
		B	0.2315	0.2630	0.2390	0.2383	0.2418	0.2535	0.2458	0.2578	0.2488	0.2430	0.2343	0.3225

			q=6			q=7			q=8			q=9		
P1	P2	Situation	LM	LDF	QDF	LM	LDF	QDF	LM	LDF	QDF	LM	LDF	QDF
	0.1	A	0.2777	0.2582	0.3240	0.2794	0.2564	0.3235	0.2782	0.2529	0.3218	0.2782	0.2563	0.3244
		B	0.2493	0.2350	0.3221	0.2514	0.2329	0.3225	0.2520	0.2287	0.3204	0.2540	0.2364	0.3209
	0.3	A	0.2550	0.2524	0.2747	0.2539	0.2521	0.2760	0.2566	0.2576	0.2778	0.2545	0.2484	0.2770
		B	0.2460	0.2444	0.2790	0.2451	0.2462	0.2763	0.2460	0.2512	0.2756	0.2470	0.2446	0.2747
0.1	0.5	A	0.2569	0.2516	0.2589	0.2563	0.2479	0.2594	0.2560	0.2413	0.2595	0.2537	0.2469	0.2600
		B	0.2437	0.2630	0.2634	0.2441	0.2617	0.2630	0.2458	0.2594	0.2641	0.2439	0.2622	0.2613
	0.7	A	0.2447	0.2609	0.2607	0.2431	0.2595	0.2615	0.2482	0.2544	0.2624	0.2437	0.2609	0.2609
		B	0.2581	0.2782	0.2638	0.2584	0.2788	0.2640	0.2604	0.2810	0.2644	0.2601	0.2778	0.2638
	0.9	A	0.2575	0.2536	0.2601	0.2578	0.2522	0.2600	0.2592	0.2513	0.2594	0.2611	0.2511	0.2588
		B	0.2661	0.2439	0.2637	0.2665	0.2459	0.2644	0.2667	0.2470	0.2633	0.2656	0.2465	0.2661
	0.3	A	0.2405	0.2428	0.3229	0.2408	0.2454	0.3227	0.2413	0.2491	0.3242	0.2419	0.2471	0.3201
		B	0.2494	0.2560	0.3188	0.2520	0.2564	0.3175	0.2482	0.2555	0.3194	0.2489	0.2605	0.3163
0.3	0.5	A	0.2561	0.2472	0.3054	0.2587	0.2466	0.3070	0.2572	0.2460	0.3060	0.2594	0.2432	0.3059
		B	0.2520	0.2472	0.3037	0.2545	0.2463	0.3050	0.2536	0.2458	0.3026	0.2544	0.2480	0.3026
	0.7	A	0.2428	0.2469	0.2854	0.2407	0.2483	0.2884	0.2434	0.2454	0.2912	0.2416	0.2529	0.2878
		B	0.2222	0.2363	0.2875	0.2316	0.2365	0.2889	0.2472	0.2290	0.2898	0.2315	0.2423	0.2871
	0.9	A	0.2346	0.2484	0.2728	0.2381	0.2498	0.2732	0.2381	0.2547	0.2731	0.2374	0.2506	0.2742
		B	0.2487	0.2630	0.2800	0.2469	0.2633	0.2816	0.2456	0.2638	0.2840	0.2488	0.2660	0.2787
	0.5	A	0.2428	0.2803	0.3154	0.2437	0.2798	0.3174	0.2422	0.2758	0.3163	0.2455	0.2826	0.3160
0.5		B	0.2450	0.2589	0.3218	0.2440	0.2567	0.3229	0.2245	0.2478	0.3203	0.2436	0.2596	0.3259
	0.7	A	0.2353	0.2582	0.3405	0.2372	0.2564	0.3426	0.2333	0.2529	0.3463	0.2372	0.2563	0.3419
		B	0.2436	0.2350	0.3391	0.2439	0.2329	0.3392	0.2435	0.2287	0.3390	0.2445	0.2364	0.3423
	0.9	A	0.2634	0.2523	0.3330	0.2606	0.2521	0.3335	0.2609	0.2576	0.3337	0.2596	0.2484	0.3349
		B	0.2493	0.2444	0.3302	0.2507	0.2462	0.3268	0.2543	0.2512	0.3270	0.2496	0.2446	0.3264
	0.7	A	0.2508	0.2516	0.3452	0.2510	0.2479	0.3477	0.2507	0.2413	0.3486	0.2514	0.2469	0.3488
0.7		B	0.2708	0.2630	0.3498	0.2690	0.2617	0.3461	0.2682	0.2594	0.3472	0.2690	0.2622	0.3417
	0.9	A	0.2778	0.2582	0.3240	0.2794	0.2564	0.3235	0.2782	0.2529	0.3218	0.2782	0.2563	0.3244
		B	0.2493	0.2350	0.3221	0.2514	0.2329	0.3225	0.2520	0.2287	0.3204	0.2540	0.2364	0.3209

q=10		
LM	LDF	QDF
0.2809	0.2548	0.3236
0.2532	0.2305	0.3222
0.2526	0.2505	0.2766
0.2520	0.2510	0.2755
0.2604	0.2492	0.2596
0.2474	0.2629	0.2639
0.2440	0.2542	0.2623
0.2629	0.2806	0.2651
0.2590	0.2534	0.2592
0.2678	0.2482	0.2637
0.2376	0.2494	0.3238
0.2584	0.2536	0.3202
0.2576	0.2479	0.3065
0.2578	0.2431	0.3026
0.2392	0.2492	0.2904
0.2406	0.2389	0.2871
0.2394	0.2490	0.2736
0.2464	0.2638	0.2823
0.2452	0.2791	0.3161
0.2428	0.2577	0.3228
0.2382	0.2548	0.3440
0.2461	0.2305	0.3403
0.2648	0.2505	0.3356
0.2490	0.2510	0.3261
0.2528	0.2492	0.3486
0.2716	0.2629	0.3451
0.2809	0.2548	0.3236
0.2532	0.2305	0.3222

Average Error Rate for $6 \leq q \leq 10$ & $0.1 \leq P_1, P_2 \leq 0.9$

Table 3.2: Summary of the First Position in the Performance of the three Classification Rules using Simulated Data

Classification Rule	Situation A	No.	Situation B	No.
Location Model	$q = 2, q = 4,$ $q = 6, q = 8$ $q = 9$, q=10	6	$q = 2, q = 3,$ $q = 5, q = 6$ $q = 7, q = 8$	6
LDF	$q = 3, q = 5$	2	$q = 10$	1
QDF	NIL	NIL	NIL	NIL

3.1 Application to real data

Table 3.3 below shows the result obtained for average error rate using the real data for the three classification rules

Table 3.3

Data	LM	LDF	QDF
Primary	0.2400	0.2450	0.2600
Secondary	0.2694	0.2796	0.2798

4 Findings

The study evaluated the performance of the Location Model (LM), Linear Discrimination Function (LDP) and Quadratic Discriminant Function (QDF) for mixture of discrete and continuous variables. The findings are as follows:

a. The Location Model came first in terms of minimizing the average error rate (lower error rate) in the simulated experiment performed under two situations (a) and (b). In situation (a), q = 2, q = 4, q = 6, q = 7, q = 8, q = 9 and q = 10, while in situation (b) when q = 2, q = 3, q = 6, q = 7, q = 8 and q = 9.
b. The Linear Discriminant Function (LDP) came first in situation (a) when q = 3, and q = 5 while in situation (b) when q = 10
c. The Quadratic Discrimination Function came first in situation (b) when q = 4.
d. The result indicated that, when there is evidence of interaction between discrete variables and populations, the average error rate from linear discriminant function and quadratic discriminant function tend to give very poor result than the location model.
e. The result from the Location model and linear discriminant function also showed that proportion of error rate which is less than 30% is optimal in minimizing the probability of misclassification.
f. The result of the application of the real data revealed that the location model came first with lower error rate real data analysis.

5 Conclusions

The analysis shows that in classification of objects, losses in estimation procedure can cause disruption in the allocation role which may lead to some inflation of the error rate incurred in the experiment. The analysis and results obtained from this work indicated that the location model is considered good in terms of minimizing the average error rate. It gave better result than the other two classifications rule in the experiment conducted and performed better on real data analysis.

Author Contributions: All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

Hamid, H. H. (2010). A new a roach for classifying large number of mixed variables. *World Academy of Science Engineering & Technology*. Vol. 46, 10-22.

Krzanowski, W. J (1975a). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association.* Vol. 70, 782-790.

Krzanowski, W. J (1975b). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association.* Vol. 70, 782-790.

Krzanowski, W. J (1975c). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association.* Vol. 70, 782-790.

Mahat, N. I., Krzanowski, W.J., Hernandez, A. (2009). Strategies for non-parametric smoothing of the location model in mixed-variable discriminant analysis. *Modern Applied Science*. Vol. 3, No. 1, 151-163.

Wenbo, C. (2015). Quadratic discriminant analysis revisited. *Graduate Center City University of New York (CUNY).*