Fjrs.futa.edu.ng
FUTA Journal of Research in Sciences
ISSN: 2315 – 8239 (Print); E-ISSN: 2489 - 0413

FUTA Journal of Research in Sciences, Vol. 12 (2) 2016:299 - 306

ON LINEAR AND QUADRATIC DISCRIMINANT RULES

Olusola Samuel Makinde and Oluwatoyin Kikelomo Bodunwa

Department of Statistics, Federal University of Technology, Akure, Nigeria.
osmakinde@futa.edu.ng; okbodunwa@futa.edu.ng

ABSTRACT

In this study, we consider some theory of Bayes rule and discriminant analysis to establish asymptotics of linear and quadratic discriminant rules. Effect of deviation from property of spherical symmetry is investigated and optimal performances of discriminant rules are demonstrated using simulation and real life applications.
Keywords: linear classifier; quadratic classifier; optimal rule; symmetry; robust discriminant function.
AMS 2010 Mathematics Subject Classification: 62H30; 60E05

INTRODUCTION

Discriminant analysis is aimed at getting maximum information about separability or distinction among classes or populations while classification is aimed at assigning each observation to one of these populations on the basis of a vector of measurements. A good classification procedure is the one that classifies observations from unknown populations correctly (Makinde and Chakraborty, 2015).

Welch (1939) proposed Bayes classification rule, which assigns an observation to population with highest posterior probability of the population given some measurement on the observation. It follows from Welch (1939) that the ratio of log likelihood functions of the two normally distributed populations forms the theoretical basis for building discriminant function that best classifies new individuals into any of the two populations given that the prior probabilities of the populations are known. Wald (1944) defined expected cost of misclassification as the product of posterior probability of the population and the cost associated with misclassifying the observation, and then proposed assigning an observation to population or class that has the highest expected cost of misclassification.

Suppose there are two populations with equal covariance matrix (this case is referred to as location shift or homogenous scale), Fisher (1936) described the separation between these two populations to be ratio of variance between the populations to variance within the populations. This postulation leads to discriminant analysis, called Fisher's discriminant analysis. Suppose there are two populations from the same family of multivariate distributions to which observations can be classified. If these populations are normally distributed and have the same covariance matrix, the discriminant analysis is referred to as linear discriminant analysis (LDA). Similarly, if these populations are normally distributed but have different covariance matrices, the optimal rule is nonlinear and referred to as quadratic discriminant analysis (QDA). QDA can be seen as the problem of scale shift or location-scale shift, depending on whether

the populations have the same location vector or not. Welch (1939) and Wald (1944) showed that linear discriminant function has optimal properties for two group classification if the populations are multivariate normally distributed.

Many researchers have worked on the estimation of probability of misclassification given that observations are from multivariate normally distributed random samples or populations. The works include Anderson (1972), Das Gupta (1972), Krzanowski (1977), Chang and Afifi (2008). Many classification methods, both parametric and nonparametric, have been compared with LDA and QDA under normality and non-normality. These include comparison with depth based methods, nearest neighbour rules among others in the studies of Ghosh and Chaudhuri(2005), Kim et al.(2011) and Li et al.(2012), Dutta and Ghosh (2012), etc. Makinde (2016) studied the theoretical misclassification probability of linear and quadratic classifiers, examined the performance of linear and quadratic discriminant under distributional variations in theory and using simulation and derived Bayes errors for some competing distributions from the same family under location shift.

In this study, we consider some theory of Bayes rule and discriminant analysis for multiclass problem to establish asymptotics of linear and quadratic discriminant rules under normality conditions. Effect of deviation from property of spherical symmetry is investigated and optimal performances of discriminant rules are demonstrated using simulation for normal and non-normal populations. We compare the performance of these classification rules with support vector machine for some real life applications.

ROBUST DISCRIMINANT ANALYSIS

assign $\boldsymbol{x}$ *to* π_1 *if*

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 - 2\boldsymbol{x}) > 0 \qquad (2)$$

Consider J populations and having density function of the form

$$f_j(\boldsymbol{x}) = g((\boldsymbol{x} - \boldsymbol{\mu}_j)^{\intercal}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

$j = 1, 2, \ldots, J$, for some strictly decreasing, continuous, non-negative scalar function g, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are mean vector and covariance matrix of jth population respectively. Assuming normality and equal cost of misclassification for the J populations, Bayes rule can be defined as

assign $\boldsymbol{x}$ *to* π_j *if* $D_j(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_j)^{\intercal}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j) + log|\boldsymbol{\Sigma}_j| - log\, p_j$ *is minimum.*
(1)

where p_j is the prior probability of the jth population. This classification rule is known as quadratic discriminant analysis (QDA). It is linear (LDA) if all $\boldsymbol{\Sigma}_j$ are equal. LDA and QDA have the limitation that they cannot be used for discriminating between distributions whose first and second moments do not exist. Furthermore, the presence of an outlying training sample point will affect the performance of LDA and QDA. Hence, both linear and quadratic classifiers are not robust against outliers and extreme values. Hubert and Van Driessen (2004) proposed replacing the estimates of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ in equation (1) above by reweighted minimum covariance determinant (MCD) (Rousseeuw, 1984) estimator of multivariate location and scatter based on FAST-MCD algorithm of Rousseeuw and Van Driessen (1999).

Let us consider two classes for simplicity. Suppose π_j has multivariate normal distribution $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), j = 1, 2$. For $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ and $p_1 = p_2$, the classification rule in (1) can be expressed as

We note that if $\boldsymbol{\Sigma}$ is a constant multiple of identity matrix, classification in (2) is a generalisation of component-wise centroid

classifier in Hall, Titterington and Xue (2009).

One way of evaluating the performance of a classification rule is to calculate its misclassification probabilities. One can define the total probability of misclassification (Δ) as

$$\Delta = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

where $R_1 = \{\boldsymbol{x} \in \mathbb{R}^d : p_1 f_1(\boldsymbol{x}) \geq p_2 f_2(\boldsymbol{x})\}$ and $R_2 = \{\boldsymbol{x} \in \mathbb{R}^d : p_1 f_1(\boldsymbol{x}) < p_2 f_2(\boldsymbol{x})\}$.

The classification regions R_1 and R_2 can be constructed only when the distributions F and G are fully known. This will rarely be the case, we have to work with the empirical versions of the classification regions and calculate the error rates.

Invariance of discriminant functions under general affine transformation of the data

The distribution of a random variable $\boldsymbol{X}$ is said to be spherically symmetric about a parameter $\boldsymbol{\theta}$ if, for any orthogonal matrix $\boldsymbol{B}$,

$$\boldsymbol{X} - \boldsymbol{\theta} \overset{d}{=} \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{\theta})$$

The density function of any spherically symmetric distribution of a random variable $\boldsymbol{X}$, if it exists, is of the form $f(\boldsymbol{x}) \propto g((\boldsymbol{x} - \boldsymbol{\theta})^{\intercal}(\boldsymbol{x} - \boldsymbol{\theta}))$ for some nonnegative scalar function $g(.)$. Similarly, the distribution of a random variable $\boldsymbol{X}$ is said to be elliptically symmetric about $\boldsymbol{\theta}$ if there exists a $d \times d$ nonsingular matrix $\boldsymbol{A}$ such that $\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\theta})$ has a spherically symmetric distribution about $\boldsymbol{0}$. Liu (1990), Liu and Singh (1993), Liu, Parelius and Singh (1999) and Serfling (2006) suggested various ideas on multivariate symmetry.

Suppose a random vector $\boldsymbol{X}$ has a distribution F with mean vector $\boldsymbol{\mu}_F$ and covariance matrix $\boldsymbol{\Sigma}_F$, define $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ for some nonsingular matrix $\boldsymbol{A}$ and a constant vector $\boldsymbol{b}$. It is straightforward to show that

$$D_G(y) = (y - \mu_G)^{\intercal}\Sigma_G^{-1}(y - \mu_G) = (x - \mu_F)^{\intercal}\Sigma_F^{-1}(x - \mu_F) = D_F(x)$$

where G is the distribution of $\boldsymbol{Y}$, $\boldsymbol{\mu}_G$ and $\boldsymbol{\Sigma}_G$ are mean vector and covariance matrix of the distribution G respectively. The implication of this is that $D_F(\boldsymbol{x})$ is invariant under affine transformation of the data, meaning that this classifier cannot be affected by correlation existing among variables or features in the data.

NUMERICAL EXAMPLES

As illustration of actual error rates of LDA and QDA, we present some simulation study. As illustration of probability of misclassification, consider the following example. Let populations π_1 and π_2 be bivariate spherically symmetric with centres of symmetry $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and covariance matrices, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Assume that the prior probabilities of π_1 and π_2 are equal. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample from π_1 and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$, a random sample from π_2. We simulate a new random sample $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_m$ from π_1 and $\boldsymbol{Z}_{m+1}, \boldsymbol{Z}_{m+2}, \ldots, \boldsymbol{Z}_{2m}$ from π_2 and take sample sizes n and m to be 100. The simulation is repeated 1000 times. Here we consider different sample sizes with equal and unequal prior probabilities for some competing distributions. The distributions are bivariate normal distribution, bivariate Laplace distribution and bivariate t distribution with 3 degrees of freedom. Here we consider some cases:

Location shift with equal and unequal prior probabilities:

Consider $\boldsymbol{\mu}_1 = (0 \quad 0)^{\intercal}$, $\boldsymbol{\mu}_2 = (\delta \quad \delta)^{\intercal}$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2$, we can estimate probabilities of misclassification corresponding to LDA based on training samples. Figure1 shows the

apparent error rates of three bivariate spherically symmetric distributions, namely normal, Laplace and *t* with 3 degrees of freedom, as δ varies in $[0,2]$. As expected, the misclassification error is nearly 0.5 when $\delta = 0$ and it decreases as δ goes away from 0 and the separation between the population increases. The error rate is least in bivariate normal and higher in non-normal distributions.

Following the above numerical example, we consider the probability of misclassification of LDA accounting for effect of sample sizes and prior probabilities on the classification rule, and the result is presented in Table 1. It is shown in the table that the error rates increase as the prior probabilities of the competing classes differ. Considering various sample size, when $n = m = 20$, the sample sizes are regarded as small and the error rates are bigger than when m and n are large for each of the three families of distributions.

Table 1: Comparison of error rate of LDA for different sample sizes with equal and unequal prior probabilities for some competing distributions.

Distributions	Sample size	Error rates $\delta = 1$			$\delta = 2$		
		$p_1 = p_2$	$\frac{p_2}{p_1} = 0.5$	$\frac{p_2}{p_1} = 2$	$p_1 = p_2$	$\frac{p_2}{p_1} = 0.5$	$\frac{p_2}{p_1} = 2$
Bivariate normal	20	0.2507	0.2753	0.2765	0.0853	0.0920	0.0924
	50	0.2445	0.2694	0.2698	0.0814	0.0873	0.0877
	100	0.2417	0.2672	0.2672	0.0802	0.0860	0.0860
	200	0.2403	0.2661	0.2654	0.0796	0.0854	0.0854
	500	0.2403	0.2651	0.2654	0.079	0.0852	0.0852
Bivariate t with 3 d.f.	20	0.2934	0.3722	0.3711	0.1426	0.1739	0.1744
	50	0.2775	0.3786	0.3801	0.1347	0.1709	0.1704
	100	0.2725	0.3849	0.3851	0.1310	0.1679	0.1682
	200	0.2695	0.3892	0.3900	0.1291	0.1670	0.1670
	500	0.2671	0.3953	0.3948	0.1277	0.1671	0.1673
Bivariate Laplace	20	0.3241	0.4000	0.3976	0.1819	0.2045	0.2121
	50	0.3096	0.4000	0. 4000	0.1750	0.2045	0.2052
	100	0.3043	0.3996	0.3991	0.1721	0.2022	0.2014
	200	0.3027	0.4004	0.4004	0.1709	0.2009	0.2004
	500	0.3016	0.4005	0.4003	0.1704	0.1997	0.1998

We observe that there is slight difference in misclassification rates as n and m increase from 100 irrespective of competing distributions and prior probabilities, meaning that the error rates converge for large sample sizes

Location and scale shift with equal and unequal prior probabilities:

Consider a case where both mean vectors and covariance matrices of the competing classes differ. Let $\boldsymbol{\mu}_1 = (0 \quad 0)'$, $\boldsymbol{\mu}_2 = (2 \quad 2)'$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_2$ and $\boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$, then the probability of misclassification corresponding to QDA is presented in Table 2. We consider the probabilities of misclassification of QDA accounting for effect of sample sizes and prior

probabilities on the classification rule, and the result is presented in Table 2. It is shown in the table that the error rates increase as the prior probabilities of the competing classes differ. Also, error rates decrease as m and n become large for each of the three families of distributions. We observe that there is slight difference in misclassification rates as n and m increase from 100 irrespective of competing distributions and prior probabilities. Similarly, error rates are generally smaller in normal populations than non-normal populations.

Table 2: Comparison of error rate of QDA for different sample sizes with equal and unequal prior probabilities for some competing distributions.

Distributions	Sample size	Error rates					
		$\delta = 1$			$\delta = 2$		
		$p_1 = p_2$	$\frac{p_2}{p_1} = 0.5$	$\frac{p_2}{p_1} = 2$	$p_1 = p_2$	$\frac{p_2}{p_1} = 0.5$	$\frac{p_2}{p_1} = 2$
Bivariate normal	20	0.2038	0.2084	0.2313	0.0883	0.0868	0.0993
	50	0.1937	0.2028	0.2148	0.0799	0.081	0.0855
	100	0.1883	0.1977	0.2096	0.0782	0.0803	0.0828
	200	0.1863	0.1964	0.2057	0.0766	0.0785	0.0824
	500	0.1862	0.1954	0.2045	0.0755	0.0788	0.0818
Bivariate t with 3 d.f.	20	0.2145	0.2411	0.2168	0.1493	0.1522	0.1412
	50	0.2099	0.2437	0.1996	0.144	0.1534	0.1360
	100	0.2092	0.2508	0.1928	0.1455	0.1582	0.1353
	200	0.2125	0.2520	0.1886	0.1456	0.1597	0.1347
	500	0.2125	0.2604	0.1855	0.1479	0.1629	0.1349
Bivariate Laplace	20	0.2357	0.2590	0.2406	0.1381	0.1390	0.1419
	50	0.2233	0.2542	0.2236	0.1318	0.1389	0.1309
	100	0.2224	0.2536	0.2161	0.1309	0.1375	0.1307
	200	0.2227	0.2536	0.2151	0.1302	0.1368	0.1287
	500	0.2228	0.2532	0.2126	0.1297	0.1367	0.1273

Comparing LDA with support vector machine based on the setting above for location shift and scale shift. Table 3 presents the comparison between LDA and SVM. It is clearly shown in the table that the error rates of LDA are noticeably smaller than that of SVM for normal populations, which support the optimality of LDA in theory. For non-normal populations, it is clearly seen that LDA still outperforms the popular SVM.

Table 3: Comparison of error rates of LDA with SVM for some competing distributions.

Distributions	Error rates			
	$\delta = 1$		$\delta = 2$	
	LDA	SVM	LDA	SVM
Biv. normal	0.2417	0.327	0.0802	0.167
Biv. Laplace	0.3043	0.325	0.1721	0.184
Biv. t with 3 d.f.	0.2725	0.285	0.131	0.139

APPLICATION TO REAL LIFE DATA

We analyse three benchmark data sets to illustrate the performances of our methods (LDA and QDA) and compare them with support vector (SVM). These datasets include iris data (Fisher, 1936), Pima Indians diabetes (PID) data (owned by the National Institute of Diabetes and Digestive and Kidney Diseases) and biomedical data (Cox et al., 1982). All data sets are taken from UCI Machine Learning Repository (available at https://archive.ics.uci.edu/ml/datasets.html) and StatLib Archive (available at http://lib.stat.cmu.edu/datasets/), except iris data. In Table 4, we present information about real data set and present the analysis of the real data sets in Table 5. For computing MCD estimate of covariance using the FAST-MCD algorithm (Rousseeuw and Van Driessen, 1999) via R package robustbase, we set $\alpha = 0.90$. For SVM, we have used *C*-support vector classification with Gaussian RBF-kernel. The default choice of parameters in the R package kernlab was used. The cost of constraint *C* is taken to be 1 with hyper-parameter σ is determined automatically by the sigest function in the same library and it returns a value between the 0.1 and 0.9 quantile of $||\boldsymbol{x} - \boldsymbol{x}'||$.

Table 4: Information about real data set.

Datasets	Number of classes	Training sample size	Validation sample size	Dimension
Iris	3	30 each	20 each	4
Biomedical	2	50 each	17 each	4
Pima Indian Diabetes	2	100 each	100 each	4

Table 5: Estimated error rates of the classifiers in the real data examples

Datasets	Classifiers		
	LDA	QDA	SVM
Iris	0.033	0.017	0.067
Biomedical	0.206	0.147	0.176
Pima Indian Diabetes	0.26	0.28	0.275

For iris data and biomedical data, QDA appears to appear best. The reason for this better performance of QDA over LDA can be attributed to different scale existing among the groups of each of iris data and biomedical data. For Pima Indian diabetes data, LDA appears to perform best while others performs competitively. We observe that LDA and QDA perform well and competitively even when the distributional assumption is violated compared to the popular support vector machine.

CONCLUSION

Linear and quadratic classifiers for multiclass problem are considered in this paper, with emphasis on the performance of these classifiers under different prior probabilities of the competing populations and different sample sizes. The optimal performance of linear and quadratic discriminant functions under normality condition is investigated based on simulation and provide solutions of some theoretical examples. The performance of linear and quadratic classifiers for non-normal distributions is examined for different prior probabilities of the competing populations and different sample sizes under location and scale shift. The non-normal distributions are bivariate Laplace distributions and bivariate t distributions with the same degree of freedom. As a yardstick for measuring the effect of deviation of discriminant rule from normality assumption, LDA is compared with support vector machine and it is found out that LDA is robust against deviation from normality assumption.

REFERENCES

Anderson, T. W. (1972): Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions. Stanford University, Department of Statistics. Technical Report No 10.

Chang, P. C. and Afifi, A. A. (2008): Classification based on dichotomous and continuous variables. *Journal of American Statistical Association*, 69, pages 336 - 339.

Cox, L. H., Johnson, M. M. and Kafadar, K. (1982): Exposition of statistical graphics technology. ASA Proceedings of the Statistical Computation Section, pp. 55 – 56.

Das Gupta, S. (1972): Probability inequalities and error in classification. University of Minnesota, School of Statistics, Technical report, No 190.

Dutta, S. and Ghosh, A. K. (2012): On classification based on L_p depth with an adaptive choice of p. Technical Report No. R5/2011, Statistics and Mathematics Unit. Indian Statistical Institute, Kolkata, India

Fisher, R.A. (1936): The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179 - 188.

Ghosh, A. K. and Chaudhuri, P. (2005): On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32, 327 - 350.

Hall, P., Titterington, D.M. and Xue, J. (2009): Median based classifiers for high dimensional data. Journal of the American Statistical Association, 104(488), 1597-1608

Hubert, M. and Van Driessen, K. (2004): Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45, 301 - 320.

Kim, K. S., Choi, H. H., Mo on, C. S. and Mun, C. W. (2011): Comparison of k-nearest neighbour, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist- motion directions. *Current Applied Physics*, 11, pages 740–745.

Krzanowski, W. J. (1977): The performance of Fisher's linear discriminant function under non-optimal conditions, *Technometrics*, 19(2), pages 191–200.

Li, J., Cuesta-Alberstos, J. A. and **Liu, R. Y.** (2012). DD-Classifier: Nonparametric Classification Procedure Based on DD-plot. *Journal of American Statistical Association*, 107, pages 737–753.

Liu, R. Y.(1990): On a notion of data depth based on random simplices. *The Annals of Statistics*, *18*, 405 – 414

Liu, R. Y. and Singh, K. (1993): A quality index based on multivariate data depth and multivariate rank tests. *Journal of the American Statistical Association*, *88*, 252 – 260.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999): Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, *27*, 783 – 858.

Makinde, O. S. (2015): On Misclassification Probabilities of Linear and Quadratic Classifiers. *Afrika Statistika.* 11(1), 943 – 953.

Makinde, O. S. and Chakraborty, B. (2015): On Some Classifiers Based on Multivariate Ranks. In Nordhausen, K and Taskinen, S.(eds): Modern Nonparametric, Robust and Multivariate Methods, Festschrift in Honour of Hannu Oja. Springer, 249-264

Rousseeuw, P. J. (1984): Least Median of Squares Regression. *Journal of the American Statistical Association*, *79*, 871 - 880

Rousseeuw, P. J. and Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, *41*, 212 - 223.

Serfling, R. (2006): Multivariate symmetry and asymmetry, In Encyclopedia of Statistical Sciences, (S. Kotz, N. Balakrishnan, C. B. Read and B. Vidakovic, eds.), 8: 5338 - 5345. Second Edition, Wiley.

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

Welch, B. L. (1939): Note on discriminant functions. *Biometrika*, *31*, 218 - 220.