ED 432 591		TM 029 967

AUTHOR	Ferrer, Alvaro J. Arce; Wang, Lin
TITLE	Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods.
PUB DATE	1999-04-00
NOTE	23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Classification; Comparative Analysis; *Discriminant Analysis; *Nonparametric Statistics; *Regression (Statistics)
IDENTIFIERS	Accuracy; Logistic Regression

ABSTRACT

This study compared the classification performance among parametric discriminant analysis, nonparametric discriminant analysis, and logistic regression in a two-group classification application. Field data from an organizational survey were analyzed and bootstrapped for additional exploration. The data were observed to depart from multivariate normality; neither the group sizes in the sample nor the covariance matrices of the two groups were equal. A crossed design of classification function by prior probability was implemented for over 244 bootstrap samples. The classification error rates for each group and the total sample were gathered for each cell of the design matrix. The major findings of this study are: (1) nonparametric discriminant functions and logistic regression performed below expectations from theory; (2) the choice of prior probabilities influenced the classification performance for the smaller and the larger group, but not for the total sample; and (3) minimization of error rates for one group implied an increment in the error rate for the other group, or vice versa. The findings do not demonstrate the expected theoretical strength of nonparametric discriminant functions when applied to data with nonnormality and unequal covariance matrices. No consistent superiority was observed in logistic regression and quadratic discriminant function over the linear discriminant function. This indicates a more complicated situation than that portrayed in previous studies on the applications of discriminant functions and logistic regression for classification purposes. (Contains 7 tables and 22 references.) (Author)

COMPARING THE CLASSIFICATION ACCURACY AMONG NONPARAMETRIC, PARAMETRIC DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION METHODS

Alvaro J. Arce Ferrer

University of Iowa/ Autonomous University of Yucatan

Lin Wang

ACT

Paper presented at the 1999 Annual meeting of the American Educational Research Association, April 13 – 17, Montreal, Canada.

ABSTRACT

This study compared the classification performance among parametric discriminant analysis, nonparametric discriminant analysis, and logistic regression in a two-group classification application. Field data from an organizational survey was analyzed and bootstrapped for additional exploration. The data were observed to depart from multivariate normality; neither the group sizes in the sample nor the covariance matrices of the two groups were equal. A crossed design of classification function by prior probability was implemented over 244 bootstrap samples. The classification error rates for each group and the total sample were gathered for each cell of the design matrix. The major findings of this study are: 1) Nonparametric discriminant functions and logistic regression performed below expectations from theory; 2) the choice of prior probabilities influenced the classification performance for the smaller and the larger group, but not for the total sample; and 3) minimization of error rates for one group implied an increment in the error rate for the other group, or vice versa. The findings do not demonstrate the expected theoretical strength of nonparametric discriminant functions when applied to data with nonnormality and unequal covariance matrices. No consistent superiority was observed in logistic regression and quadratic discriminant function over the linear discriminant function. This indicates a more complicated situation than that portrayed in previous studies on the applications of discriminant functions and logistic regression for classification purpose.

Discriminant analysis and logistic regression methods have been frequently used in educational research either to classify individuals or organizations into different groups or to predict their group membership using a set of measures. Although the two methods appear to have the same utility, they are two different models. Generally speaking, discriminant analysis is part of the general linear model (Knapp, 1978; Thompson, 1991), whereas logistic regression models the nonlinear probabilistic function of a binary dependent variable, using a single or a set of independent variables. Readers who are not yet familiar with these two methods may wish to go over the section of the brief review of two methods in Fan and Wang's (1998) paper.

Both parametric and nonparametric models are can be used in discriminant analysis. The studies comparing the performance between discriminant analysis and logistic regression have typically focused on parametric discriminant analysis and logistic regression (Dattalo, 1994; Fan and Wang, 1998; Harrell & Lee, 1985; Wilson and Hardgrave, 1995). Using parametric discriminant analysis requires three assumptions: multivariate normal distribution in each group, equal covariance structure in each group, and reasonably large sample size (Johnson and Wichern, 1988). When these assumptions are met, parametric discriminant analysis can be expressed in the same form as logistic regression (Clearly and Angel, 1984; Harrell & Lee, 1985). Empirical investigations using real data and simulation data have also reported that the two methodologies give similar results when the assumptions are basically met (Fan and Wang, 1998; Meshbane and Morris, 1996).

In Educational research, situation often arises when a desired distribution of data such as normality is not possible, or the group covariance structures are different. Where this situation occurs, nonparametric discriminant analysis should be a more logical choice since it does not require distributional assumptions (SAS, 1988). One particular nonparametric discriminant method used in this study is the k-nearest-neighbor method. This method uses the pooled covariance matrix to calculate the Mahalanobis distances and generates a classification criterion. The choice of k is said to be relatively uncritical; one can practically try different k values and selects the one that yields the best classification result (Hand, 1982; SAS Institute, 1988; Silverman, 1986). For some unknown reasons, very little is found in literature that has compared the performance of nonparametric discriminant analysis with that of parametric discriminant analysis or logistic regression.

The purpose of this study was to use classification error rates as the criteria to compare the classification performance among parametric discriminant analysis, nonparametric discriminant analysis, and logistic regression for a two-group classification problem.

Methods

Design

A crossed two-factor design was implemented in analyzing the classification error rates for the total sample and for each of the two groups. This design permits a systematical assessment of the contribution by each factor to reducing the explained variance of the classification error rates. The first factor is the classification method that consists of parametric discriminant analysis, nonparametric discriminant analysis, and logistic regression. For parametric discriminant analysis, both a linear function and a quadratic function are included. Three k values (4, 5, 6) are used in the k-nearest-neighbor method in the nonparametric discriminant analysis. Therefore, the factor of classification method actually contains six levels: linear discriminant analysis, quadratic discriminant analysis,

4-nearest-neighbor method, 5-nearest-neighbor method, 6-nearest-neighbor method, and logistic regression.

The second factor is the prior probability (two levels: equal priors, and priors estimated from the sample). Priors are used to help derive a reasonable classification rule or criterion. When prior probabilities are unknown, they are set to equal (.50 to .50). In practice, the ratio of the two group sizes is often taken as the ratio of the their respective populations. This ration is then used as priors in finding the classification rule. As is described later, the data for this study contains two groups and their sizes are approximately .30 to .70. Therefore, these two priors (.50 to .50 and .30 to .70) were adopted in the design.

Given the nature of this study and for simplicity, costs of misclassification were assumed to be equal in discriminant analysis. Also, sample size was not manipulated in the design; the achieved sample size (n = 244) from the field was used instead. Model fitting was performed over 244 bootstrap samples. The bootstrap procedure was included to evaluate generalizability of the results across different configurations of subjects (Thompson, 1993).

Data Source

The data for this study was taken from an organizational study in an international project in which human resources managers in various organizations in a North American country were surveyed. The organizations were a priori classified, using certain benchmarks, as high performance organizations or otherwise. The sample was not strictly a random sample of the two types of organizations in that country due to voluntary participation even though the initial sampling plan was meant to select a random sample. The achieved sample contained responses from 244 organizations.

For this study, all the 244 organizations were included. Seventy-one (29.1%) were high performance organizations, 171 (70.9%) were not. The knowledge that an organization was either a high performance one or otherwise was used to define the criterion, or dependent variable. Eight predictor variables were constructed from the responses to the survey items. In this data, the number of variables, the distributions of the variables (Table 1), and the sample size are close to those frequently found in a typical classification studies (Meshbane and Morris, 1996; Huberty and Curry, 1978). The distributions of the variables in Table 1 and the covariance matrices in Table 2 suggest the violation of the assumptions, such as multivariate normality of predictors and equal covariance matrices in the group populations, that are often invoked in classification analyses (Johnson and Wichern, 1988).

In preprocessing the data, the eight predictor variables were standardized to reduce their dispersion resulting from the unequal number of items in each survey section. No theoretical reason existed to weight the contribution of each item differently.

Insert Table 1 and Table 2 about here

Following some common advise in classification theory (Jonhson and Wichern, 1988), the assumptions of multivariate normality and equal covariance structure of the two groups were assessed. Multivariate normality was assessed by evaluating the distribution of each variable first. It is known that multivariate normality is not possible if univariate normality is not observed in the distribution of each individual variable. In Table 1, under skewness and kurtosis, predictors in both groups (called Class 1 for the high performance group, and Class 2 for the other group) exhibit varying degrees of departure from the expected values (0s) for the normal distribution. The skewness ranged from -0.21 to 2.57 and the kurtosis ranged from -1.36 to 6.35. Across most predictors, values of the skewness and kurtosis for class 1 are relatively closer to those for the normal distribution than those for class 2. However, Table 1 also indicates that one predictor, $\underline{x}1$, in Class 2 shows the greatest departure from normality.

Table 2 shows that the variances of the predictors in Class 1 were larger than those in Class 2. The median difference between the variances of the two classes is about 37% and this indicates some overlapping at the tails of the two class distributions. The degree of inequality of the class covariance matrices was evaluated using the Bartlett-Box test of equality of population covariance matrices (Tatsuoka, 1988). The data supported the rejection of the null hypothesis of equal population covariance matrices (chi-square = 90.94; df=36, p<0.05).

Bootstrap samples and cross-validation schema

Each condition of the design was replicated over 244 random configurations of cases. Each configuration represents a sample drawn with replacement from the original data set. The cases from each bootstrap sample were randomly assigned to a training and a testing sample using the 2/3 - 1/3 rule as specified by the general cross-validation schema (Weiss and Kulikowski, 1991). While the leaving-one-out approach is a preferable technique (Lachenbruch, 1967; Huberty, 1994; Johnson and Wichern, 1988), it is not only computationally expensive, but may also produce estimates of error rate with high variance for small samples. Therefore, only the general cross-validation schema was chosen.

The training data set was defined to calibrate the functions, and the testing data set was reserved strictly for testing the classification rule derived from the training data. Using independent testing samples may reduce the upward bias of classification error rates that would otherwise be present if cross-validation had also used the training samples (Huberty and Curry, 1978; and Meshbane and Morris, 1996). The above procedure was implemented using SAS interactive matrix language (IML).

Model fitting

The parametric and nonparametric discriminant functions and logistic regression function were fit to each training sample using the two prior probabilities. The parametric discriminant functions were fit using SAS PROC DISCRIM and the linear and quadratic classification rules. The nonparametric discriminant function was fit by requesting the nonparametric method in the SAS PROC DISCRIM procedure, and invoked three classification rules, i.e., the 4th, 5th, and 6th nearest neighbor rules. Finally, the logistic regression function was fit with SAS PROC LOGISTIC procedure and the prior probability for Class 1 (the high performance group) was specified for the classification.

The classification error rates for each combination of the levels of the two factors were estimated from the testing data. For both parametric and nonparametric discriminant functions, SAS PROC DISCRIM allows choosing a data set as the testing sample to cross-validate error rates, but this flexibility is not available in PROC LOGISTIC. Cross-validation error rates for the logistic functions were obtained following these four steps. First, we estimated logistic regression parameters from a merged data set (i.e., training and testing data sets) in which the testing portion of the data was copied into another variable and replaced with missing values. Second, we got an estimate of the probability of observing Class 1 given the set of predictors, for each case in the testing data. Third, Class 1 prior probability was used to determine whether observations in the testing set belong Class 1 or Class 2. Fourth, we cross-tabulated the predicted class and the observed class. A classification error occurred when there was a mismatch between the expected class and a predicted class. The classification error rates for each class and the total were collected for later analyses.

Results and Discussions

Comparing predictive classification error rates among models

Tables 3, 4, and 5 summarize the classification error rates for Class 1 (the smaller group), Class 2 (the larger group), and the total (both groups). The marginal distributions of the classification functions in Table 3 show that, independent of the size of the prior probability, the error rates fall in three groups. The lowest error rate group contains only the logistic regression with a mean error rate of 0.33. The second group includes the two parametric functions, with the mean error rates of 0.44 and 0.40 for the linear and quadratic functions, respectively. The last group has the k-nearest-neighbor nonparametric discriminant methods. These methods yielded the largest mean error rates, ranging from 0.45 to 0.49.

The marginal error rates in Table 3 also indicate that predicting Class 1 membership using the set of 8 predictor variables performed better than chance. For example, in a training sample, the average proportions of Class 1 and Class 2 are 0.30 and 0.70. By pure chance, one would expect to predict Class 1 membership with a 30% success rate, which means a 70% error rate. None of the marginal error rates in Table 3 is larger than 50%.

The results of predicting Class 2 membership in Table 4 portray a somewhat different pattern in terms of the marginal error rate distributions of the three methods. The lowest error rate group includes the two parametric discriminant functions. The nonparametric method group has the medium error rate; the highest error rate is from the logistic regression function. As mentioned above, the expected error rate in predicting Class 2 membership would be 30%. Again, none of the observed marginal error rates is greater than 30%, although the logistic regression error rate (28%) is very close to the expected error rate. The differences in the marginal error rates in Table 3 and Table show that the loss of classification accuracy for one class can turn into a gain in the classification accuracy for the other class. For example, the logistic regression did the worst in predicting Class 2 membership, but had the best accuracy in predicting Class 1 membership. Similar results were also reported by Fan and Wang (1998). Finally, in Table 5, all the classification functions are found to have performed much alike in terms of the marginal error rates for predicting both Class 1 and Class.

Insert Tables 3, 4, and 5 about here

Comparing predictive classification error rates between the two priors

The bottom rows in Table 3, Table 4, and Table 5 give the marginal distributions of the predictive classification error rates for Class 1, Class 2, and both classes, respectively, using two prior probabilities (.50 and .30 for Class 1). The results suggest that the size of a prior may affect classification error rates for predicting each class, but not for predicting both classes. In Table 3, the error rate using a prior of .50 was smaller than that using a prior of .30. This indicates a positive effect of choosing a prior probability around the middle point of the distribution. However, the logistic regression error rate was in fact smaller when using .30 instead of .50 for the prior probability. This suggests that, when information about group sizes is used to estimate priors, the logistic regression method may do a better job than both the parametric and nonparametric discriminant functions. The opposite is true in predicting Class 2 where the two discriminant functions outperformed the logistic regression method when the sample proportion of Class 2 (.70) was used as the prior (Table 4).

The decision between using equal priors or using sample information to estimate priors seems to be less relevant when researchers are concerned with the overall predictive accuracy. In Table 5, the marginal error rates are the same for the two prior probabilities.

When priors are unknown, there are two approaches that researchers often utilize. The first approach is based on the concept of sampling and advocates the use of proportions of each category as estimates of the population priors. The second approach is less systematic and it requires to assume that the population priors are the same for each class. There is no total agreement on which one should be used (Lindeman, Merenda, and Gold, 1980). Most of the literature supports using equal priors except when there is enough confidence in the accuracy of prior probabilities estimated from sample data (Huberty, 1994; Johnson and Wichern, 1988).

The results presented above show that a wrong choice of prior may unduly increase the error rate of most functions. In this study, for example, when the interest is in classifying members of the smaller group (i.e., Class 1), choosing a prior proportional to the sample size generally increased error rates. This did not, however, apply to the logistic regression method. This finding may have some practical significance in educational, behavioral, and psychological research where correct identification of a small-size special population is very important (Fan & Wang, 1998). On the other hand, when predicting Class 2 membership, using equal priors generally reduces error rates. Finally, choice of priors does not affect the overall predictive accuracy. Therefore, when the concern is the correct classification of members of both classes, it makes little difference to assume equal priors or to estimate them from the sample.

Comparing Predictive Classification Error Rates Across Functions and Priors

Theoretically, the classification results from the nonparametric discriminant functions and the logistic regression function should be less sensitive to departures from normality and inequality of covariance matrices. For the parametric discriminant functions, normality is required of both linear and quadratic functions. Equality of covariance is, however, not necessary for the quadratic discriminant function (Anderson, 1984; Huberty, 1994; Johnson and Wichern, 1988). With this in mind, the following situations can be expected.

First, when prior probabilities are assumed to be equal, the relative efficiency of the parametric and nonparametric discriminant functions and logistic regression function would be alike. However, when data does not satisfy various assumptions, the quadratic function may outperform the linear discriminant function and, in turn, the nonparametric discriminant (i.e., the k-nearest-neighbor) method and logistic regression function may do better than the parametric quadratic discriminant function. Between the nonparametric discriminant function and the logistic regression function, little is found in literature that compares their performances in classification. It is therefore difficult to set expectations about their performance comparison. Intuitively, the nonparametric discriminant method may have an edge over the logistic regression, because the nonparametric rule assigns observations based on their closeness to groups of observations and requires no ordering of cases.

Second, the expectations given above also stand when prior probabilities are estimated from the sample proportions of the groups and the proportions are not .50. In addition, the superiority of the logistic regression over parametric discriminant functions would be more evident. Namely, the logistic regression function is expected to perform relatively better than the linear and quadratic discriminant functions because the logistic regression is believed to be more precise in modeling the extreme regions in a probabilistic function (Hosmer, 1989; Huberty, 1994; Dattalo, 1994).

The results in Table 3 support some, but not all of the expectations just mentioned. What is surprising is that the logistic regression performed below expectations and had the largest error rate of all the functions. This may be due to the overlapping of the two classes in the region where the logistic regression probabilistic function reaches 0.5. It is also noticed that, when using the sample proportion as the prior, the quadratic discriminant function performed relatively better than the linear discriminant function, but both had higher error rates than the logistic regression method. The nonparametric discriminant functions performed very poorly when compared with both parametric discriminant functions and the logistic regression.

Table 6 presents the percentage improvement (or deterioration) of the classification error rates for each function when changing priors from equal to sample size proportional. When decisions are targeted to classify members of the smaller group (i.e., Class 1), it is observed that, except for the logistic regression, the classification error rates increased to different degrees for the other functions. The k-nearest-neighbor methods were found to be most sensitive to changes in priors. The next most sensitive function was the linear discriminant whose error rate went up about 67%. The similar poor performance by the nonparametric and linear discriminant functions may be due to the use of pooled covariance in estimation of the distance. Because the current data set involves classes with different covariance matrices, pooling the covariance ignored this fact when computing the Malahanobis' distance, thus probably biasing the estimates of distance.

In Table 6, when classifying members of the smaller class (i.e., Class 1), the error rate increased about 25% for the quadratic discriminant function but decreased about 39% for the logistic regression classification error rate. This difference supports the expectation about the superiority of the logistic regression in classifying observations in the extreme regions of a probabilistic function. The other results in Table 6 show agreement with the discussions given earlier on Tables 3, 4, and 5 with regard to the performances of the different classification methods.

Insert Table 6 about here

Sources of variation in the classification error rates

To summarize the contribution of the factors considered in this study on the category error rate and overall error rate, the variance of error rates was partitioned using the analysis of variance method in a two-by-two orthogonal design. Table 7 presents the results of this analysis.

For Class 1 (the smaller group), it is observed that the design factors accounted for 50% of the variation in the classification error rate. Of the 50% variance, the prior probability contributed the 21.11%. The classification function contributed only 8.63%. The interaction between the discrimination method and the prior probability contributed to the reduction of 20.22% of the variance in the classification error rate. This agrees with the findings discussed earlier that the relative performances of the function is related to the prior sizes. For Class 2, the larger group, the design factors explained 46.20% of the variance in the error rate. The interaction between the discriminant function and the prior probability also accounted for 28.5% of the variance out of the total 46.20% of the variance for Class 2. For both classes, only the interaction between prior probability and classification function explained most of the variance in the error rate. As was mentioned earlier, the prior probability contributed little to the explained variance in the total error rate.

Insert table 7 about here

Limitations of this study and suggestions for future research

This research has some limitations. First, the use of secondary data did not allow us to separate the effect of nonnormality from inequality of covariance. Future studies using either experimental data or simulated data may help solve this problem. Second, the size of the data set and the method chosen to estimate classification error rate may affect the performance of the logistic regression. Particularly, the chosen method removed cases from the estimation and then used them in assessing the predictive accuracy of the classification methods. With moderate sample sizes, this method leaves one with either insufficient number of cases for the training sample or for the testing sample. Third, the efficiency of the parameter estimation method that the logistic regression utilizes is based on large sample sizes. The size of the samples might influence the accuracy of the estimates of the classification error rates for the logistic regression. It is desirable to replicate this study using larger sample sizes. Finally, the quality of the predictor variables could be another source of interest to control in other studies.

Summary and Conclusions

This study examined the effects of parametric (linear and quadratic discriminant functions), nonparametric discriminant functions, and logistic regression function on their classification error rates for a two-group classification problem. To evaluate stability of classification results, 244 bootstrap samples were drawn from the original sample. For each combination of prior probability and classification functions, classification rules were estimated using a set of training samples. The cross-validation approach used a fixed percentage of cases for the training and the testing samples. Classification error rates for each class as well as for both classes were analyzed.

The following findings were obtained from this study:

1. The nonparametric discriminant function did not perform as expected and, in several cases, it performed worse than the parametric discriminant functions. The superiority of the parametric discriminant function over the nonparametric function was more noticeable when priors were estimated from sample sizes. However, when equal priors were used in predicting the smaller class (Class 1), parametric and nonparametric discriminant functions performed more or less alike. Given the complexity in application, the nonparametric discriminant function used in this study (k-nearest neighbor method) might not be a good alternative for use with nonnormal data and unequal group covariance matrices in a two-group classification problem .
2. The superiority of the logistic regression was not impressive in this study. The logistic regression and the nonparametric discriminant functions performed somewhat similarly under certain combination of prior probabilities and targeted class. Under other combinations, the logistic regression outperformed nonparametric discriminant function.
3. Classification error rates for linear and quadratic functions were relatively close. In several design cells, however, the linear classification rule yielded smaller error rates than the quadratic rule.
4. The results from this study helped to realize the complexity of the dynamics in the classification

process. Classification literature typically focuses on model assumptions (Johnson and Wichern, 1988). However, sizes of prior probabilities and a priori selection of a class are two additional factors to be considered when evaluating the performance of parametric classification methods and the logistic regression (Fang and Wang, 1998; Huberty, 1994; Press and Wilson, 1978; Wilson and Hargrave, 1995).

References

Anderson, T. W. (1984). An introduction to multivariate statistical analysis (2nd ed.). New York: Wiley.

Cleary, P. D., & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. Journal of Health and Social Behavior, 25, 334-348.

Dattalo, P. (1994). A comparison of discriminant analysis and logistic regression. Journal of Social Service Research, 19, 121-144.

Fan, X., & Wang, L. (1998). Comparing linear discriminant function with logistic regression for the two-group classification problem. Paper presented at the 1998 Annual Meeting of the American Educational Research Association, San Diego, CA.

Hand, D. J. (1982). Kernel discriminant analysis. New York: John Wiley & Sons, Inc.

Harrell, F. E., & Lee, K. L. (1985). A comparison of the discrimination of discrimnant analysis and logistic regression under multivariate normality. In P. K. Sen (Ed.), Biostatistics: statistics in biomedical, public health and environmental sciences (pp. 333-343). Amsterdam: Elsevier Science Publishers B.V. (North Holland).

Hosmer, D. W. (1989). Applied logistic regression. New York: Wiley.

Huberty , C. J., & Curry, A. R. (1978). Linear versus quadratic multivariate classification. Multivariate behavioral research, 13, 237-245.

Huberty, C. J. (1994). Applied discriminant analysis. New York: Wiley

Johnson, R. A., & Wichern, D. W. (1988). Applied multivariate statistical analysis (2nd ed.) Englewood Cliffs, New Jersey: Prentice Hall.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 23, 639-645.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). Introduction to bivariate and multivariate analysis. Glencoe, IL: Scott, Foresman.

Meshbane, A., & Morris, J. (1996). Journal of Experimental Education, 63, 263-273.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. Journal of American Statistical Association, 73, 699-705.

SAS Institute Inc. (1988). SAS/STAT user's guide (Release 6.03 ed.). Cary, NC: Author.

SAS Institute Inc. (1997). SAS/STAT software: Changes and enhancements through release 6.12. Cary, NC: SAS Institute, Inc.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. New York: Chapman and Hall.

Tatsuoka, M. J. (1988). Multivariate analysis: Techniques for educational and psychological research (2nd ed.). New York: Macmillan.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. The Journal of Experimental Education, 61, 361-377.

Weiss, S. M., & Kulikowski, C. (1991). Computed systems that learn. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Wilson, R. L., & Hardgrave, B. C. (1995). Predicting graduate student success in MBA program: Regression versus classification. Educational and Psychological Measurement, 55, 186-195.

Table 1

Descriptive Statistics of the Eight Predictor Variables.

	Class 1 (n = 71)								Class 2 (n = 173)							
	x1	x2	x3	x4	x5	x6	x7	x8	x1	x2	x3	x4	x5	x6	x7	x8
Mean	-0.24	0.47	0.67	0.51	0.51	0.49	0.55	0.41	-0.60	-0.19	-0.27	-0.21	-0.21	-0.20	-0.23	-0.17
SD	1.03	1.04	1.07	1.01	0.91	1.17	1.02	1.09	0.64	0.92	0.83	0.92	0.96	0.85	0.90	0.91
Skewness	0.65	-0.16	0.89	0.38	0.18	0.31	-0.21	0.17	2.57	0.78	1.10	0.40	0.35	0.46	0.43	1.19
Kurtosis	-1.28	-1.29	1.03	0.24	0.64	-0.51	-0.21	-1.36	6.35	-0.69	0.84	-0.62	-0.42	-0.49	-0.60	0.74

Table 2

Covariance Matrices of the Eight Predictor Variables (n = 244)

	x1	x2	x3	x4	x5	x6	x7	x8
Class 1	1.060	0.653	0.360	0.442	0.299	0.404	0.454	0.763
	0.653	1.080	0.320	0.680	0.389	0.484	0.548	0.610
	0.360	0.320	1.140	0.495	0.357	0.315	0.320	
	0.442	0.680	0.495	1.020	0.486	0.639	0.493	0.321
	0.299	0.389	0.357	0.486	0.819	0.638	0.460	0.324
	0.404	0.484	0.357	0.639	0.638	1.369	0.634	0.211
	0.454	0.548	0.315	0.493	0.460	0.634	1.036	0.337
	0.753	0.610	0.320	0.321	0.324	0.211	0.337	1.203

	x1	x2	x3	x4	x5	x6	x7	x8
Class 2	0.416	0.298	0.248	0.243	0.258	0.229	0.255	0.421
	0.298	0.843	0.407	0.369	0.385	0.338	0.374	0.355
	0.248	0.407	0.686	0.342	0.312	0.239	0.303	0.270
	0.243	0.369	0.342	0.844	0.405	0.267	0.339	0.244
	0.258	0.385	0.312	0.405	0.927	0.386	0.445	0.352
	0.229	0.338	0.239	0.267	0.386	0.717	0.384	0.200
	0.255	0.374	0.303	0.339	0.445	0.384	0.817	0.310
	0.421	0.355	0.270	0.244	0.352	0.200	0.310	0.826

Table 3

Classification Error Rates for Predicting Class 1 (the smaller group)

	Priors					
	Equal (0.50)		Proportional (0.30)			
Function					Marginal	
Parametric Discriminant						
Linear	0.33	(0.12)	0.55	(0.13)	0.44	(0.17)
Quadratic	0.36	(0.12)	0.45	(0.13)	0.40	(0.13)
Nonparametric Disarm.						
4 nearest neighbor	0.35	(0.13)	0.63	(0.12)	0.49	(0.19)
5 nearest neighbor	0.32	(0.12)	0.57	(0.14)	0.45	(0.18)
6 nearest neighbor	0.31	(0.12)	0.65	(0.13)	0.48	(0.21)
Logistic Regression	0.41	(0.13)	0.25	(0.15)	0.33	(0.16)
Marginal	0.35	(0.13)	0.51	(0.19)		

Note: Each table entry is the mean classification error rate. The standard deviation of the classification error rate is in the parentheses. Classification error rates are based on 244 random samples with replacement from a finite population.

Table 4

Classification Error Rates for Predicting Class 2 (the larger group)

Method (M)	Priors Equal (0.50)		Proportional (0.70)		Marginal	
Parametric Discriminant						
Linear	0.23	(0.07)	0.12	(0.06)	0.18	(0.08)
Quadratic	0.21	(0.07)	0.15	(0.06)	0.18	(0.07)
Nonparametric Disarm.						
4 nearest neighbor	0.24	(0.07)	0.18	(0.07)	0.21	(0.07)
5 nearest neighbor	0.27	(0.09)	0.14	(0.06)	0.20	(0.10)
6 nearest neighbor	0.28	(0.09)	0.16	(0.07)	0.22	(0.10)
Logistic Regression	0.19	(0.10)	0.38	(0.10)	0.28	(0.14)
Marginal	0.24	(0.09)	0.19	(0.11)		

Note: Each table entry is the mean classification error rate. The standard deviation of the classification error rate is in the parentheses. Classification error rates are based on 244 random samples with replacement from a finite population.

Table 5

Classification Error Rates For Predicting Both Classes

	Priors		
	Equal	Proportional	
Method (M)	(0.50 : .50)	(0.30 : 0.70)	Marginal
Parametric Discriminant			
Linear	0.28 (0.06)	0.25 (0.06)	0.26 (0.07)
Quadratic	0.29 (0.07)	0.24 (0.05)	0.26 (0.06)
Nonparametric Disarm.			
4 nearest neighbor	0.29 (0.07)	0.31 (0.06)	0.30 (0.06)
5 nearest neighbor	0.29 (0.06)	0.26 (0.05)	0.28 (0.06)
6 nearest neighbor	0.29 (0.06)	0.30 (0.05)	0.30 (0.06)
Logistic Regression	0.24 (0.6)	0.34 (0.07)	0.29 (0.08)
Marginal	0.28 (0.09)	0.28 (0.07)	

Note: Each table entry is the mean classification error rate. The standard deviation of the classification error rate is in the parentheses. Classification error rates are based on 244 random samples with replacement from a finite population.

Table 6

Percentage of Improvement (or Deterioration) on the Classification Error Rates of Six Functions when Changing from Equal to Proportional Priors

	Classes		
Function	Class 1	Class 2	Both
Parametric Discriminant			
Linear	-67.0	48.0	11.0
Quadratic	-25.0	29.0	17.0
Nonparametric Disc.			
4 nearest neighbor	-80.0	25.0	-7.0
5 nearest neighbor	-78.0	48.0	10.3
6 nearest neighbor	-110.0	43.0	-3.0
Logistic Regression	39.0	-100.0	-42.0

Table 7

Percentage of Variance Partitioning for Classification Error Rates

	Classes		
Source	Class 1	Class 2	Both
Total R-sq.	50.0	46.2	18.7
Prior Probability (P)	21.1	5.5	0.0
Function (F)	8.6	12.2	5.9
P * F	20.3	28.5	12.8

Note: table entries are the ETA squared, which was computed by the ratio of each source sum of squares and the total sum of squares.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029967

REPRODUCTION RELEASE

(Specific Document)

AERA

I. DOCUMENT IDENTIFICATION:

Title: COMPARING THE CLASSIFICATION ACCURACY AMONG NONPARAMETRIC, PARAMETRIC DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION METHODS

Author(s): Alvaro J. Arce Ferrer, and Lin Wang

Corporate Source: University of Iowa/Autonomous University of Yucatan ACT

Publication Date: April 1999

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Lin Wang

Printed Name/Position/Title: Lin Wang, Research Associate

Organization/Address:
Work Keys, ACT - 96
2201 N. Dodge St., P.O. Box 168
Iowa City, IA 52243

Telephone: (319) 337-1906

FAX:

E-Mail Address: WANGL@ACT.ORG

Date: 6-9-99

(over)