BULL. AUSTRAL. MATH. SOC. 62H30
VOL. 37 (1988) [313–315]

DISCRIMINANT ANALYSIS WITH CORRELATED TRAINING DATA

CHARLES RICHARD OGIKI LAWOKO

In this thesis consideration is given to the problem of allocating an object to one of two groups on the basis of measurements on the object. The performance of several sample discriminant functions are considered in terms of their associated errors of allocation. Although the performances of these discriminant functions have been studied extensively in the past, most of these studies have concentrated on the situation where the training observations from each possible group are independently distributed. These discriminant functions are studied here under more general conditions. Initial attention is focused on Anderson's sample linear discriminant function W. The performance of this discriminant function is investigated in the situation where the training observations within each group are correlated. Models of intraclass correlation which can be represented by time series processes are considered. In particular, autoregressive moving average processes of various levels of complexity are used. It is concluded from both theoretical and simulation studies that positive correlation among the training observations increases the error rates beyond their expected values in the case of independent observations. On the other hand, when training observations are negatively correlated, the error rates associated with W tend to be less than those for independent training data.

The error rates associated with the sample linear discriminant function are usually unknown in practice and therefore have to be estimated. The performances of the estimators of the error rates are thus investigated as well, and the behaviour of the estimated error rates under intraclass correlation is compared with their behaviour when the training observations are independent. It is shown that, similarly to the results for independent training observations, the plug-in estimator of the actual error rate gives too favourable an estimate of the actual error rate. The optimism, however is greater with positively correlated observations than with independent observations because the effect of positive correlation is to increase the expectation of the actual error rate and to decrease the expectation of the plug-in estimator.

Received 6 July 1987. Thesis submitted to University of Queensland, September 1986.Degree approved April 1987. Supervisor: Associate Professor G.J. McLachlan

The error rates are studied mainly through asymptotic expansions because the exact distribution of the sample linear discriminant function is extremely complicated even for independent training observations. Wherever possible, the applicability of the asymptotic expansions for small sample sizes is assessed through simulation experiments. Results from these experiments show that the asymptotic expansions provide reasonable approximations to the true error rates for the combinations of parameters considered.

The other allocation statistic which is investigated is Z, the quadratic discriminant function formed from the likelihood ratio criterion. The performance of the statistic Z is studied in the situation when the training observations are correlated and compared with the performance of W. In particular,the performance of the statistics W and Z relative to each other under intraclass correlation is compared with their relative performance when the training observations are independent. It is concluded that neither Z nor W is absolutely superior. It is found that their relative performance depends on the extent of the correlation among the training observations and the size of the separation between the classes, as measured by the Mahalanobis distance between them. However, on the basis of the asymptotic expansions of the error rates, the Z statistic is recommended over W for positively correlated training observations which follow an autoregressive process of order one or a moving average process of order one.

Having investigated the effects of correlated training observations on sample discriminant functions W and Z, consideration is given to the performance of the two discriminant functions when they have been formed under models which take the correlation among the training data into consideration. The error rates associated with the sample discriminant functions so formed, denoted as W_m and Z_m (under intraclass correlation model m), are compared with the corresponding error rates associated with the statistics W and Z which are formed under the assumption of independent training observations. Because of the complexity of the problem, only univariate observations are considered. It is found that the performances of Z_m and W_m (which are based on maximum likelihood estimates of the class parameters under intraclass correlation model m) are not necessarily better than those of Z and W. In particular, it is established that the overall asymptotic expected error rate for W_m and W are equal when the training observations follow an autoregressive process, although individual error rates are different. From numerical evaluations of the asymptotic expansions, it is found that for the moving average process of order one, there is a slight improvement in the overall error rate when W_m is used instead of W. The estimated error rates associated with W_m are also evaluated and found to provide better estimates of the actual error rates than the corresponding estimated error rates associated with W.

The problem of estimating mixing proportions in situations where the population

consists of two component classes mixed in different proportions is also considered. The model assumes that a sample of observations is drawn from a mixture of observations from the two classes and that there are available training observations sampled separately from each class which provide information on the unknown class parameters but not on the unknown mixing proportions. Consideration is given to the problem of estimating the mixing proportions and other unknown parameters of the model, when the training observations are correlated. The two estimators of mixing proportions which are considered are the discriminant analysis and the maximum likelihood estimators. The variation of the asymptotic relative efficiency of the discriminant analysis estimator with the level of correlation among the training observations is examined. It has been found that the asymptotic relative efficiency of the discriminant analysis estimator when the training observations are independent can be quite low if one of the mixing proportions is small. It is concluded from studies here that the efficiency of the discriminant analysis estimator is generally lowered by positively correlated training observations. The asymptotic bias of the discriminant analysis estimator is also derived. It is shown that the asymptotic bias of the discriminant analysis estimator is generally not an issue, although it can be quite serious when the classes are close together and mixed disproportionally.

Some preliminary results on the estimation problem which would be encountered if a general model of correlation were to be adopted in the allocation problem (as, for example, in the spatial allocation rule) are also reported.

Department of Mathematics and Statistcs
Massey University
Palmerston North
New Zealand