Influence Analysis of Error Rates: Logistic Discrimination

C. Croux[1], G. Haesbroeck[2], and K. Joossens[1]

[1] Dept. of Applied Economics, K. U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium.
[2] Dept. of Mathematics, University of Liège (B37), Grande Traverse 12, B-4000 Liège, Belgium

Keywords: Diagnostics, Discrimination, Error rate, Influence function, Logistic regression, Robustness.

1 Error rates of classification rules

In supervised classification analysis one wants to classify multivariate observations into two different populations, using the outcome of a classification or discrimination rule. The rule is constructed from a training sample, being observations for which it is known to which population they belong. Then the rule is applied to observations for which it is not known to which population they belong. The error rate is then the total probability of misclassification for these observations to classify.

If outliers are present in the training data, they will influence this error rate. Several measures have been introduced to asses the impact of outliers in the training data on the performance of classification rules. See, among others, Critchley and Vitiello (1991) for linear discriminant analysis and Fung (1996) for quadratic discriminant analysis. Most of these measures focus on the effect of outliers on the estimated parameters of the discriminant rule, and not on the effect on the error rate. This may sometimes give misleading results: the Maximum Likelihood estimator in logistic regression, for example, remains bounded when outliers are added to the training data, but its error rate goes to 50%, the same error rate as one gets with a random guess (Croux, Flandre, and Haesbroeck, 2002). In this paper an influence function approach, common in robust statistics, will be followed. This approach does not seem to have been used much for error rates, exceptions being Croux and Dehon (2001), and Croux and Joossens (2005).

It will be shown that when a classification rule is optimal, i.e. has the lowest possible value for the error rate, then the influence function becomes degenerate. In such a case, one needs to resort to the second order influence function, which appear to be a very natural concept in this setting. As an example, consider a mixture of two normal distributions, both having the same covariance matrix. Then both linear discriminant analysis and logistic discriminant analysis (as well as most of their robust variants) are consistently estimating the optimal classification rule (e.g. Efron 1974). However, at the finite sample level their robustness and classification performance will be different. It turns out that the second order influence function can help us to assess both robustness and classification efficiency. In this paper we will present such a a second order influence analysis for logistic discrimination.

2 The logistic discrimination model

The classical linear discriminant rule of Fisher is well-known and treated in every textbook on multivariate analysis. Many applied researchers, however, give preference to logistic regression as a tool for allocating observations to one out of two populations. Although not optimal at mixtures of normal distributions, it is a flexible method that can deal with different types of variables. Discriminant analysis resulting from an estimated logistic regression model is called logistic discrimination. Over the last decade, several more sophisticated classification methods like support vector machines and random forests have been proposed, but logistic discrimination remains a benchmark method performing well in many applications.

We study the robustness of logistic discriminant analysis, since there may be outlying observations in the training data set who may affect the estimated classification rule and the associated

error rate. Our focus is on the effect of such observations on the error rate, which will be measured by the second order influence function. It is shown that the use of robust estimators for the logistic regression model reduces the effect of outliers on the classification error rate. As robust estimators we consider the Bianco and Yohai (1996) estimator, its weighted version and also the weighted Maximum Likelihood estimator. All these estimator are easily computable (Croux and Haesbroeck 2003). We also illustrate how this influence function can be used as diagnostic tool to pinpoint outlying observations. Moreover, we also compute asymptotic relative error rates of robust logistic discrimination with respect to the Maximum Likelihood approach.

References

A.M. Bianco and V.J. Yohai (1996). Robust estimation in the logistic regression model. In H. Rieder, Ed. *Robust Statistics, Data Analysis and Computer Intensive Methods,* pp 17–34.

F. Critchley and C. Vitiello (1991). The infuence of observations on misclassifcation probability estimates in linear discriminant analysis. *Biometrika*, 78, 677-690.

C. Croux and C. Dehon (2001). Robust Linear Discriminant Analysis using S-estimators. *The Canadian Journal of Statistics*, 29, 473-492.

C. Croux, C. Flandre and G. Haesbroeck (2002). The Breakdown Behavior of the Maximum Likelihood Estimator in the Logistic Regression Model. *Statistics and Probability Letters,* 60, 377-386.

C. Croux and G. Haesbroeck (2003) Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis,* 44, 273-295.

C. Croux and K. Joossens (2005). Influence of Observations on the Misclassification Probability in Quadratic Discriminant Analysis', *Journal of Multivariate Analysis*, to appear.

B. Efron (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70, 892-898.

W.K. Fung (1996). Diagnosing influential observations in quadratic discriminant analysis, *Biometrics*, 52, 1235–1241.