ROBUST AND SPARSE MULTICLASS CLASSIFICATION BY THE OPTIMAL SCORING APPROACH

I. Hoffmann[1], P. Filzmoser, C. Croux
Institute of Statistics and Mathematical Methods in Economics, TU Wien
Vienna, AUSTRIA
e-mail: [1]irene.hoffmann@tuwien.ac.at

Abstract

We present a robust and sparse linear classifier for multiclass problems. Regression methods with the desirable properties of outlier detection and variable selection have got a lot of attention recently. The optimal scoring approach enables a propagation of such regression methods to classification problems.

1 Introduction

Linear discriminant analysis (LDA) is a very simple and popular method for classification, but the model can be distorted by a single anomalous observation. Several robust methods for linear classification have been introduced (see for an overview [3]) to address this issue. These methods are restricted to settings with more observations than variables.

One popular approach in regression analysis for data with a large number of variables compared to the number of observations is Lasso regression. An L_1 norm penalty on the coefficient estimate favours exact zero entries and so excludes uninformative variables from the model. This idea has been incorporated in least trimmed squares (LTS) regression and yields a sparse and robust regression method [1].

2 Optimal Scoring

Optimal scoring is an approach where the class labels are modelled as continuous values and so the classification problem is recast into a regression framework. Let $\boldsymbol{X}$ be an $n \times p$ data matrix, $\boldsymbol{Y}$ an $n \times G$ matrix of dummy variables coding the class membership of the observations, G be the number of classes and $H = G - 1$. For $h = 1, ..., H$

$$\min_{\boldsymbol{\beta}_h, \boldsymbol{\theta}_h} \{\|\boldsymbol{Y}\boldsymbol{\theta}_h - \boldsymbol{X}\boldsymbol{\beta}_h\|^2\} \quad \text{s.t.} \quad \frac{1}{n}\boldsymbol{\theta}_h^T \boldsymbol{Y}^T \boldsymbol{Y}\boldsymbol{\theta}_h = 1, \quad \boldsymbol{\theta}_h^T \boldsymbol{Y}^T \boldsymbol{Y}\boldsymbol{\theta}_l = 0 \quad \forall l < h. \tag{1}$$

Adding an L_1 penalty for $\boldsymbol{\beta}_h$ to the minimization problem leads to sparse discriminant analysis as proposed by [2].

3 Methodology

We propose to propagate the robustness and sparsity properties of sparse LTS to classification problems via the optimal scoring approach. The optimal scoring problem is solved iteratively for $\boldsymbol{\beta}_h$ and $\boldsymbol{\theta}_h$. For fixed $\boldsymbol{\theta}_h$ we replace the least squares minimization in (1) by trimmed least squares with L_1 penalty and solve it with a fast sparse LTS algorithm. This leads to a sparse and robust estimation of $\boldsymbol{\beta}_h$ for $h = 1, \ldots, H$. Then LDA with robustly estimated scatter matrix and centre is applied to $(\boldsymbol{X}\hat{\boldsymbol{\beta}}_1, ..., \boldsymbol{X}\hat{\boldsymbol{\beta}}_H)$.

4 Evaluation

A simulation study is conducted to illustrate the properties of the proposed algorithm. Its performance is compared to classical sparse discriminant analysis by means of correctly selected variables and the ratio of misclassified observations. For simulation settings with more observations than variables further comparison is made with LDA and robust LDA methods.

5 Acknowledgements

This work is supported by the Austrian Science Fund (FWF), project P 26871-N20.

References

[1] Alfons A., Croux C., Gelper S. (2013). Sparse least trimmed squares regression for analysing high-dimensional large data sets. *The Annals of Applied Statistics*. Vol. **7**(1), pp. 226-248.

[2] Clemmensen L., Hastie T., Witten D., Ersboll B. (2011). Sparse discriminant analysis. *Technometrics*. Vol. **53**, pp. 406-413.

[3] Todorov V., Pires A.M. (2007). Comparative performance of several robust linear discriminant analysis methods. *REVSTAT Statistical Journal*. Vol. **5**, pp. 63-83.