Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

Trabalho apresentado no XL CNMAC, Evento Virtual - Co-organizado pela Universidade do Mato Grosso do Sul (UFMS).

Linear Programming Applied to Separation Detection in Polytomous Logistic Regression

Inácio Andruski-Guimarães[1]
DAMAT-UTFPR, Curitiba, PR
Thiago Schinda Bubniak[2]
PIBIC-UTFPR, Curitiba, PR

Abstract. The Logistic Regression Model is widely used in Discriminant Analysis. However, parameter estimation is affected by the data configuration and may not be achieved when there is separation between the groups in the data set, which is a common problem in Discriminant Analysis. The use of linear programming to detect the separation between groups was proposed by [1], and a large number of linear programming approaches have been used to detect separate data in discriminant analysis. However, most research focuses on models for two groups and there are few models for classification problems in multiple groups. In this paper, a linear programming formulation is proposed to detect the separation between groups for the polytomous logistic regression model. The proposed model has a non-negative objective function that has a positive value when the separation is detected and allows to classify the data as completely separate, almost separated or overlapped, and can be used as part of the parameter estimation. A simulation, using data sets from the literature, shows that the proposed approach can be an efficient alternative for mathematical programming applied to problems with multiple groups.

Key-Words. Polytomous Logistic Regression, Discriminant Analysis, Linear Programming, Complete Separation.

1 Introduction

The Discriminant Analysis (DA) is interested in determining the groups of observations based on their observed scores and developing rules for the allocation of new observations into groups. The most popular techniques are Fisher's Linear Discriminant Function (FLDF) and Logistic Regression Model (LRM). The LRM is a method applied to model the relationship between a categorical - or ordinal - dependent variable and a set of explanatory variables, or covariates, that may be either continuous or discrete. The accuracy of the LRM has been reported in many studies involving bankruptcy prediction, marketing applications and cancer classification, among others applications. However, the parameter estimation is known to be dependent on the data configuration. While the model work well for many situations, may not work when the data set has no overlapping.

Mathematical Programming approaches have been used for detecting separated data in discriminant analysis, but almost all researches have focused on the two group problem. Alternative procedures, using a set of interrelated goal programming formulations are suggested by [5], but this approach requires that the data sets needs to be ordered. An algebraic approach, suggested

[1] andruski@utfpr.edu.br.

[2] thiago.bubniak@hotmail.com.

DOI: 10.5540/03.2021.008.01.0427

by [1], uses ideas of linear programming and specifies the necessary constraints, but not an objective function. A mixed integer linear program, presented by [9], determines whether data is separated or overlapped. [10] uses linear programming to check the necessary conditions for the existence of a finite maximum likelihood estimate for the logistic model. A single linear programming formulation, proposed by [2], generates a plan that minimizes an average sum of misclassified points belonging of two disjoint point sets in n-dimensional real space.

In this paper we propose an algorithm based on a Linear Programming Model with a nonnegative objective function that has a positive optimal value when separation is detected. The proposed approach allows to classify the data as completely separated, quasi-separated or overlapped, and can be used as part of the parameter estimation. A comparative analysis using different data sets taken from the literature shows that our linear programming formulation may suggest an efficient alternative to traditional statistical methods and mathematical programming formulations for the multi-group classification problem.

This paper is organized as follows. First, we revisit the Classical Logistic Regression model. Next we present a brief review of Separation. Then we give an overview of the use of Linear Programming for detecting separation. After that, we propose a Linear Programming formulation to detect separation in polytomous logistic regression. Last, we apply the formulation on data sets taken from the literature in order to observe its performance when applied to detect separation. Finally we give a brief conclusion about the results achieved.

2 Polytomous Logistic Regression Model

Let us consider a sample of n independent observations, available from the groups $G_1, ..., G_s$, and a vector $\mathbf{x}$ of $(p+1)$ explanatory continuous variables, given by $\mathbf{x^T} = (x_0, x_1, ..., x_p)$, where $x_0 \equiv 1$, for convenience. In this case, we know the membership of each observation with respect to the groups. Furthermore, we also assume that each group has n_j observations, $j = 1, ..., s$, such that $n = \sum_{j=1}^{s} n_j$. Let Y denote the polytomous dependent variable with s possible outcomes. We will summarize the n observations in a matrix form given by:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ ... & ... & ... & ... \\ 1 & x_{1n} & ... & x_{pn} \end{bmatrix}$$

The Classical Logistic Regression (CLR) Model assumes that the posterior probabilities have the form:

$$P\left(G_k \mid \mathbf{x}\right) = \frac{exp\left(\mathbf{B}_k\right)}{\sum\limits_{i=1}^{s} exp\left(\mathbf{B}_i\right)}\ , \tag{1}$$

where $\mathbf{B}_k = \beta_{k0} + \sum\limits_{j=1}^{p} \beta_{kj} x_j,\ k = 1\,, 2\,, ...\,, s-1$ and $\mathbf{B}_s = \mathbf{0}$. In this paper the group s is called reference group. The model involves $(s-1)\,(p+1)$ unknown parameters and the conditional likelihood function is:

$$L\left(\mathbf{B} \mid \mathbf{Y}, \mathbf{x}\right) = \prod_{i=1}^{n} \prod_{k=1}^{s} \left[P\left(G_k \mid \mathbf{x}_i\right)\right]^{Y_{ki}}\ , \tag{2}$$

DOI: 10.5540/03.2021.008.01.0427

where $\mathbf{Y} = (\mathbf{Y}_1 \, , \, ... \, , \mathbf{Y}_n)^{\mathbf{T}}$ and $\mathbf{Y}_i = (Y_{1i}, ..., Y_{si})$, with $Y_{ki} = 1$ if $Y = k$, and $Y_{ki} = 0$ otherwise. Taking the logarithm, the log-likelihood function is given by:

$$\ell\left(\mathbf{B} \mid \mathbf{Y}, \mathbf{x}\right) = \sum_{i=1}^{n} \sum_{k=1}^{s} Y_{ki} ln \left[P\left(G_k \mid \mathbf{x}_i\right)\right] \, . \tag{3}$$

Thus:

$$\frac{\partial}{\partial \beta_{kj}} \ell\left(\mathbf{B} \mid \mathbf{Y}, \mathbf{x}\right) = \sum_{i=1}^{n} x_{ij} \left(Y_{ki} - P\left(G_k \mid \mathbf{x}_i\right)\right) \, . \tag{4}$$

The Maximum Likelihood Estimator (MLE) $\hat{\mathbf{B}}$ is obtained by setting the derivatives (4) to zero and solving for $\mathbf{B}$. The solution is found using an iterative procedure, such as Newton-Raphson method.

3 Separation

Separation is a common problem in discriminant analysis, and occurs quite frequently. In practice, the estimation of unknown parameters in logistic regression should be considering the possible configurations of the sample points. An approach proposed by [1] suggested a sample classification into three mutually exclusive categories: Overlapped, completely separated and quasi-completely separated. They also proved that the MLE do not exists if there is complete separation among the groups, that is, when the groups are linearly separable. If there is complete, or quasi-complete separation, existing iterative methods fail to converge, or give a wrong answer. In binary logistic regression, if there is complete separation, the MLE do not exist. However, the problem can be easily avoided by using other methods, such as Fisher's Linear Discriminant Function (FLDF) or Decision Trees (DT), for example. But, in polytomous logistic regression, complete separation does not make the same sense, although the parameter estimation is not necessarily affected.

We say that two groups G_i and G_j are linearly separable if there exists a vector given by $\mathbf{B} = (\beta_1, ..., \beta_p)$ and a real number δ such that $\mathbf{B}\,\mathbf{x}_k > \delta$ if $\mathbf{x}_k \in G_i$ and $\mathbf{B}\,\mathbf{x}_k < \delta$ if $\mathbf{x}_k \in G_j$, where $i, j = 1, ..., s$, $i \neq j$ and $k = 1, ..., n$. When there are more than two groups, the difference between linear separability and separation becomes more important. In this case linear separability means the existence of a set of vectors $\mathbf{B}_1, ..., \mathbf{B}_s$ satisfying $s(s-1)$ inequalities given by:

$$(\mathbf{B}_j - \mathbf{B}_t)^{\mathbf{T}} \mathbf{x}_i \geq \delta \, , \tag{5}$$

for all $i = 1, ..., n$, and $j, t = 1, ..., s \, (j \neq t)$.

4 Linear Programming for Detecting Separation

The use of linear programming for detecting separation among the sample points was proposed by [1]. A mixed integer linear program proposed by [9] classifies a data set as completely separated, quasicompletely separated and overlapped and, in case of quasicomplete separation, identifies the minimal set of quasiseparated points. According to [9], the model is always feasible.

A goal programming model, introduced by [5], consists of a model which can be posed as: Given the groups G_j, $j = 1, ..., s$ and the vectors $\mathbf{x}_i$, $i = 1, ..., n$, the goal is to find a linear transformation T and the boundaries L_j and U_j to classify each $\mathbf{x}_i$, where L_j and U_j represent the lower and upper boundaries for observations assigned to the j-th group. The objective is to determine a linear predictor $\mathbf{T}_k = (t_{k1}, ..., t_{kp})$, $k = 1, ..., s$, and the breakpoints L_j and U_j, such that:

$$L_j \leq \mathbf{T}_k \mathbf{x} \leq U_j \Leftrightarrow \mathbf{x} \in G_j \tag{6}$$

and

$$L_1 < U_1 < L_2 < U_2 < \ldots < L_s < U_s \,. \tag{7}$$

Let C_j, $j = 1, ..., s$, be the costs of classifying an observation $\underline{\mathbf{x}}$ as belonging to G_j, when, in fact, the observation is not. Then the problem can be given as:

$$Min \;\; \sum_{j=1}^{s} C_j \mathbf{T}_j$$

$$s.t \;\; \left\{ \begin{array}{l} \mathbf{T}_j \mathbf{x} \geq L_j \\ \mathbf{T}_j \mathbf{x} \leq U_j \end{array} \right. \;\; (8)$$

An obstacle to the application of this approach is the need for the classification scores, given by $S_i = \sum_{j=1}^{p} w_j x_{ij}$, for $i \in G_k$, to be ordered in some way. Another problem, according to [6], is that the resulting classification rules may not cover each segment of the decision space.

In order to introduce our approach we start by considering a matrix $\mathbf{X}_j$ with rows $\mathbf{x}_i^{\mathbf{T}}$ such that $i \in G_j$. We can define the $(s-1) \times s$ matrix $\tilde{\mathbf{X}}_j$ to have blocks $\mathbf{X}_j$ in each element of column j, blocks $-\mathbf{X}_j$ in row k and column k, for $k < j$, and in row $k-1$ and column k, for $j < k$, and to be zero otherwise. For example, if the problem has four groups, the matrix $\tilde{\mathbf{X}}_3$, is given by:

$$\tilde{\mathbf{X}}_3 = \left[\begin{array}{cccc} -\mathbf{X}_3 & 0 & \mathbf{X}_3 & 0 \\ 0 & -\mathbf{X}_3 & \mathbf{X}_3 & 0 \\ 0 & 0 & \mathbf{X}_3 & -\mathbf{X}_3 \end{array} \right].$$

As stated by [7], if we let

$$\tilde{\mathbf{X}} = \left[\begin{array}{c} \tilde{\mathbf{X}}_1 \\ \vdots \\ \tilde{\mathbf{X}}_s \end{array} \right],$$

then $\tilde{\mathbf{X}}\mathbf{B} \geq 0$ implies that $\mathbf{B}$ satisfies the conditions for quasi-complete separation, where $\mathbf{B} = \left(\mathbf{B}_1^{\mathbf{T}}, \ldots, \mathbf{B}_s^{\mathbf{T}}\right)^{\mathbf{T}}$ is the parameter vector. If $\tilde{\mathbf{X}}_{\mathbf{j}}\mathbf{B} \geq 0$, for a given j, then

$$(\mathbf{B}_j - \mathbf{B}_t)^{\mathbf{T}} \mathbf{x}_i \geq 0 \,. \tag{9}$$

Because $\mathbf{B}_s = \mathbf{0}$, the corresponding columns of $\tilde{\mathbf{X}}_s$ does not need to be stored, so the matrix $\tilde{\mathbf{X}}_j$ can be stored in an $n(s-1) \times p(s-1)$ matrix. Let us consider s groups $G_1, ..., G_s$, and a vector of explanatory variables, given by $\mathbf{x}^{\mathbf{T}} = (x_1, ..., x_p)$. Suppose there is complete separation among two groups, G_r and G_t, $r = 1, ..., s-1, t = 2, ..., s \; (r < t)$. Then there is a hyperplane H_{rt} such that all of the sample points in G_r lie on one side of H_{rt} and all of the sample points in G_t lie on the other side of the hyperplane. The distance of the $\mathbf{x}_k$ point from the hyperplane is given by $d_{krt} = \mathbf{x}_k^{\mathbf{T}} \mathbf{u}_{rt}$,

DOI: 10.5540/03.2021.008.01.0427

$k = 1, ..., n$, $r = 1, ..., (s-1)$, $t = 2, ..., s$, $r < t$, where $\mathbf{u}_{rt}$ is a unit vector normal to H_{rt} showing separation among G_r and G_t, when it is present. If there is complete separation among G_r and G_t, then there is a vector $\mathbf{u}_{rt}$ such that $d_{krt} < 0$, if $\mathbf{x}_k \in G_r$, and $d_{krt} > 0$, if $\mathbf{x}_k \in G_t$. If there is quasi-complete separation among the groups G_r and G_t, then $d_{krt} \leq 0$, if $\mathbf{x}_k \in G_r$, and $d_{krt} \geq 0$, if $\mathbf{x}_k \in G_t$, with equality for at least one value $k = 1, ..., n$. Let $S(\mathbf{u}_{rt}) = \sum_{k=1}^{n} d_{krt}$ and a hyperplane P_{rt} with normal vector $\mathbf{u}_{rt}^*$ such that $S(\mathbf{u}_{rt}^*)$ is maximum. In this case, finding $\mathbf{u}_{rt}^*$ can be posed as:

$$Max \quad S(\mathbf{u}_{rt}) = \sum_{k=1}^{n} d_{krt}$$

$$s.t \ \left\{ \begin{array}{c} d_{krt} = \mathbf{x}_k^{\mathbf{T}} \mathbf{u}_{rt} \geq 0 \\ d_{krt} = \mathbf{x}_k^{\mathbf{T}} \mathbf{u}_{rt} \leq 0 \\ \mathbf{u}_{rt}^{\mathbf{T}} \mathbf{u}_{rt} = 1 \end{array} \right. \quad (10)$$

If there is no vector $\mathbf{u}_{rt}$ satisfying the constraints, then there is overlap among G_r and G_t. If the model above is feasible, its solution provides a vector $\mathbf{u}_{rt}^*$ such that $S(\mathbf{u}_{rt}^*)$ is maximum. The third constraint, given by $\mathbf{u}_{rt}^T \mathbf{u}_{rt} = 1$, forces $\mathbf{u}_{rt}$ to have unit length. However is not considered because our purpose is to determine if the sample point is completely separated, or quasi-completely separated, by $\mathbf{u}_{rt}$, hence the length of $\mathbf{u}_{rt}$ is not relevant. Furthermore, the referred constraint gives a non linearly constrained optimization problem, which is not appropriated for our purpose. Therefore, the proposed model is given by:

$$Max \quad S(\mathbf{u}_{rt}) = \sum_{k=1}^{n} d_{krt}$$

$$s.t \ \left\{ \begin{array}{c} d_{krt} = \mathbf{x}_k^{\mathbf{T}} \mathbf{u}_{rt} \geq 0 \\ d_{krt} = \mathbf{x}_k^{\mathbf{T}} \mathbf{u}_{rt} \leq 0 \end{array} \right. \quad (11)$$

Let $\mathbf{U} = \left(u_{12}, ..., u_{(s-1)s}\right)$ be the matrix whose columns are the vectors $\mathbf{u}_{rt}$. Taking into account that $d_{krt} = \mathbf{x}_k^T \mathbf{u}_{rt}$, $S(\mathbf{u}_{rt}) = \sum_{k=1}^{n} d_{krt}$ can be expressed as $\sum_{k=1}^{n} d_{krt} = \mathbf{c}_n^T \tilde{\mathbf{X}}_k \mathbf{U}$, where $\mathbf{c}_n$ is a vector with n ones. Thus, the resulting linear programming problem can be posed as:

$$Max \quad S(\mathbf{u}_{rt}) = \sum_{k=1}^{n} d_{krt}$$

$$s.t \ \left\{ \begin{array}{c} \mathbf{X}_r^{\mathbf{T}} \mathbf{u}_{rt} \geq \mathbf{0} \\ \mathbf{X}_t^{\mathbf{T}} \mathbf{u}_{rt} \leq \mathbf{0} \end{array} \right. \quad (12)$$

where $\mathbf{0} = (0, ...0)^{\mathbf{T}}$. For s groups there are $s(s-1)/2$ models, and each one, which can be solved using techniques of linear programming, such as Simplex Method or Interior Points Method, has $n_r + n_t$ constraints and the same number of slack variables. Furthermore we should to taking into account that observations with the same values, or which can be expressed as linear combinations of other observations, a phenomenon also known as multicollinearity, results in redundant equations. In this case the elimination of these equations leads to a equivalent system having fewer equations and the same number of variables. Another approach, suggested by [8], is to express the free decision variables as linear combinations of the slack variables.

5 Applications

In this section we consider two benchmark data sets, taken from the literature. Iris Data, taken from [4], and Fatty Acid Composition Data, taken from [3]. We have applied the proposed model to both data sets. The results achieved are given in the sequence.

Example 1: Iris Data. There are three groups of Iris flowers: *Iris Setosa* (G_1), *Iris Versicolor* (G_2) and *Iris Virginica* (G_3). For each group there are 50 observations and four independent variables: Sepal Length (x_1), Sepal Width (x_2), Petal Length (x_3) and Petal Width (x_4). In this paper, the reference group is Iris Virginica. The vectors are shown in Table 1.

Table 1: Vectors for Iris Data.

$\mathbf{u}_{12}$	$\mathbf{u}_{13}$	$\mathbf{u}_{23}$
1.22	0	∄
0	0.12	∄
2.03	1.19	∄
0	0	∄

Our results showed that two groups, G_2 (*Iris Versicolor*) and G_3 (*Iris Virginica*), overlap and form a cluster completely separated from G_1 (*Iris Setosa*). In this case, it would be possible to replace the polytomous logistic regression model with a decision tree using a binary model for groups G_2 and G_3, as shown by [**?**].

Example 2: Fatty Acid Data. There are 120 observations, five groups and seven variables, representing the percentage levels of seven fatty acids, namely palmitic (x_1), stearic (x_2), oleic (x_3), linoleic (x_4), linolenic (x_5), eicosanoic (x_6) and eicosenoic (x_7) acids. In this paper we consider five groups: rapeseed (G_1), sunflower (G_2), peanut (G_3), corn (G_4) and pumpkin (G_5) oils. In this paper the reference group is G_5 (pumpkin oil). The original data set have eight groups, and the complete table of the original data can be found in [3].

Table 2: Vectors for Fatty Acid Data.

$\mathbf{u}_{12}$	$\mathbf{u}_{13}$	$\mathbf{u}_{14}$	$\mathbf{u}_{15}$	$\mathbf{u}_{23}$	$\mathbf{u}_{24}$	$\mathbf{u}_{25}$	$\mathbf{u}_{34}$	$\mathbf{u}_{35}$	$\mathbf{u}_{45}$
33.28	∄	∄	2.60	0.00	∄	∄	∄	∄	∄
21.83	∄	∄	0.00	51.84	∄	∄	∄	∄	∄
139.38	∄	∄	4.05	38.03	∄	∄	∄	∄	∄
268.18	∄	∄	0.00	260.72	∄	∄	∄	∄	∄
0.00	∄	∄	0.00	484.90	∄	∄	∄	∄	∄
1.05	∄	∄	2.09	10.42	∄	∄	∄	∄	∄
0.00	∄	∄	0.28	0.00	∄	∄	∄	∄	∄

Our results showed that the group G_1 (rapeseed oil) is completely separated from G_2 (rapeseed oil) and G_5 (pumpkin oil). Furthermore, the group G_2 (sunflower oil) is completely separated from G_3 (peanut oil). Furthermore, we can see that G_4 has overlap with all other groups. In this case the parameter estimation would not necessarily affected, since it would be possible to estimate all the parameters for a model that uses G_4 (corn oil) as a reference group.

6 Conclusion

The main purpose with this job is to develop and implement a simple and direct model based on Linear Programming which allows the detection of separation among sample points, in order to assist the parameter estimation for the Polytomous Logistic Regression Model, and to explore the performance of the referred approach. The referred linear programming model has a non-negative objective function that has a positive optimal value when separation is detected. The results achieved suggest that the approach is a promising alternative to detecting separation, even when a large number of dimensions have to be considered. Furthermore the proposed approach provides a simple and easy-to-implement solution and does not need any particular ordering arrangement of data, which is an advantage for practical purposes, and there are no computational difficulties for its implementation, since it uses common algorithms for Linear Programming problems. In the next step of our study we intend to evaluate the applicability of this approach as an alternative method to the variable selection for the Polytomous Logistic Regression Model.

References

[1] A. Albert, J. A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, **71**, 1–10 , 1984

[2] K. P. Bennet, O. L. Mangasarian, Multicategory discrimination via linear programming, *Optimization Methods and Software*, **1**, 23–24, 1992

[3] D.Brodnjak – Vončina, Z.C.Kodba, C.Novič, Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. *Chemometrics and Intelligent Laboratory Systems* **75**, 31–43, 2005.

[4] R.A. Fisher, The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **3**, 179–188, 1936

[5] N. Freed, F. Glover, Simple but powerful goal programming models for discriminant problems. *European Journal of Operational Research* **7**, 44–60, 1981.

[6] W. Gochet, V. Srinivasan, A. Stam, S. Chen, Multi-group discriminant analysis using linear programming. *Operations Research* **45** (2), 213–225, 1997.

[7] K. Konis, Linear programming algorithms for detecting separated data in binary logistic regression models. DPhill in Computational Statistics, Worcester College, University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, United Kingdom, 2007.

[8] D. G. Luenberger, *Linear and Nonlinear Programming*, 4th. Ed.. Addison-Wesley Publishing Company. Reading, MA. (1973).

[9] T. J. Santner, D. E. Duffy, A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **73**, 1–10, 1986.

[10] M. J. Silvapulle, J. Burridge, Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *J. R. Statist. Soc. B*, **48**, 100–106, 1986.

DOI: 10.5540/03.2021.008.01.0427