Logistic regression for potential modeling

Samuel Kost¹,*, Oliver Rheinbach¹,**, and Helmut Schaeben²,***

¹ TU Bergakademie Freiberg, Department of Numerics and Optimization, Akademiestraße 6, 09599 Freiberg
² TU Bergakademie Freiberg, Department of Geophysics and Geoinformatics, Gustav-Zeuner-Str. 12, 09599 Freiberg

Regression or regression-like models are often employed in potential modeling, i.e., for the targeting of resources, either based on 2D map images or 3D geomodels both in raster mode or based on spatial point processes. Recently, machine learning techniques such as artificial neural networks have gained popularity also in potential modeling. Using artificial neural networks, decent results in the prediction of the target event are obtained. However, insight into the problem, e.g., about importance of specific covariables, is difficult to obtain. On the other hand, logistic regression has a well understood statistical foundation and works with an explicit model from which knowledge can be gained about the underlying problem. However, establishing such an explicit model is rather difficult for real world problems. We propose a model selection strategy for logistic regression, which includes nonlinearities for improved classification results. At the same time, interpretability of the results is preserved.

© 2019 The Authors Proceedings in Applied Mathematics & Mechanics published by Wiley-VCH Verlag GmbH & Co. KGaA Weinheim

1 Introduction

The objective of potential modeling in geosciences is to predict locations for which the estimated conditional probability of the occurrence of a target event (e.g., an ore deposit) is maximal, given some favorable or unfavorable predictor variables. Presuming (i) a spatial resolution which is adequate for sound economic and ecologic decisions, and (ii) a reasonable total number of covariables capturing geological, geophysical, geochemical and other features of the region, the computational cost for many of the mathematical-statistical methods becomes demanding if not prohibitive.

The logistic regression model is a widely used tool in statistics for the classification of a two-class dependent variable. It has been extensively studied, e.g., in [1] and information about the underlying structure of the problem can be gained from the model. The application of truncated Newton methods [2, 3], i.e., the use of Krylov methods, to obtain the maximum likelihood estimate (MLE), has enabled logistic regression methods to tackle large scale applications. Introducing nonlinearities improves the performance of a logistic regression model. However, a model selection is necessary for the selection of appropriate nonlinearities.

2 Logistic Regression

The logistic regression model is a special case of a generalized linear model. It consists of the linear predictor $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$ and a link function which links the conditional probability $P(y|x) = \mu(\eta)$ to the linear predictor. This link function is the logarithm of the odds, called the logit, given as logit $= \log \frac{\mu}{1-\mu}$. The MLE $\beta$ is obtained by minimizing the negative logarithm of the likelihood function, i.e.

$$\min_{\beta} L(\beta) = -\sum_{i=1}^{n} \{ y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i) \}$$

This equation is solved using the Newton-Raphson algorithm where the new iterate can be obtained by solving a system of linear equations. Using the conjugate gradient method with an early stopping criterion, logistic regression methods can tackle also large scale applications [2].

The application of logistic regression in potential modeling has some specific challenges. First, one needs to deal with rare events. This means that the occurrence of one class has a significantly lower frequency than the majority class. It was shown, e.g., in [4], that logistic regression underestimates the probabilities of the rare events because it tends to be biased towards the more frequent class, which in many practical applications is the less important class. This can be solved through endogenous sampling which means taking all positive events and a random sample of the negative events to get balanced training data. This makes some corrections necessary such as using a weighted likelihood [5] and calculating the robust variance [6].

* Corresponding author: e-mail Samuel.Kost@math.tu-freiberg.de, phone +49 3731 393394
** Corresponding author: e-mail oliver.rheinbach@math.tu-freiberg.de
*** Corresponding author: e-mail Helmut.Schaeben@geophysik.tu-freiberg.de

This is an open access article under the terms of the Creative Commons Attribution License 4.0, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/pamm.201900039

© 2019 The Authors Proceedings in Applied Mathematics & Mechanics published by Wiley-VCH Verlag GmbH & Co. KGaA Weinheim

Received: 3 April 2019 Accepted: 7 June 2019
3 Model selection strategy

Since logistic regression is only a linear classifier we provide a dictionary with some nonlinearities in the data $x$. After choosing the appropriate nonlinearities, these are added to the linear predictor, i.e. $\eta = x^T \beta + \sum_{j=m+1}^{\tilde{m}} \beta_j f(x)$ with $f(x) \in \mathbb{R}^m \rightarrow \mathbb{R}$. This is still linear in the parameters $\beta_i$, $i = 1, \ldots, \tilde{m}$ where $\tilde{m}$ is the total number of variables, including the nonlinearities. To obtain such a model a selection of variables needs to be carried out. The proposed model selection is performed in two main steps. The first step is a coarse selection using the $p$-value of the Wald test, which tests for the significance of a variable. Because of known problems with the Wald test in rare events and large samples, e.g. [7], many unimportant variables will be left in the model. Therefore a second selection step is needed. This uses the Bayes’ information criterion (BIC). The BIC for a given model is calculated as

$$\text{BIC} = -2 \ln(\hat{L}) + \ln(n)k$$

where $n$ and $k$ describe the number of datapoints and number of variables, respectively. Furthermore, $\hat{L}$ is the function value of the likelihood for the MLE $\hat{\beta}$. Because the model should be able to achieve good results on unseen data, the BIC is calculated using a validation dataset. After calculating the BIC values of all models with one variable dropped we sort the unimportant variables. Then we apply the following method to discard more than one variable: We calculate the BIC for the models where one-half, one-quarter and one-eights of the most unimportant variables are dropped. The model with the smallest BIC is used as the starting model for the next iteration.

4 Results

Experiments on synthetic data, where the true model is known, show, that our suggested method is able to detect the true model if possible or approximates it using the given nonlinearities if it cannot be fully recovered. In both cases, it improves the performance of a simple logistic regression. It comes close to the prediction accuracy of a neural network while remaining interpretable. Due to space limitations the results on synthetic data are not presented here.

Our experiments with real world data show a similar behavior. Results for the datasets described in Table 1 are presented in Figure 1 and Figure 2. The model selection improves the simple logistic regression in both cases and for the Cod-RNA dataset it even gives the same prediction result as a neural network with 20 hidden layers and the logistic function as activation function.

![Cod-RNA: precision-recall curve](image1.png)  
**Fig. 1:** Precision-Recall curve for the cod-RNA dataset

![Covertype: precision-recall curve](image2.png)  
**Fig. 2:** Precision-Recall curve for the covertype dataset

References

[1] D. W. Hosmer, S. Lemeshow and R. X. Sturdivant, Applied Logistic Regression (Wiley, New Jersey, 2013)
[2] P. Komarek and A. Moore, Institute, Carnegie Mellon University, (2005).
[3] C. Lin, R. C. Weng, and S. S. Keerthi, JMLR 9, p. 627–650 (2008)
[4] G. King and L. Zeng, Political Analysis 9, p. 137–163 (2001)
[5] P. McCullagh and J. A. Nelder, Generalized Linear Models (Chapman & Hall/CRC, 1989)
[6] R. J. Carroll and G. Kauermann, JASA 96, p. 1387–1396 (2001)
[7] M. Lin, H. C. Lucas and G. Shmueli, Inf. Syst. Res. 24 (2013)