Logistic regression in data analysis: An overview

Article *in* International Journal of Data Analysis Techniques and Strategies · July 2011

DOI: 10.1504/IJDATS.2011.041335 · Source: RePEc

1 author:

Maher Maalouf
Khalifa University
31 PUBLICATIONS **354** CITATIONS

SEE PROFILE

Logistic Regression in Data Analysis: An Overview

Maher Maalouf*

School of Industrial Engineering
University of Oklahoma
202 W. Boyd St., Room 124
Norman, OK. 73019 USA
E-mail: mcm@ou.edu
*Corresponding author

Abstract: Logistic regression (LR) continues to be one of the most widely used methods in data mining in general and binary data classification in particular. This paper is focused on providing an overview of the most important aspects of LR when used in data analysis, specifically from an algorithmic and machine learning perspective and how LR can be applied to imbalanced and rare events data.

Keywords: data mining, logistic regression, classification, rare events, imbalanced data

Reference to this paper should be made as follows: Maalouf, M. (xxxx) 'Logistic Regression in Data Analysis: An Overview', *International Journal of Data Analysis Techniques and Strategy (IJDATS)*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Maher Maalouf received his PhD in Industrial Engineering in 2009 from the University of Oklahoma, OK. He is a Post Doctorate Research Associate in the Department of Industrial Engineering at the University of Oklahoma. His research interests include operations research, data mining and machine learning methods, multivariate statistics, optimization methods, and robust regression.

1 Introduction

Logistic Regression (LR) is one of the most important statistical and data mining techniques employed by statisticians and researchers for the analysis and classification of binary and proportional response data sets [2, 27, 29, 40]. Some of the main advantages of LR are that it can naturally provide probabilities and extend to multi-class classification problems [27, 37]. Another advantage is that most of the methods used in LR model analysis follow the same principles used in linear regression [31]. What's more, most of the unconstrained optimization

techniques can be applied to LR [45]. Recently, there has been a revival of LR importance through the implementation of methods such as the truncated Newton. Truncated Newton methods have been effectively applied to solve large scale optimization problems. Komarek and Moore [42] were the first to show that the Truncated-Regularized Iteratively Re-weighted Least Squares (TR-IRLS) can be effectively implemented on LR to classify large data sets, and that it can outperform the Support Vector Machines (SVM) [64], which is considered a state-of-the-art algorithm. Later on, trust region Newton method [45], which is a type of truncated Newton, and truncated Newton interior-point methods [41] were applied for large scale LR problems.

With regard to imbalanced and rare events data, and/or small samples as well as certain sampling strategies (such as choice-based sampling), however, the standard binary methods, including LR, are inconsistent unless certain corrections are applied. The most common correction techniques are *prior correction* and *weighting* [39]. King and Zeng [39] applied these corrections to the LR model, and showed that they can make a difference when the population probability of interest is low.

This paper provides an overview of some of the algorithms and the corrections that enable LR to be both fast and accurate from a machine learning point of view. It is by no means an exhaustive survey of all the LR techniques in data mining. Rather, the objective is to enable researchers to choose the right techniques based on the type of data they deal with while at the same time expose them to the flexibility and effectiveness of the LR method. In fact, binary LR models are the foundation from which more complex models are constructed [46]. The techniques presented in this paper apply mainly to continuous, large-scale categorical response as well as imbalanced and rare-event data sets with no missing values.

Section 2 provides a description of the LR method. Section 3 discusses fitting LR with the Iteratively-Reweighted Least Squares (IRLS) technique. In Section 4 we discuss LR in imbalanced and rare events data. Section 5 provides a brief survey of LR pertaining to other common data mining challenges and Section 6 states the conclusion.

2 The Logistic Regression Model

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix where n is the number of instances (examples) and d is the number of features (parameters or attributes), and $\mathbf{y}$ be a binary outcomes vector. For every instance $\mathbf{x}_i \in \mathbb{R}^d$ (a row vector in $\mathbf{X}$), where $i = 1 \dots n$, the outcome is either $y_i = 1$ or $y_i = 0$. Let the instances with outcomes of $y_i = 1$ belong to the positive class (occurrence of an event), and the instances with outcomes $y_i = 0$ belong to the negative class (non-occurrence of an event). The goal is to classify the instance $\mathbf{x}_i$ as positive or negative. An instance can be thought of as a Bernoulli trial (the *random component*) with an expected value $E[y_i]$ or probability p_i.

A linear model to describe such a problem would have the matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where ε is the error vector, and where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1\, x_{11}\, x_{12} \cdots x_{1d} \\ 1\, x_{21}\, x_{22} \cdots x_{2d} \\ \vdots\ \vdots\ \ \vdots\ \ \ \ \vdots \\ 1\, x_{n1}\, x_{n2} \cdots x_{nd} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}, \text{and } \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}. \tag{2}$$

The vector $\boldsymbol{\beta}$ is the vector of unknown parameters such that $\mathbf{x}_i \leftarrow [1, \mathbf{x}_i]$ and $\boldsymbol{\beta} \leftarrow [\beta_0, \boldsymbol{\beta}^{\mathrm{T}}]$. From now on, the assumption is that the intercept is included in the vector $\boldsymbol{\beta}$. Now, since $\mathbf{y}$ is a Bernoulli random variable with a probability distribution

$$P(y_i) = \begin{cases} p_i, & \text{if } y_i = 1; \\ 1 - p_i, & \text{if } y_i = 0; \end{cases} \tag{3}$$

then the expected value of the response is

$$E[y_i] = 1(p_i) + 0(1 - p_i) = p_i = \mathbf{x}_i\boldsymbol{\beta}, \tag{4}$$

with a variance

$$V(y_i) = p_i(1 - p_i). \tag{5}$$

It follows from the linear model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \tag{6}$$

that

$$\varepsilon_i = \begin{cases} 1 - p_i, & \text{if } y_i = 1 \text{ with probability } p_i; \\ -p_i, & \text{if } y_i = 0 \text{ with probability } 1 - p_i; \end{cases} \tag{7}$$

Therefore, ε_i has a distribution with an expected value

$$E[\varepsilon_i] = (1 - p_i)(p_i + (-p_i)(1 - p_i) = 0, \tag{8}$$

and a variance

$$V(\varepsilon_i) = E[\varepsilon_i^2] - E[\varepsilon_i] = (1 - p_i)^2(p_i) + (-p_i)^2(1 - p_i) - (0) \tag{9}$$

$$= p_i(1 - p_i). \tag{10}$$

Since the expected value and variance of both the response and the error are not constant (heteroskedastic), and the errors are not normally distributed, the least squares approach cannot be applied. In addition, since $y_i \in \{0, 1\}$, linear regression would lead to values above one or below zero. Thus, when the response vector is binary, the logistic response function, as shown in Figure 2, is the appropriate one.

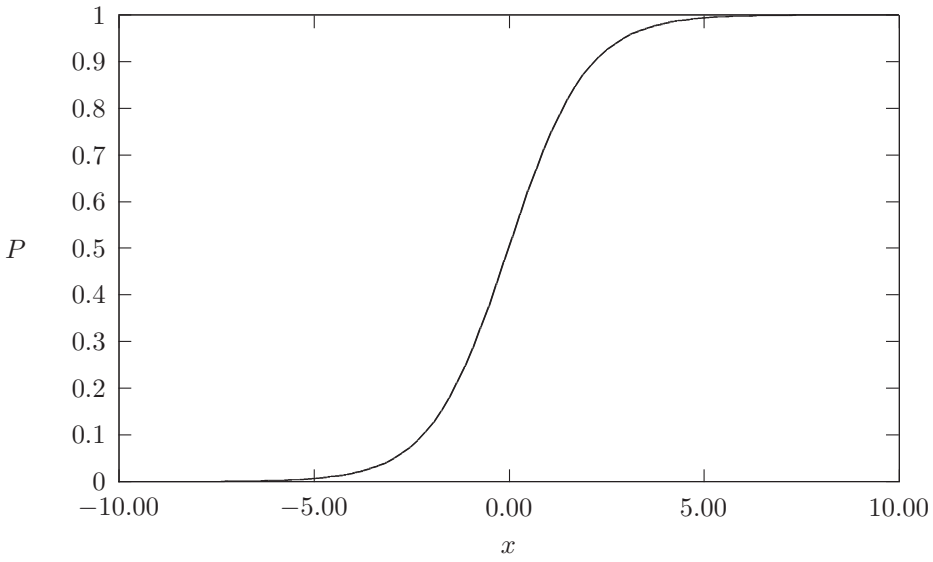

Figure 1 Logistic Response Function

The logistic function commonly used to model each positive instance $\mathbf{x}_i$ with its expected binary outcome is given by

$$E[y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}] = p_i = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}} = \frac{1}{1 + e^{-\mathbf{x}_i\boldsymbol{\beta}}}, \qquad \text{for } i = 1, \dots n. \tag{11}$$

The logistic (logit) transformation is the logarithm of the odds of the positive response, and is defined as

$$\eta_i = g(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_i\boldsymbol{\beta}. \tag{12}$$

In matrix form, the logit function is expressed as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \tag{13}$$

The logit transformation function is important in the sense that it is linear and hence it has many of the properties of the linear regression model. In LR, this function is also called the *canonical link function*, which relates the linear predictor η_i to $E[y_i] = p_i$ through $g(p_i)$. In other words, the function $g(.)$ links $E[y_i]$ to $\mathbf{x}_i$ through the linear combination of $\mathbf{x}_i$ and $\boldsymbol{\beta}$ (the *systematic component*). Furthermore, the logit function implicitly places a separating hyperplane, $\beta_0 + \langle \mathbf{x}, \boldsymbol{\beta} \rangle = 0$, in the input space between the positive and non-positive instances.

The most widely used general method of estimation is the method of *Maximum Likelihood* (ML). The ML method is based on the joint probability density of the observed data, and acts as a function of the unknown parameters in the model [26].

Now, with the assumption that the observations are independent, the likelihood function is

$$\mathbb{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} (p_i)^{y_i} (1-p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right)^{1-y_i}, \tag{14}$$

and hence, the log-likelihood is then

$$\ln \mathbb{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \ln \left(\frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right) + (1-y_i) \ln \left(\frac{1}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right) \right). \tag{15}$$

Amemiya [3] provides formal proofs that the ML estimator for LR satisfies the ML estimators' desirable properties. Unfortunately, there is no closed form solution to maximize $\ln \mathbb{L}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The LR *Maximum Likelihood Estimates* (MLE) are therefore obtained using numerical optimization methods, which start with a guess and iterate to improve on that guess. One of the most commonly used numerical methods is the Newton-Raphson method, for which, both the gradient vector and the Hessian matrix are needed:

$$\frac{\partial}{\partial \beta_j} \ln \mathbb{L}(\beta) = \sum_{i=1}^{n} \left(y_i \left(\frac{x_{ij}}{1+e^{\mathbf{x}_i \beta}} \right) + (1-y_i) \left(\frac{-x_{ij} e^{\mathbf{x}_i \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right) \right) \tag{16}$$

$$= \sum_{i=1}^{n} \left(y_i x_{ij} \left(\frac{1}{1+e^{\mathbf{x}_i \beta}} \right) - (1-y_i) x_{ij} \left(\frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right) \right) \tag{17}$$

$$= \sum_{i=1}^{n} (y_i x_{ij} (1-p_i) - (1-y_i) x_{ij} (p_i)) \tag{18}$$

$$= \sum_{i=1}^{n} (x_{ij} (y_i - p_i)) = 0, \tag{19}$$

where $j = 0, ...d$ and d is the number of parameters. Each of the partial derivatives is then set to zero. In matrix form, equation (19) is written as

$$g(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \ln \mathbb{L}(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}} (\mathbf{y} - \mathbf{p}) = \mathbf{0}. \tag{20}$$

Now, the second derivatives with respect to β are given by

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} \ln \mathbb{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{-x_{ij} x_{ik} e^{\mathbf{x}_i \boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i \boldsymbol{\beta}})(1+e^{\mathbf{x}_i \boldsymbol{\beta}})} \right) \tag{21}$$

$$= \sum_{i=1}^{n} (-x_{ij} x_{ik} (p_i (1-p_i))). \tag{22}$$

If v_i is defined as $p_i(1-p_i)$ and $\mathbf{V} = diag(v_1, ...v_n)$ then the Hessian matrix can be expressed as

$$\mathbf{H}(\boldsymbol{\beta}) = \nabla^2_{\boldsymbol{\beta}} \ln \mathbb{L}(\boldsymbol{\beta}) = -\mathbf{X}^{\mathrm{T}} \mathbf{V} \mathbf{X}. \tag{23}$$

Since the Hessian matrix is negative definite, then the objective function is strictly concave, with one global maximum. The LR *information matrix* is given by

$$\mathbf{I}(\boldsymbol{\beta}) = -E[\mathbf{H}(\boldsymbol{\beta})] = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X}. \tag{24}$$

The variance of $\boldsymbol{\beta}$ is then $\mathbf{V}(\boldsymbol{\beta}) = \mathbf{I}(\boldsymbol{\beta})^{-1} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$.

Over-fitting the training data may arise in LR [31], especially when the data are very high dimensional and/or sparse. One of the approaches to reduce over-fitting is through quadratic *regularization*, known also as *ridge regression*, which introduces a penalty for large values of $\boldsymbol{\beta}$ and to obtain better generalization [8]. The regularized log-likelihood can be defined as

$$\ln \mathbb{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \ln \left(\frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} \right) + (1-y_i) \ln \left(\frac{1}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} \right) \right) - \frac{\lambda}{2}||\boldsymbol{\beta}||^2 \tag{25}$$

$$= \sum_{i=1}^{n} \ln \left(\frac{e^{y_i\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} \right) - \frac{\lambda}{2}||\boldsymbol{\beta}||^2, \tag{26}$$

where $\lambda > 0$ is the regularization parameter and $\frac{\lambda}{2}||\boldsymbol{\beta}||^2$ is the regularization (penalty) term. For binary outputs, the loss function or the deviance (DEV), also useful for measuring the goodness-of-fit of the model, is the negative log-likelihood and is given by the formula [31, 42]

$$\mathbb{DEV}(\hat{\boldsymbol{\beta}}) = -2 \ln \mathbb{L}(\boldsymbol{\beta}). \tag{27}$$

Minimizing the deviance $\mathbb{DEV}(\hat{\boldsymbol{\beta}})$ given in (27) is equivalent to maximizing the log-likelihood [31]. Recent studies showed that the *Conjugate Gradient* (CG) method, when applied to the method of *Iteratively Re-weighted Least Squares* (IRLS) provides better results to estimate $\boldsymbol{\beta}$ than any other numerical method [51, 55].

3 Iteratively Re-weighted Least Squares

One of the most popular techniques used to find the MLE of $\boldsymbol{\beta}$ is the iteratively re-weighted least squares (IRLS) method, which uses Newton-Raphson algorithm to solve LR score equations. Each iteration finds the *Weighted Least Squares* (WLS) estimates for a given set of weights, which are used to construct a new set of weights [26]. The gradient and the Hessian are obtained by differentiating the regularized likelihood in (26) with respect to $\boldsymbol{\beta}$, obtaining, in matrix form

$$\nabla_{\beta} \ln \mathbb{L}(\beta) = \mathbf{X}^{\mathrm{T}}(\mathbf{y}-\mathbf{p}) - \lambda\boldsymbol{\beta} = \mathbf{0}, \tag{28}$$

$$\nabla^2_{\beta} \ln \mathbb{L}(\boldsymbol{\beta}) = -\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} - \lambda\mathbf{I}, \tag{29}$$

where $\mathbf{I}$ is a $d \times d$ identity matrix. Now that the first and second derivatives are obtained, the Newton-Raphson update formula on the $(c+1)-th$ iteration is

given by

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \hat{\boldsymbol{\beta}}^{(c)} + (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{p}) - \lambda\hat{\boldsymbol{\beta}}^{(c)}). \tag{30}$$

Since $\hat{\boldsymbol{\beta}}^{(c)} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})\hat{\boldsymbol{\beta}}^{(c)}$, then (30) can be rewritten as

$$\hat{\boldsymbol{\beta}}^{(c+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{V}\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + (\mathbf{y} - \mathbf{p})) \tag{31}$$

$$= (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{z}^{(c)}, \tag{32}$$

where $\mathbf{z}^{(c)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{V}^{-1}(\mathbf{y} - \mathbf{p})$ and is referred to as the adjusted response [27].

Despite the advantage of the regularization parameter, λ, in forcing positive definiteness, if the matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})$ were dense, the iterative computation could become unacceptably slow [42]. This necessitates the need for a "trade off" between convergence speed and accurate Newton direction [44]. The method which provides such a trade-off is known as the truncated Newton's method.

3.1 TR-IRLS Algorithm

The WLS subproblem, $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{z}^{(c)}$, is a linear system of d equations and variables, and solving it is equivalent to minimizing the quadratic function

$$\frac{1}{2}\hat{\boldsymbol{\beta}}^{(c+1)}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})\hat{\boldsymbol{\beta}}^{(c+1)} - \hat{\boldsymbol{\beta}}^{(c+1)}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{z}^{(c)}). \tag{33}$$

Komarek and Moore [43] were the first to implement a modified linear CG to approximate the Newton direction in solving the IRLS for LR. This technique is called *Truncated-Regularized Iteratively-Reweighted Least Squares* (TR-IRLS). The main advantage of the CG method is that it guarantees convergence in at most d steps [44]. The TR-IRLS algorithm consists of two loops. Algorithm 1 represents the outer loop which finds the solution to the WLS problem and is terminated when the relative difference of deviance between two consecutive iterations is no larger than a specified threshold ε_1. Algorithm 2 represents the inner loop, which solves the WLS subproblems in Algorithm 1 through the linear CG method, which is the Newton direction. Algorithm 2 is terminated when the residual

$$\mathbf{r}^{(c+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{I})\hat{\boldsymbol{\beta}}^{(c+1)} - \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{z}^{(c)}$$

is no greater than a specified threshold ε_2. For more details on the TR-IRLS algorithm and implementation, see Komarek [42].

Default parameter values are given for both algorithms [43] and are shown to provide adequate accuracy on very large datasets. For Algorithm 1, the maximum number of iterations is set to 30 and the relative difference of deviance threshold, ε_1, is set to 0.01. For Algorithm 2, the ridge regression parameter, λ, is set to 10 and the maximum number of iterations for the CG is set to 200 iterations. In addition, the CG convergence threshold, ε_2, is set to 0.005, and no more than three non-improving iterations are allowed on the CG algorithm.

Algorithm 1: LR MLE using IRLS

Data: $\mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\beta}}^{(0)}$
Result: $\hat{\boldsymbol{\beta}}$
begin
  $c = 0$
  while $|\frac{DEV^{(c)} - DEV^{(c+1)}}{DEV^{(c+1)}}| > \varepsilon_1$ **and** $c \leq$ *Max IRLS Iterations* **do**
    for $i \leftarrow 1$ **to** n **do**
      $\hat{p}_i = \frac{1}{1+e^{-\mathbf{x}_i\hat{\boldsymbol{\beta}}}}$; /* Compute probabilities */
      $v_i = \hat{p}_i(1-\hat{p}_i)$; /* Compute weights */
      $z_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(c)} + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}$; /* Compute the adjusted response */
    $\mathbf{V} = diag(v_1, ..., v_n)$
    $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{X})\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{z}^{(c)}$; /* Compute $\hat{\boldsymbol{\beta}}$ via WLS */
    $c = c + 1$
end

Algorithm 2: Linear CG. $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X} + \lambda\mathbf{X}$, $\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{z}$

Data: $\mathbf{A}, \mathbf{b}, \hat{\boldsymbol{\beta}}^{(0)}$
Result: $\hat{\boldsymbol{\beta}}$ such that $\mathbf{A}\hat{\boldsymbol{\beta}} = \mathbf{b}$
begin
  $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\hat{\boldsymbol{\beta}}^{(0)}$; /* Initialize the residual */
  $c = 0$
  while $||\mathbf{r}^{(c+1)}||^2 > \varepsilon_2$ **and** $c \leq$ *Max CG Iterations* **do**
    if $c = 0$ **then**
      $\zeta^{(c)} = 0$
    else
      $\zeta^{(c)} = \frac{\mathbf{r}^{\mathrm{T}(c+1)}\mathbf{r}^{(c+1)}}{\mathbf{r}^{\mathrm{T}(c+1)}\mathbf{r}^{(c)}}$; /* Update $\mathbf{A}$-Conjugacy enforcer */
    $\mathbf{d}^{(c+1)} = \mathbf{r}^{(c+1)} + \zeta^{(c)}\mathbf{d}^{(c)}$; /* Update the search direction */
    $s^{(c)} = \frac{\mathbf{r}^{\mathrm{T}(c)}\mathbf{r}^{(c)}}{\mathbf{d}^{\mathrm{T}(c)}\mathbf{A}\mathbf{d}^{(c)}}$; /* Compute the optimal step length */
    $\hat{\boldsymbol{\beta}}^{(c+1)} = \hat{\boldsymbol{\beta}}^{(c)} + \zeta^{(c)}\mathbf{d}^{(c+1)}$; /* Obtain approximate solution */
    $\mathbf{r}^{(c+1)} = \mathbf{r}^{(c)} - s^{(c)}\mathbf{A}\mathbf{d}^{(c+1)}$; /* Update the residual */
    $c = c + 1$
end

Once the optimal MLE for $\hat{\boldsymbol{\beta}}$ are found, classification of any given $i-th$ instance, $\mathbf{x}_i$, is carried out according to the following rules

$$\hat{y}_i = \begin{cases} 1, & \text{if } \hat{\eta}_i \geq 0 \text{ or } \hat{p}_i \geq 0.5 \ ; \\ 0, & \text{otherwise.} \end{cases} \tag{34}$$

Aside from the implementation simplicity of TR-IRLS, the main advantage of the algorithm is that it can process and classify large datasets with little time compared to other methods such as SVM. In addition, the TR-IRLS is robust to linear dependencies and data scaling, and its accuracy is comparable to that of

SVM. Furthermore, the algorithm does not require parameter tuning. This is an important characteristic when the goal is to classify large and balanced datasets.

Despite all of the aforementioned advantages of TR-IRLS, the algorithm is not designed to handle rare events data, and it is not designed to handle small-to-medium size datasets that are highly non-linearly separable [47].

4 Logistic Regression in Imbalanced and Rare Events Data

4.1 Endogenous (Choice-Based) Sampling

Almost all of the conventional classification methods are based on the assumption that the training data consist of examples drawn from the same distribution as the testing data (or real-life data) [65, 68]. Likewise in *Generalized Linear Models* (GLM), likelihood functions solved by methods such as LR are based on the concepts of random sampling or *exogenous* sampling [39, 67]. To see why this is the case [3, 10], under random sampling, the true joint distribution of $\mathbf{y}$ and $\mathbf{X}$ is $P(\mathbf{y}|\mathbf{X})P(\mathbf{X})$, and the likelihood function based on n binary observations is given by

$$\mathbb{L}_{Random} = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \boldsymbol{\beta})P(\mathbf{x}_i). \tag{35}$$

Under exogenous sampling, the sampling is on $\mathbf{X}$ according to a distribution $f(\mathbf{X})$, which may not reflect the actual distribution $P(\mathbf{X})$, and then $\mathbf{y}$ is sampled according to its true distribution probability $P(\mathbf{y}|\mathbf{X})$. The likelihood function would then be

$$\mathbb{L}_{Exogenous} = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \boldsymbol{\beta})f(\mathbf{x}_i). \tag{36}$$

As long as the ML estimator is not related to $P(\mathbf{X})$ or $f(\mathbf{X})$, then maximizing $\mathbb{L}_{Random}$ or $\mathbb{L}_{Exogenous}$ is equivalent to maximizing

$$\mathbb{L} = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \boldsymbol{\beta}), \tag{37}$$

which is exactly the likelihood maximized by LR in (14) [68].

While the ML method is the most important method of estimation with a great advantage in general applicability, it is well-known that MLE of the unknown parameters, with exception to the normal distribution, are *asymptotically biased* in small samples. The ML properties are satisfied mainly asymptotically, meaning with the assumption of large samples [26, 12]. In addition, while it is ideal that sampling be either random or exogenous since it is reflective of the population or the testing data distribution, this sampling strategy has three major disadvantages when applied to REs. First, in data collection surveys, it would be very time consuming and costly to collect data on events that occur rarely. Second, in data

mining, the data to be analyzed could be very large in order to contain enough REs, and hence computational time could be big. Furthermore, while the ML estimator is consistent in analyzing such data, it is asymptotically biased in the sense that the probabilities generated underestimate the actual probabilities of occurrence. In other words, the results are asymptotically biased. Cox and Hinkley [18] provided a general rough approximation for the asymptotic bias, developed originally by Cox and Snell [19], such that

$$E[\hat{\beta}-\beta]=-\frac{1}{2n}\frac{i_{30}+i_{11}}{i_{20}^2}, \tag{38}$$

where $i_{30}=E\left[\left(\frac{\partial L}{\partial \beta}\right)^3\right]$, $i_{11}=E\left[\left(\frac{\partial L}{\partial \beta}\right)\left(\frac{\partial^2 L}{\partial \beta^2}\right)\right]$, and $i_{20}=E\left[\left(\frac{\partial L}{\partial \beta}\right)^2\right]$, are evaluated at $\hat{\beta}$. Following King and Zeng [39], if $p_i=\frac{1}{1+e^{-\beta_0+x_i}}$, then the asymptotic bias is

$$E[\hat{\beta}_0-\beta_0]=-\frac{1}{n}\frac{E[(0.5-\hat{p}_i)((1-\hat{p}_i)^2y_i+\hat{p}_i^2(1-y_i))]^2}{(E[(1-\hat{p}_i)^2y_i+\hat{p}_i^2(1-y_i)])} \tag{39}$$

$$\approx \frac{\overline{p}-0.5}{n\overline{p}(1-\overline{p})}, \tag{40}$$

where $\overline{p}$ is the proportion of events in the sample. Therefore, as long as $\overline{p}$ is less than 0.5 and/or n is small, the bias in (40) will not be equal to zero. Furthermore, the variance would be large. To see this mathematically, consider the variance matrix of the LR estimator, $\hat{\boldsymbol{\beta}}$, given by

$$\mathbf{V}(\hat{\boldsymbol{\beta}})=\left[\sum_{i=1}^{n}p_i(1-p_i)\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i\right]^{-1}. \tag{41}$$

The variance given in (41) is smallest when the part $p_i(1-p_i)$, which is affected by rare events, is closer to 0.5. This occurs when the number of ones is large enough in the sample. However, the estimate of p_i with observations related to rare events is usually small, and hence additional ones would cause the variance to drop while additional zeros at the expense of events would cause the variance to increase [39, 58]. The strategy is to select on $\mathbf{y}$ by collecting observations for which $y_i=1$ (the cases), and then selecting random observations for which $y_i=0$ (the controls). The objective then is to keep the variance as small as possible by keeping a balance between the number of events (ones) and non-events (zeros) in the sample under study. This is achieved through *endogenous* sampling or $choice-based$ sampling. Endogenous sampling occurs whenever sample selection is based on the dependent variable ($\mathbf{y}$), rather than on the independent (exogenous) variable ($\mathbf{X}$).

However, since the objective is to derive inferences about the population from the sample, the estimates obtained by the common likelihood using pure endogenous sampling are inconsistent. King and Zeng [39] recommend two methods of estimation for choice-based sampling, *prior correction* and *weighting*.

4.2 Correcting Estimates Under Endogenous Sampling

4.2.1 Prior Correction

Consider a population of N examples with τ as the proportion of events and $(1-\tau)$ as the proportion of non-events. Let the event of interest be $y=1$ in the population with a probability $\tilde{p}$. Let n be the sample size with $\overline{y}$ and $(1-\overline{y})$ representing the proportions of events and non-events in the sample, respectively. Then, let $\hat{p}$ be the probability of the event in the sample, and $s=1$ be a selected event. By the Bayesian formula [58, 20],

$$\hat{p} = P(y=1|s=1) = \frac{P(s=1|y=1)P(y=1)}{P(s=1|y=1)P(y=1) + P(s=1|y=0)P(y=0)} \tag{42}$$

$$= \frac{\left(\frac{\overline{y}}{\tau}\right)\tilde{p}}{\left(\frac{\overline{y}}{\tau}\right)\tilde{p} + \left(\frac{1-\overline{y}}{1-\tau}\right)(1-\tilde{p})}. \tag{43}$$

If the sample is random, then $\overline{y}=\tau$ and $1-\overline{y}=1-\tau$, hence $\hat{p}=\tilde{p}$ and there is no inconsistency. When endogenous sampling is used to analyze imbalanced or rare events data, $\tau<(1-\tau)$, and $\overline{y}\approx(1-\overline{y})$, and hence $\hat{p}\neq\tilde{p}$, regardless of the sample size.

Now, assuming that $\hat{p}$ and $\tilde{p}$ are *logit* probabilities, and let $\nu_1 = \left(\frac{\overline{y}}{\tau}\right)$, and $\nu_0 = \left(\frac{1-\overline{y}}{1-\tau}\right)$, then equation (43) can be rewritten as

$$\hat{p} = \frac{\nu_1\tilde{p}}{\nu_1\tilde{p} + \nu_0(1-\tilde{p})}. \tag{44}$$

The odd of (44) is then

$$O = \frac{\hat{p}}{1-\hat{p}} = \frac{\nu_1\tilde{p}}{\nu_0(1-\tilde{p})}, \tag{45}$$

and the log odds is

$$\ln(O) = \ln\left(\frac{\nu_1}{\nu_0}\right) + \ln(\tilde{p}) - \ln(1-\tilde{p}), \tag{46}$$

which implies that

$$\mathbf{x}\hat{\boldsymbol{\beta}} = \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right] + \mathbf{x}\tilde{\boldsymbol{\beta}}. \tag{47}$$

Prior correction is therefore easy to apply as it involves only correcting the intercept [39, 20], β_0, such that

$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right], \tag{48}$$

thereby making the corrected logit probability be

$$\tilde{p}_i = \frac{1}{1 + e^{\ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right] - \mathbf{x}_i\boldsymbol{\beta}}}, \qquad \text{for } i = 1 \dots n. \tag{49}$$

Prior correction requires knowledge of the fraction of events in the population, τ. The advantage of prior correction is its simplicity. However, the main disadvantage of this correction is that if the model is misspecified, then estimates on both $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ are less robust than weighting [39, 67].

4.2.2 Weighting

Under pure endogenous sampling, the conditioning is on $\mathbf{X}$ rather than $\mathbf{y}$ [10, 54], and the joint distribution of $\mathbf{y}$ and $\mathbf{X}$ in the sample is

$$f_s(\mathbf{y}, \mathbf{X}|\boldsymbol{\beta}) = P_s(\mathbf{X}|\mathbf{y}, \boldsymbol{\beta}) P_s(\mathbf{y}), \tag{50}$$

where β is the unknown parameter to be estimated. Yet, since $\mathbf{X}$ is a matrix of exogenous variables, then the conditional probability of $\mathbf{X}$ in the sample is equal to that in the population, or $P_s(\mathbf{X}|\mathbf{y}, \boldsymbol{\beta}) = P(\mathbf{X}|\mathbf{y}, \boldsymbol{\beta})$. However, the conditional probability in the population is

$$P(\mathbf{X}|\mathbf{y}, \boldsymbol{\beta}) = \frac{f(\mathbf{y}, \mathbf{X}|\boldsymbol{\beta})}{P(\mathbf{y})}, \tag{51}$$

but

$$f(\mathbf{y}, \mathbf{X}|\boldsymbol{\beta}) = P(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) P(\mathbf{X}), \tag{52}$$

and hence, substituting and rearranging yields

$$f_s(\mathbf{y}, \mathbf{X}|\boldsymbol{\beta}) = \frac{P_s(\mathbf{y})}{P(\mathbf{y})} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) P(\mathbf{X}) \tag{53}$$

$$= \frac{H}{Q} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) P(\mathbf{X}), \tag{54}$$

where $\dfrac{H}{Q} = \dfrac{P_s(\mathbf{y})}{P(\mathbf{y})}$. The likelihood is then

$$\mathbb{L}_{Endogenous} = \prod_{i=1}^{n} \frac{H_i}{Q_i} P(y_i|\mathbf{x}_i, \boldsymbol{\beta}) P(\mathbf{x}_i), \tag{55}$$

where $\dfrac{H_i}{Q_i} = \left(\dfrac{\overline{y}}{\tau}\right) y_i + \left(\dfrac{1-\overline{y}}{1-\tau}\right)(1 - y_i)$. Therefore, when dealing with REs and imbalanced data, it is the likelihood in (55) that needs to be maximized [3, 67, 10, 52, 33]. Several consistent estimators of this type of likelihood have been proposed in the literature. Amemiya [3] and Ben Akiva and Lerman [4] provide an excellent survey of these methods.

Manski and Lerman [52] proposed the *Weighted Exogenous Sampling Maximum Likelihood* (WESML), and proved that WESML yields a consistent and asymptotically normal estimator so long as knowledge of the population probability is available. More recently, Ramalho and Ramalho [60] extended the work of Manski and Lerman [52] to cases where such knowledge may not be available. Knowledge of population probability or proportions, however, can be acquired from previous surveys or existing databases. The log-likelihood for LR can then be rewritten as

$$\ln \mathbb{L}(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} \frac{Q_i}{H_i} \ln P(y_i|\mathbf{x}_i,\boldsymbol{\beta}) \tag{56}$$

$$= \sum_{i=1}^{n} \frac{Q_i}{H_i} \ln \left(\frac{e^{y_i \mathbf{x}_i \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right) \tag{57}$$

$$= \sum_{i=1}^{n} w_i \ln \left(\frac{e^{y_i \mathbf{x}_i \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i \boldsymbol{\beta}}} \right), \tag{58}$$

where $w_i = \dfrac{Q_i}{H_i}$. Thus, in order to obtain consistent estimators, the likelihood is multiplied by the inverse of the fractions. The intuition behind weighting is that if the proportion of events in the sample is more than that in the population, then the ratio $\left(\dfrac{Q}{H}\right) < 1$ and hence the events are given less weight, while the non-events would be given more weight if their proportion in the sample is less than that in the population. This estimator, however, is not fully efficient, because the information matrix equality does not hold. This is demonstrated as

$$-E\left[\frac{Q}{H}\nabla^2_{\boldsymbol{\beta}} \ln P(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})\right] \neq E\left[\left(\frac{Q}{H}\nabla_{\boldsymbol{\beta}} \ln P(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})\right)\left(\frac{Q}{H}\nabla_{\boldsymbol{\beta}} \ln P(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})\right)^{\mathrm{T}}\right], \tag{59}$$

and for the LR model it is

$$-\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_i}{H_i}\right) p_i(1-p_i)\mathbf{x}_i\mathbf{x}_j\right] \neq \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_i}{H_i}\right)^2 p_i(1-p_i)\mathbf{x}_i\mathbf{x}_j\right]. \tag{60}$$

Let $\mathbf{A} = \dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{Q_i}{H_i}\right) p_i(1-p_i)\mathbf{x}_i\mathbf{x}_j$, and $\mathbf{B} = \dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{Q_i}{H_i}\right)^2 p_i(1-p_i)\mathbf{x}_i\mathbf{x}_j$, then the asymptotic variance matrix of the estimator $\boldsymbol{\beta}$ is given by the *sandwich estimate*, such that $\mathbf{V}(\boldsymbol{\beta}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ [3, 67, 52].

Now that consistent estimators are obtained, finite-sample/rare-event bias corrections could be applied. King and Zeng [39] extended the small-sample bias corrections, as described by McCullagh and Nelder [53], to include the weighted likelihood (58), and demonstrated that even with choice-based sampling, these corrections can make a difference when the population probability of the event of interest is low. According to McCullagh and Nelder [53], and later Cordeiro and McCullagh [15], the bias vector is given by

$$bias(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}\xi, \tag{61}$$

where $\xi_i = Q_{ii}(\hat{p_i} - \frac{1}{2})$, and Q_{ii} are the diagonal elements of $\mathbf{Q} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, which is the approximate covariance matrix of the logistic link function $\boldsymbol{\eta}$. The second-order bias-corrected estimator is then

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - bias(\hat{\boldsymbol{\beta}}). \tag{62}$$

As for the variance matrix $\mathbf{V}(\tilde{\boldsymbol{\beta}})$ of $\tilde{\boldsymbol{\beta}}$, it is estimated using

$$\mathbf{V}(\tilde{\boldsymbol{\beta}}) = \left(\frac{n}{n+d}\right)^2 \mathbf{V}(\hat{\boldsymbol{\beta}}). \tag{63}$$

Since $\left(\frac{n}{n+d}\right)^2 < 1$, then $\mathbf{V}(\tilde{\boldsymbol{\beta}}) < \mathbf{V}(\hat{\boldsymbol{\beta}})$, and hence both the variance and the bias are now reduced.

The main advantage then of the bias correction method proposed by McCullagh and Nelder [53] is that it reduces both the bias and the variance [39]. The disadvantage of this bias correction method is that it is corrective and not preventive, since it is applied after the estimation is complete, and hence it does not protect against infinite parameter values that arise from perfect separation between the classes [28, 66]. Hence, this bias correction method can only be applied if the estimator, $\hat{\boldsymbol{\beta}}$, has finite values. Firth [23] proposed a preventive second-order bias correction method by penalizing the log-likelihood such that

$$\ln \mathbb{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln\left(\frac{e^{y_i \mathbf{x}_i \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i \boldsymbol{\beta}}}\right) + \frac{1}{2}\ln|\mathbf{I}(\boldsymbol{\beta})|, \tag{64}$$

which leads to a modified score equation given by

$$\frac{\partial}{\partial \beta_j} \ln \mathbb{L}(\beta) = \sum_{i=1}^{n} (x_{ij}(y_i - p_i + h_i(0.5 - p_i))) = 0, \tag{65}$$

where h_i is the i-th diagonal element of the hat matrix

$$\mathbf{H} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{\frac{1}{2}}. \tag{66}$$

A recent comparative simulation study by Maiti and Pradhan [50] showed that the bias correction of McCullagh and Nelder [53], provides the smallest *Mean Squared Error* (MSE) when compared to that of Firth [23] and others using LR. Cordeiro and Barroso [14] more recently derived a third-order bias corrected estimator and showed that in some cases it could deliver improvements in terms of bias and MSE over the usual ML estimator and that of Cordeiro and McCullagh [15].

Now, as mentioned earlier, LR regularization is used in the form of the ridge penalty $\frac{\lambda}{2}||\boldsymbol{\beta}||^2$. When regularization is introduced, none of the coefficients is set to zero [59], and hence the problem of infinite parameter values is avoided. In addition, the importance of the parameter λ lies in determining the bias-variance trade-off of an estimator [17, 49]. When λ is very small, there is less bias but more variance. On the other hand, larger values of λ would lead to more bias

but less variance [5]. Therefore, the inclusion of regularization in the LR model is very important to reduce any potential inefficiency. However, as regularization carries the risk of a non-negligible bias, even asymptotically [5], the need for bias correction becomes inevitable [48]. In sum, bias correction is needed to account for any bias resulting from regularization, small samples, and rare events.

The challenge remains on finding the best class distribution in the training dataset. First, when both the events and non-events are easy to collect and both are available, then a sample with equal number of ones and zeros would be generally optimum [16, 32]. Second, when the number of events in the population is very small, the decision is then how many more non-events to collect in addition to the events. If collecting more non-events is inexpensive, then the general judgment is to collect as many non-events as possible. However, as the number of non-events exceed the number of events, the marginal contribution to the explanatory variables' information content starts to drop, and hence the number of zeros should be no more than two to five times the number of ones [39].

Applying the above corrections, offered by King and Zeng [39], along with the recommended sampling strategies, such as collecting all of the available events and only a matching proportion of non-events, could (1) significantly decrease the sample size under study, (2) cut data collection costs, (3) increase the rare event probability, and, (4) enable researchers to focus more on analyzing the variables.

5 LR and Other Data Mining Challenges

Data quality problems, such as noise, missing data, collinearity, redundant attributes, and over-fitting, are common in data mining. Several strategies and LR models have been proposed in the literature to deal with those issues. Noise is defined as any random error in the data mining process [56]. Bi and Jeske [7] compared LR to the Normal Discriminant Analysis (NDA) method and showed that LR is more efficient and less deteriorated than NDA in the presence of class-conditional classification noise (CCC-Noise). With regard to missing data, Hotron and Kleinman [30] described different methods to fit LR models with missing data and reviewed their implementation using general-purpose statistical software. Applying LR models in missing data has also been addressed by Agresti [1].

The problem of LR in the presence of collinearity (when two or more variables are highly correlated) has been addressed in Rossi [61]. The Stepwise Logistic Regression method, well described by Hosmer [31], is an effective method for *feature selection* and for reducing the number of redundant and/or irrelevant variables. The method is relatively easy to apply as it includes or excludes variables based on the resulting deviance of the fitted model that is associated with those variables. Active learning, a recent and popular method for reducing the size of the training data based on incremental learning, is very useful for mining massive data sets, specially when labeling can be expensive [57]. Schein [63] provides an excellent evaluation of different active learning techniques for LR. To achieve good generalization and to avoid the problem of over-fitting, an n-fold cross-validation (ten folds are usually adequate) is performed [21]. The n-fold cross-validation divides the data set into n folds, retaining $n-1$ folds for training while the remaining fold is used for testing. This is done iteratively until all of the folds

have been used as testing data sets [6]. The results are then averaged to produce the final accuracy. For data with dichotomous, polychotomous (multinomial), and continuous independent variables, Agresti [2] and Hosmer [31] provide the appropriate analyses and LR models.

Boosting [9] is an important machine-learning method and is used to improve the accuracy of any given classifier. Iterative algorithms such as AdaBoost [24] assign different weights to the training distribution in each iteration. After each iteration, boosting increases the weights associated with incorrectly classified examples and decreases the weights associated with the correctly classified ones. A variant of AdaBoost, AdaCost [22], has been shown useful in addressing the problem of rarity and imbalance in data. Analysis of boosting techniques, however, showed that boosting is tied to the choice of the base learning algorithm [36, 35]. Thus, if the base learning algorithm is a good classifier without boosting, then boosting would be useful when that base learner is used in REs. With regard to LR, Friedman et al. [25] suggested using the logistic function to estimate the probabilities of the AdaBoost output. The authors showed that the properties of their proposed method, LogitBoost, are identical to those of AdaBoost. Collins et al. [13] proposed a more direct modification of AdaBoost for the logistic loss function by deriving their algorithm using a unification of LR and boosting based on Bregman distances. Furthermore, they also generalized the algorithm for multiple classes.

LR linearity may be an obstacle to handling highly nonlinearly separable small-to-medium size data sets [42]. When analyzing highly non-linear data sets, the underlying assumption of linearity in the LR model, as evident in its logit function, is often violated [27]. With the advancement of kernel methods, the search for an effective non-parametric LR model, capable of classifying non-linearly separable data, has become possible. Like LR, Kernel Logistic Regression (KLR) can naturally provide probabilities and extend to multi-class classification problems [27, 37]. Interested readers should consult papers written on Kernel Logistic Regression (KLR) [11, 34, 27, 38] and how some of the methods described in this overview can be effectively extended to KLR [62, 47, 48].

6 Conclusions

Logistic regression provides a great means for modeling binary as well as multiple class response variable dependence on one or more independent variables. Those independent variables could be categorical or continuous, or both. The fit of the resulting model can be assessed using a number of methods, the most important of which is the IRLS method, which in turn is best solved using the CG method in the form of the truncated Newton method. Furthermore, with regard to imbalanced and rare events data sets, certain sampling strategies and appropriate corrections should be applied to the LR method. The most common correction techniques are *prior correction* and *weighting*.

In addition, LR is adaptable to handle other data mining challenges, such as the problems of collinearity, missing data, redundant attributes and nonlinear separability, among others, making LR a powerful and resilient data mining method. It is our hope that this overview of the LR method and the developed

state-of-the-art techniques in it, as provided by the literature, would shed further light on this method as well as encourage and direct future theoretical and applied research in it.

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2 edition, 2002.

[2] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley-Interscience, 2007.

[3] T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[4] M. Ben-Akiva and S. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, 1985.

[5] R. Berk. *Statistical Learning from a Regression Perspective*. Springer, 1st edition, 2008.

[6] M. R. Berthold and D. J. Hand, editors. *Intelligent Data Analysis*. Springer, 2 edition, 2010.

[7] Y. Bi and D. R. Jeske. The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. *Journal of Multivariate Analysis*, 101(7):1622–1637, 2010.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta. *Metalearning: Applications to Data Mining*. Springer, 1 edition, 2008.

[10] A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005.

[11] S. Canu and A. J. Smola. Kernel methods and the exponential family. In *ESANN*, pages 447–454, 2005.

[12] D. Collett. *Modelling Binary Data*. Chapman & Hall/CRC, 2nd edition, 2003.

[13] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48:253–285, 2002.

[14] G. Cordeiro and L. Barroso. A third-order bias corrected estimate in generalized linear models. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 16(1):76–89, 2007.

[15] G. M. Cordeiro and P. McCullagh. Bias correction in generalized linear models. *Journal of Royal Statistical Society*, 53(3):629–643, 1991.

[16] Stephen R. Cosslett. *Structural Analysis of Discrete Data and Econometric Applications*. Cambridge: The MIT Press, 1981.

[17] G. Cowan. *Statistical Data Analysis*. Oxford University Press, 1998.

[18] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall/CRC, 1979.

[19] D. R. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society*, 30(2):248–275, 1968.

[20] J. S. Cramer. *Logit Models From Economics And Other Fields*. Cambridge University Press, 2003.

[21] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

[22] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan. Adacost: misclassification cost-sensitive boosting. In *Prococeedings to the 16th International Conference on Machine Learning*, pages 97–105. Morgan Kaufmann, 1999.

[23] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

[24] Y. Freund and R. E. Schapire. A brief introduction to boosting. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406. Morgan Kaufmann, 1999.

[25] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 38(2):337–374, 2000.

[26] P. Garthwaite, I. Jolliffe, and J. Byron. *Statistical Inference*. Oxford University Press, 2002.

[27] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, 2 edition, 2009.

[28] G. Heinze and M. Schemper. A solution to the problem of monotone likelihood in cox regression. *Biometrics*, 57:114–119, 2001.

[29] J. M. Hilbe. *Logistic Regression Models*. Chapman & Hall/CRC, 2009.

[30] N. J. Horton and K. P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.

[31] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, second edition, 2000.

[32] G. W. Imbens. An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, 60(5):1187–214, September 1992.

[33] G. W. Imbens and T. Lancaster. Efficient estimation and stratified sampling. *Journal of Econometrics*, 74:289–318, 1996.

[34] T. Jaakkola and D. Haussler. Probabilistic kernel regression models, 1999.

[35] M. V. Joshi, R. C. Agarwal, and V. Kumar. Predicting rare classes: can boosting make any weak learner strong? In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 297–306, New York, NY, USA, 2002. ACM.

[36] M. V. Joshi, V. Kumar, and R. C. Agarwal. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 257–264, Washington, DC, USA, 2001.

[37] P. Karsmakers, K. Pelckmans, and J. A. K. Suykens. Multi-class kernel logistic regression: a fixed-size implementation. *International Joint Conference on Neural Networks*, pages 1756–1761, 2007.

[38] S. S. Keerthi, K. B. Duan, S. K. Shevade, and A. N. Poo. A fast dual algorithm for kernel logistic regression. *Mach. Learn.*, 61(1-3):151–165, 2005.

[39] G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, 2001.

[40] D. G. Kleinbaum, L. L. Kupper, A. Nizam, and K. E. Muller. *Applied Regression Analysis and Multivariable Methods*. Duxbury Press, 4 edition, 2007.

[41] K. Koh, S. Kim, and S. Boyd. An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

[42] P. Komarek. *Logistic Regression for Data Mining and High-Dimensional Classification*. PhD thesis, Carnegie Mellon University, 2004.

[43] P. Komarek and A. Moore. Making logistic regression a core data mining tool: A practical investigation of accuracy, speed, and simplicity. Technical report, Carnegie Mellon University, 2005.

[44] J. M. Lewis, S. Lakshmivarahan, and S. Dhall. *Dynamic Data Assimilation: A Least Squares Approach*. Cambridge University Press, 2006.

[45] C. Lin, R. C. Weng, and S. S. Keerthi. Trust region newton methods for large-scale logistic regression. *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[46] J. S. Long and J. Freese. *Regression Models for Categorical Dependent Variables Using Stata*. Stata Press, 2 edition, 2005.

[47] M. Maalouf and T. B. Trafalis. Kernel logistic regression using truncated newton method. In C. H. Dagli, D. L. Enke, K. M. Bryden, H. Ceylan, and M. Gen, editors, *Intelligent Engineering Systems Through Artificial Neural Networks*, volume 18, pages 455–462, New York, NY, USA, 2008. ASME Press.

[48] M. Maalouf and T. B. Trafalis. Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, 55(1):168–183, 2011.

[49] O. Maimon and L. Rokach, editors. *Data Mining and Knowledge Discovery Handbook*. Springer, 2005.

[50] T. Maiti and V. Pradhan. A comparative study of the bias corrected estimates in logistic regression. *Statistical Methods in Medical Research*, 17(6):621–634, 2008.

[51] R. Malouf. A comparison of algorithms for maximum etropy parameter estimation. In *Proceedings of Conference on Natural Language Learning*, volume 6, 2002.

[52] C. F. Manski and S. R. Lerman. The estimation of choice probabilities from choice based samples. *Econometrica*, 45(8):1977–88, 1977.

[53] P. McCullagh and J. Nelder. *Generalized Linear Model*. Chapman and Hall/CRC, 1989.

[54] M. Milgate, J. Eatwell, and P. K. Newman, editors. *Econometrics*. W. W. Norton & Company, 1990.

[55] T. P. Minka. A comparison of numerical optimizers for logistic regression. Technical report, Deptartment of Statistics, Carnegie Mellon University, 2003.

[56] T. Mitsa. *Temporal Data Mining*. Chapman & Hall/CRC, 2010.

[57] S. K. Pal and P. Mitra. *Pattern recognition algorithms for data mining*. Chapman and Hall/CRC, 1 edition, 2004.

[58] K. Palepu. Predicting takeover targets: A methodological and empirical analysis. *Journal of Accounting and Economics*, 8:3–35, 1986.

[59] M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2008.

[60] E. A. Ramalho and J. J. S. Ramalho. On the weighted maximum likelihood estimator for endogenous stratified samples when the population strata probabilities are unknown. *Applied Economics Letters*, 14:171–174, 2007.

[61] R. J. Rossi. *Applied Biostatistics for the Health Sciences*. Wiley, 1 edition, 2009.

[62] V. Roth. Probabilistic discriminative kernel classifiers for multi-class problems. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, pages 246–253, London, UK, 2001. Springer-Verlag.

[63] A. I. Schein and L. H. Ungar. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

[64] V. Vapnik. *The Nature of Statistical Learning*. Springer, NY, 1995.

[65] S. Visa and A. Ralescu. Issues in mining imbalanced data sets - a review paper. In *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, 2005*, pages 67–73, 2005.

[66] S. Wang and T. Wang. Precision of warm's weighted likelihood for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, 25(4):317–331, 2001.

[67] Y. Xie and C. F. Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17:283–302, 1989.

[68] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 114, New York, NY, USA, 2004. ACM.