Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 201 (2017) 779–788

Procedia Engineering

www.elsevier.com/locate/procedia

3rd International Conference "Information Technology and Nanotechnology", ITNT-2017, 25-27 April 2017, Samara, Russia

Minimizing the effects of floor and ceiling to improve the convergence of log-likelihood

Ilona Kulikovskikh[a,]*, Sergej Prokhorov[a]

[a]*Samara National Research University, 34 Moskovskoye sh., 443086 Samara, Russia*

Abstract

The problem of complete separation between classes may produce serious difficulties with the successful implementation of logistic regression due to the presence of floor and ceiling effects. To address this problem, the present study proposes two modifications of ordinary log-likelihood. To reveal the benefits of these modifications, we provided a strong theoretical and experimental basis for comparison with the mostly reported way of penalizing of log-likelihood – the regularization method. From these comparisons, we concluded that the proposed modifications produced less biased estimates and reached higher accuracy on prediction compared to the regularized log-likelihood under more unstable conditions: on samples with fewer observations and more predictors.

Keywords: logistic regression; log-likelihood; regularization; floor effect; ceiling effect

1. Introduction

Logistic regression (LR) has proved to show remarkable performance among machine learning methods. This success is often attributed to having lower bias [1], making fewer assumptions in comparison with other linear classifiers [2-5], and deeper understanding the role of predictors due to the high level of simplicity and interpretability [5-8].

* Corresponding author. Tel.: +7-846-267-4672.
E-mail address: kulikovskikh.i@gmail.com

1877-7058 © 2017 The Authors. Published by Elsevier Ltd.
Peer-review under responsibility of the scientific committee of the 3rd International Conference "Information Technology and Nanotechnology".
10.1016/j.proeng.2017.09.627

Nevertheless, the successful implementation of LR seems to crucially depend on the accurate identification of complete separation between classes [5-11]. Although the problem of separation primarily arises in small datasets with several unbalanced, highly predictive features [10] and results in a log-likelihood's (LL) failure to converge [5-17], it may also occur with small or medium-sized datasets when at least one LR estimate is infinite even if the likelihood converges. This means that classes can be perfectly separated by a single feature or by a non-trivial linear combination of features. Finally, the problem of clear separation may arise if the underlying model parameters are low in an absolute value. Consequently, creating a proper measure to handle perfect separability is of high importance.

A comprehensive review of the literature on this problem suggested a good deal of solutions such as partial least squares [9, 13], iteratively reweighted least squares [14, 15]. Another approach to deal with the separable data is to apply prior distributions to the likelihood function as suggested in [7, 12, 14]. In particular, Jeffreys prior distribution (Firth) that is developed to reduce the bias of maximum likelihood estimates in generalized linear models has been shown to provide an ideal solution to separation [7, 10, 14, 15]. However, in spite of reliable computational results, these estimates are not clearly interpretable as prior information in a regression context. The problem of proper interpretability was highlighted in [11] where the authors considered complete separability in terms of floor or ceiling effects [7, 8]. The logit transformation permits to move proportions away from the ceiling or floor by adding half a success and half a failure. Even if empirical logit analysis helps to cope with convergence issues, it little addresses the real problem: the estimates of model parameters from logistic regression and empirical logit analysis rest on different assumptions.

The most reported solution to the problem of perfect separation consists in penalizing the maximum likelihood [6, 7, 9, 16, 17]. This method implies penalizing of LL to make the estimates finite. But, adopting regularization may lead to not asymptotically normal and highly biased estimates even if the regularized LL tends to produce lower prediction errors [6, 7]. Thus, the present research is an attempt to fill this gap by proposing two promising modifications of LL. These modifications should ensure less biased and more interpretable estimates under unstable conditions in which either floor or ceiling is present. To support the theoretical outcomes, this study provided a strong theoretical and experimental basis for comparison with the mostly reported way of penalizing of log-likelihood – the regularization method.

2. Problem statement

Let $\{x_i, y_i\}_{i=1}^{m}$ denote independent and identically distributed observations with binary responses $y_i \in \{0,1\}$. The matrix $X \in \mathbf{R}^{m\times n}$ can be viewed either as $X = \left[x_1, \quad , x_n\right]^T$, with vectors of predictors $x_i \in \mathbf{R}^n$, or as $X = \left[x^1, \quad , x^m\right]$, with vectors of features $x^j \in \mathbf{R}^m$. Let $y = \left[y_1, \quad , y_n\right]^T$ be the response vector. Then, for any vector of regression coefficients $\theta \in \mathbf{R}^n$ LR models the class conditional probabilities $p(x_i,\theta) = P(y_i = 1 \mid x_i,\theta)$ by

$$\ln\left(\frac{p(x_i,\theta)}{1-p(x_i,\theta)}\right) = \theta^T x_i .$$

Having defined the main parameters, let us now move on to pose the LR problem.

Problem 1. *Let* g *be the link (logit) function that defines the relationship between the class conditional expectation of the response variable and the underlying linear model* $g\left(E\left[y_i \mid x_i\right]\right) = \theta^T x_i$. *Taking into account that* $E\left[y_i \mid x_i\right] = p(x_i,\theta)$ *for LR,*

$$g\left(p(x_i,\theta)\right) = \ln\left(\frac{p(x_i,\theta)}{1-p(x_i,\theta)}\right). \tag{1}$$

Under the model $p(x_i,\theta)$, *the negative log-likelihood (LL) expressed as*

$$\ln L(\theta) = -\sum_i y_i \ln\left(p(x_i,\theta)\right) + (1-y_i)\ln\left(1-p(x_i,\theta)\right) \tag{2}$$

or, what is the same,

$$\ln L(\theta) = -\sum_i y_i \theta^T x_i - \ln\left(1+\exp\left(\theta^T x_i\right)\right) \tag{3}$$

needs to be minimized to solve the following problem

$$\theta^* = \arg\min_{\theta \in \mathbf{R}^n} \ln L(\theta). \tag{4}$$

As a result of complete separation or no separation between the classes, the logit function (1) may go to $-\infty$ for 0 successes (floor effect) or ∞ for 0 failures (ceiling effect) [6, 11], respectively. But, this means that (4) fails to converge. The traditional approach to deal with floor and ceiling effects is to penalize LL (4) for very large estimates and, thus, to shrink these estimates toward 0. In particular, a widely-used way [6, 7] to do this adds an extra term to (2)

$$\theta^* = \arg\min_{\theta \in \mathbf{R}^n} \left\{\ln L(\theta) + \lambda P(\theta)\right\}, \tag{5}$$

where λ is a regularization parameter, $P(\theta)$ is a function that penalizes coefficients θ as they get further away from zero.

Considering some shortcomings of (5) that were pointed out in Section 1, let us now propose two new modifications of LL to address the problem of floor and ceiling effects.

3. Modifications of log-likelihood

Before explaining these modifications, it is necessary to present the extension of *Problem 1*. Thus, *Problem 2* involves an extra parameter $c \in [0,1]$ to describe floor (S_1) and ceiling effects (S_2).

Problem 2. *Let the logit function* $g(p(x_i,\theta))$ *be extended to* $g\left(p(x_i,\theta,c)^S\right)$, *where* $c \in [0,1]$, $S = \{S_1, S_2\}$, as follows

$$g\left(p(x_i,\theta,c)^S\right) = \begin{cases} \ln\left(\dfrac{p(x_i,\theta,c)^{S_1} - c}{1 - p(x_i,\theta,c)^{S_1}}\right) & \text{if floor;} \\ \ln\left(\dfrac{p(x_i,\theta,c)^{S_2}}{1 - p(x_i,\theta,c)^{S_2} - c}\right) & \text{if ceiling.} \end{cases} \tag{6}$$

Then, minimizing the negative LL $\ln L(\theta,c)^S$ *solves the following two-parameter problem*

$$(\theta,c)^* = \arg\min_{\substack{\theta \in \mathbf{R}^n \\ c \in [0,1]}} \ln L(\theta,c)^S. \tag{7}$$

The present study considers two modifications of LL $\ln L(\theta,c)$ with regard to (6) and the way of passing c into (2).

3.1. 1-form modification

This form of modification implies introducing the parameter c in the LL loss function (2) as

$$\ln L(\theta,c)^{S^I} = \begin{cases} -\sum_i y_i \ln\left(p(x_i,\theta,c)^{S_1} - c\right) + (1-y_i)\ln\left(1 - p(x_i,\theta,c)^{S_1}\right) & \text{if floor;} \\ -\sum_i y_i \ln\left(p(x_i,\theta,c)^{S_2}\right) + (1-y_i)\ln\left(1 - p(x_i,\theta,c)^{S_2} - c\right) & \text{if ceiling.} \end{cases} \tag{8}$$

Based on the definition (8), let us state the following lemma.

Lemma 1. *For each* $x_i \in \mathbf{R}^n$ *,* $\theta \in \mathbf{R}^n$ *, and* $c \in [0,1]$ *, the LL loss function* $\ln L(\theta,c)^{S^I}$ *based on 1-form modification* (9) *for floor and ceiling effects are the same and equal to*

$$\ln L(\theta,c)^{S^I} = -\sum_i y_i \theta^T x_i - \ln\left(1+\exp\left(\theta^T x_i\right)\right) - \ln(1-c). \tag{9}$$

Proof. See Appendix.

Corollary 2. For $c = 0$, (9) results in (3).

3.2. 2-form modification

This modification, in contrast, employs the same definition for presenting floor and ceiling effects

$$\ln L(\theta,c)^{S^{II}} = -\sum_i y_i \ln\left(p(x_i,\theta,c)^S\right) + (1-y_i)\ln\left(1-p(x_i,\theta,c)^S\right). \tag{10}$$

The lemma that is posed below seems to clearly underline the difference between the proposed modifications of LL.

Lemma 3. *For each* $x_i \in \mathbf{R}^n$ *,* $\theta \in \mathbf{R}^n$*, and* $c \in [0,1]$*, the LL loss function* $\ln L(\theta,c)^{S^{II}}$ $\ln L(\theta,c)$ *based on 2-form modification* (10) *for floor and ceiling effects are equal to*

$$\ln L(\theta,c)^{S^{II}} = \begin{cases} -\sum_i y_i \ln\left(c+\exp\left(\theta^T x_i\right)\right) + (1-y_i)\ln(1-c) - \ln\left(1+\exp\left(\theta^T x_i\right)\right) & \textit{if } \text{floor;} \\ -\sum_i y_i \ln\left((1-c)\exp\left(\theta^T x_i\right)\right) + (1-y_i)\ln\left(1+c\exp\left(\theta^T x_i\right)\right) - \ln\left(1+\exp\left(\theta^T x_i\right)\right) & \textit{if } \text{ceiling.} \end{cases} \tag{11}$$

Proof. See Appendix.

Corollary 4. For $c = 0$, (11) gives (3).

The similarity between (3), (9), and (11) invites the following comparison: while *2-form* modification (11) implies both the inclusion of the estimates θ and the parameter c to penalize LL, *1-form* modification (9) includes only one extra term $-\ln(1-c)$ compared to the known definition (3). It should be noted, though, that the problem (7) produces the estimates θ based on (6), i.e. the parameter c modifies the coefficients θ and, thus, is implicitly presented in both modifications regardless of the form or the type of effect introduced (floor or ceiling).

The detailed analysis of the estimates' bias and the issues of convexity and convergence is beyond the scope of this paper, but still seems promising direction for further research. In the present study, to confirm the theoretical outcomes, we conducted a series of computational experiments the results of which are given in the following section.

4. Results

To highlight the benefits of the proposed solution to the problem of minimizing floor and ceiling effects, the present study is intended to compare the results proposed in this paper with previously reported in the literature. For this purpose, we considered ridge regression with the penalty $P(\theta) = \sum_j \theta_j^2$ in (5) to introduce the regularized LL. Before we go any further, let us first describe datasets chosen to support these comparisons.

4.1. Datasets

The datasets *Haberman's Survival* [18], *Liver Disorder* [19], *Pima Indians Diabetes* [20], *Breast Cancer Wisconsin* [21], *Heart Statlog* [22] were taken from UCI Machine Learning Repository. Table 1 presents the relevant information on these datasets.

Table 1. A brief description of datasets.

	m			n
dataset	$y \in \{0,1\}$	$y=0$	$y=1$	
Haberman's Survival (*haberman*)	306	225	81	3
Liver Disorder (*liver*)	345	145	200	6
Pima Indians Diabetes (*pima*)	768	500	268	8
Breast Cancer Wisconsin (*breast*)	683	444	239	10
Heart Statlog (*heart*)	270	150	120	13

The given samples of data $X^m = (x_i, y_i)_{i=1}^m$ were divided into the training subset $X^{l_1} = (x_i, y_i)_{i=1}^{l_1}$ and the validation subset $X^{l_2} = (x_i, y_i)_{i=1}^{l_2}$ using 3-fold cross validation. To increase a chance of identifying the effects of floor and ceiling, the experiments suggested varying the limited number of observations: $m = \{5k \mid k \in [5,10]\}$. As for small to moderate sample sizes the resampling estimates are better than the asymptotic estimates, the bootstrap method was adopted to provide reliable results.

4.2. Computational experiments

The computational experiments were designed to: 1) estimate the accuracy of classification with regard to a form of LL; 2) compare the modified LL with the ordinary and regularized LL based on the classification results. Table 2-6 demonstrate the results of classification subject to $m = \{5k \mid k \in [5,10]\}$ for each dataset. For clarity, we highlighted in bold those classification results which reveal the advantages the proposed modified LL over both the ordinary LL and the regularized LL.

If we look at these values, we can see that all the penalized LL (5), (9), (11) produced better results – an increase in accuracy is up to 5% – than the ordinary LL (3) except for *haberman* and *liver* datasets (see Table 2, 3). This fact may be explained by having fewer predictors (see Table 1) compared to other datasets as the worst results of classification are obtained for *haberman* dataset that includes only 3 predictors (see Fig. 2).

Table 2. The accuracy of classification subject to $m = \{5k \mid k \in [5,10]\}$ (*haberman* dataset).

A form of LL	$m=25$	$m=30$	$m=35$	$m=40$	$m=45$	$m=50$
LL	67.3154	68.3744	69.6062	69.9102	70.3023	72.289
Regularized LL	70.0349	71.9228	72.8543	72.4385	73.0967	74.7039
1-form modified LL (floor)	67.0908	67.3098	68.6445	69.4777	69.3328	71.2442
2-form modified LL (floor)	67.8912	68.0861	69.0074	69.6420	70.2876	72.0978
1-form modified LL (ceiling)	67.3154	68.3744	69.6062	69.9102	70.3023	72.2289
2-form modified LL (ceiling)	67.8663	68.0639	68.3361	69.0931	69.7177	71.6188

Applying the modified LL to *liver* dataset seems more attractive: both LL modifications performed better than the regularized LL, but, still, worse than the ordinary LL (see Table 3). The probable explanation for this phenomenon may lie in more biased estimates in case of the regularized LL in comparison with the modified LL due to the high level of variation in small datasets.

The results of prediction on *pima* and *breast* datasets are clearly indicative of the presence of floor and ceiling effects and, as a result, reveal the real benefits of the proposed solution (see Table 4, 5). As the values of accuracy are not high enough for *pima* dataset, the modifications that describe a floor effect allowed us to yield more marked improvement on the results than the modifications of LL for a ceiling effect. Moreover, *1-form* modified LL performed better than *2-form* modified LL (see Fig. 2).

Table 3. The accuracy of classification subject to $m = \{5k \mid k \in [5,10]\}$ (*liver* dataset).

A form of LL	$m = 25$	$m = 30$	$m = 35$	$m = 40$	$m = 45$	$m = 50$
LL	60.9032	62.1867	62.7291	63.2735	63.7154	64.8303
Regularized LL	53.6677	54.7572	53.9829	55.8716	54.2201	55.7951
1-form modified LL (floor)	57.0858	59.1484	59.2996	59.5309	59.6094	60.7585
2-form modified LL (floor)	57.4850	58.7048	60.0617	60.1464	60.4220	62.0492
1-form modified LL (ceiling)	57.9840	58.6605	59.3359	59.1983	60.1796	60.4258
2-form modified LL (ceiling)	56.9361	57.1524	59.1000	59.1317	59.9800	61.1178

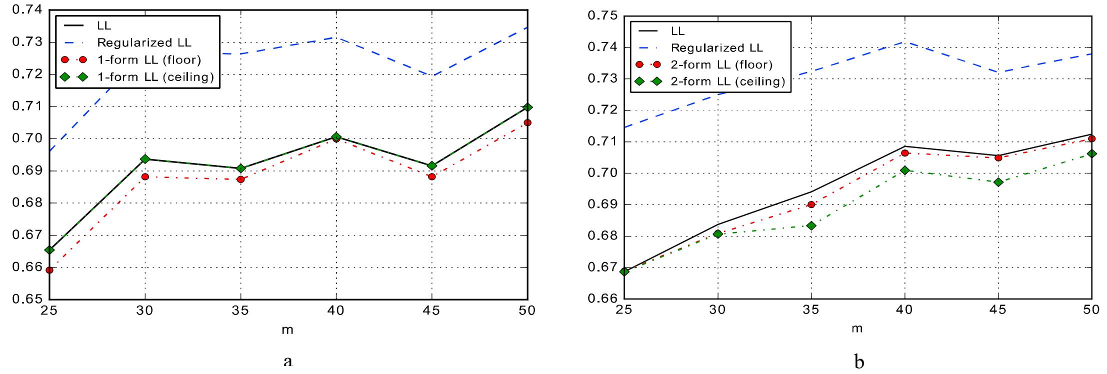


Fig.1. Comparison the modified LL with both the ordinary and regularized LL subject to $m = \{5k \mid k \in [5,10]\}$ (*haberman* dataset): a) 1-form; b) 2-form.

Table 4. The accuracy of classification subject to $m = \{5k \mid k \in [5,10]\}$ (*pima* dataset).

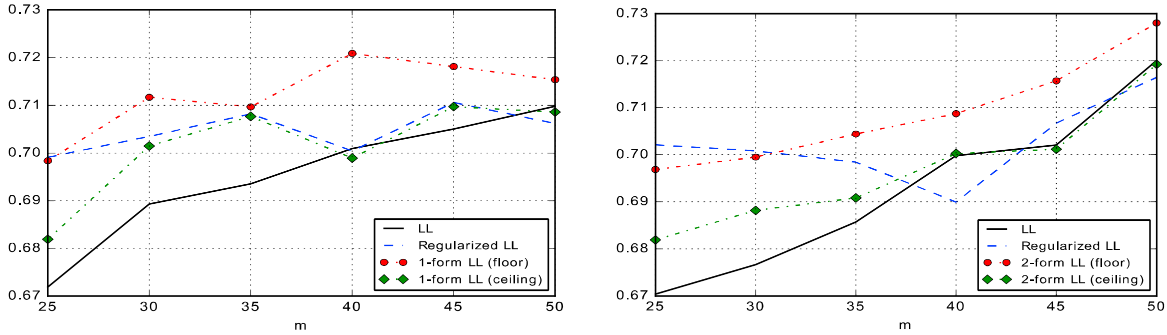
A form of LL	$m = 25$	$m = 30$	$m = 35$	$m = 40$	$m = 45$	$m = 50$
LL	67.1906	68.9288	69.3522	70.0931	70.5019	70.9780
Regularized LL	69.3102	70.3482	70.8220	70.0432	71.0579	70.6188
1-form modified LL (floor)	**69.8353**	**71.1688**	**70.9672**	**72.0892**	**71.8135**	**71.5369**
2-form modified LL (floor)	**69.6856**	70.0821	70.4409	**70.8749**	**71.5711**	**72.8011**
1-form modified LL (ceiling)	68.1886	70.1486	70.7676	69.8935	70.9723	**70.8583**
2-form modified LL (ceiling)	68.1886	69.9490	69.0811	70.0266	70.1169	**71.9228**

Table 5. The accuracy of classification subject to $m = \{5k \mid k \in [5,10]\}$ (*breast* dataset).

A form of LL	$m = 25$	$m = 30$	$m = 35$	$m = 40$	$m = 45$	$m = 50$
LL	94.5359	94.9918	95.1430	95.0931	95.2096	95.6354
Regularized LL	95.4840	95.9535	96.0523	96.1577	96.0793	96.2608
1-form modified LL (floor)	94.5110	95.2640	95.0543	95.5256	95.4234	95.5822
2-form modified LL (floor)	95.0213	**96.2874**	**96.5536**	95.7818	**96.1710**	**96.3540**
1-form modified LL (ceiling)	95.0848	95.8447	95.7862	**96.6234**	**96.3644**	**96.9261**
2-form modified LL (ceiling)	**96.2854**	**96.7798**	**97.1391**	**96.5669**	**96.8128**	**96.9661**

The results of prediction on *breast* dataset, in contrast, mostly point to the presence of a ceiling effect as well as they demonstrate the advantages of *2-form* modified LL. In addition, as *breast* dataset has more predictors compared to *pima* dataset, *2-form* modification that describes a floor effect also allowed us to improve the accuracy of classification.

Finally, Table 6 and Fig. 3 present the results of prediction on *heart* dataset. This dataset consists of much less observations and more predictors. As can be seen, almost all classification results are highlighted in bold. This means that the proposed modifications ensure less biased and more interpretable estimates under unstable conditions in which either floor or ceiling is present.



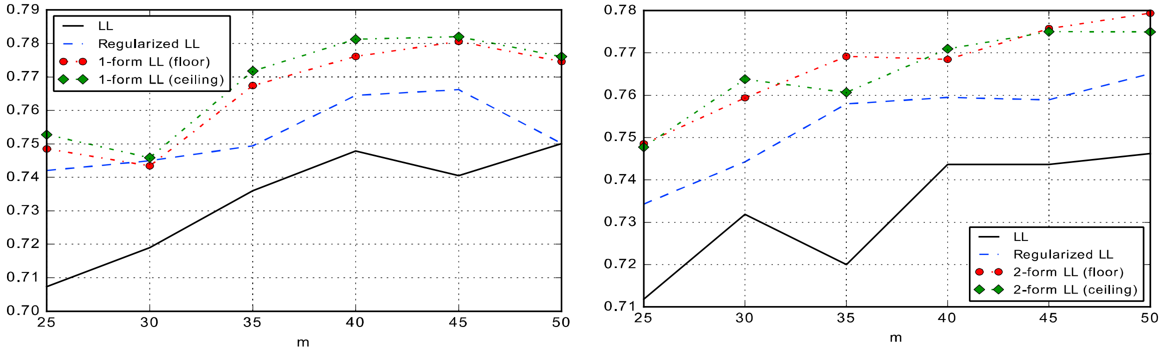
a

b

Fig.2. Comparison the modified LL with both the ordinary and regularized LL subject to $m=\{5k\,|\,k\in[5,10]\}$ (*pima* dataset): a) 1-form; b) 2-form.

Table 6. The accuracy of classification subject to $m=\{5k\,|\,k\in[5,10]\}$ (*heart* dataset).

A form of LL	$m=25$	$m=30$	$m=35$	$m=40$	$m=45$	$m=50$
LL	70.7335	71.9006	73.5983	71.7838	74.0519	75.0110
Regularized LL	74.2016	74.4955	74.9410	76.4471	76.6182	75.0110
1-form modified LL (floor)	**74.8503**	74.3402	**76.7374**	**77.6114**	**78.0582**	**77.4584**
2-form modified LL (floor)	**74.8503**	**74.9370**	**77.9189**	**76.8463**	**77.5734**	**76.5136**
1-form modified LL (ceiling)	**75.2745**	**74.5842**	**77.1729**	**78.1271**	**78.2007**	**77.6048**
2-form modified LL (ceiling)	**74.7754**	**75.3806**	**77.0660**	**77.0958**	**77.5021**	**77.4983**



a

b

Fig.3. Comparison the modified LL with both the ordinary and regularized LL subject to $m=\{5k\,|\,k\in[5,10]\}$ (*heart* dataset): a) 1-form; b) 2-form.

5. Conclusion

The present study was aimed at proposing a proper measure based on LL to directly address the issue of floor and ceiling effects in classification problems. For this reason, we offered two promising modifications of LL: *1-form* modification and *2-form* modification. In support of these modifications, we provided a theoretical and experimental basis for comparison with the known ways of penalizing of LL reported in the literature, in particular, the regularization method. From these comparisons we may draw the following conclusions: the proposed modifications produced less biased estimates and reached higher accuracy on prediction compared to the regularized LL under more unstable conditions: on samples with fewer observations and more predictors. Consequently, the proposed modifications help to minimize the effects of floor and ceiling and, thus, to improve the convergence of log-likelihood. The purpose, stated in this paper, is accomplished.

Appendix: Proofs

Proof of Lemma 1. According to the definition (8), 1-form modified LL in the presence of a floor effect (S_1) may be presented as

$$\ln L(\theta,c)^{S_1^I} = -\sum_i y_i \ln\left(p(x_i,\theta,c)^{S_1} - c\right) + (1-y_i)\ln\left(1-p(x_i,\theta,c)^{S_1}\right)$$

or, what is the same,

$$\ln L(\theta,c)^{S_1^I} = -\sum_i \ln\left(1-p(x_i,\theta,c)^{S_1}\right) + y_i \ln\left(\frac{p(x_i,\theta,c)^{S_1} - c}{1-p(x_i,\theta,c)^{S_1}}\right). \tag{A1}$$

The proposed extension to the logit function (6)

$$\ln\left(\frac{p(x_i,\theta,c)^{S_1} - c}{1-p(x_i,\theta,c)^{S_1}}\right) = \theta^T x_i$$

implies

$$1-p(x_i,\theta,c)^{S_1} = \frac{1-c}{1+\exp(\theta^T x_i)}. \tag{A2}$$

Using (A2) in (A1) gives (9).
By analogy, 2-form modified LL in the case of ceiling effect (S_2) (8) can be written as

$$\ln L(\theta,c)^{S_2^{II}} = -\sum_i y_i \ln\left(p(x_i,\theta,c)^{S_2}\right) + (1-y_i)\ln\left(1-p(x_i,\theta,c)^{S_2} - c\right),$$

$$\ln L(\theta,c)^{S_2^{II}} = -\sum_i \ln\left(1-p(x_i,\theta,c)^{S_2} - c\right) + y_i \ln\left(\frac{p(x_i,\theta,c)^{S_2}}{1-p(x_i,\theta,c)^{S_2} - c}\right). \tag{A3}$$

According to the definition (6),

$$1-p(x_i,\theta,c)^{S_2} - c = \frac{1-c}{1+\exp(\theta^T x_i)}. \tag{A4}$$

Using (A4) to simplify (A3) results in (9). **Lemma 1** is supported.

Proof of Lemma 3. Let us present the definition of 2-form modified LL (10) in the following form:

$$\ln L(\theta,c)^{S^{II}} = -\sum_i \ln\left(1-p(x_i,\theta,c)^{S}\right) + y_i \ln\left(\frac{p(x_i,\theta,c)^{S}}{1-p(x_i,\theta,c)^{S}}\right). \tag{A5}$$

With regard to (A2),

$$\frac{p\left(x_i,\theta,c\right)^{S_1}}{1-p\left(x_i,\theta,c\right)^{S_1}}=\frac{c+\exp(\theta^T x_i)}{1-c}\,. \tag{A6}$$

Then, applying (A2) and (A6) to (A5) results in

$$\ln L\left(\theta,c\right)^{S_1^{II}}=-\sum_i y_i \ln\left(c+\exp\left(\theta^T x_i\right)\right)+\left(1-y_i\right)\ln\left(1-c\right)-\ln\left(1+\exp\left(\theta^T x_i\right)\right).$$

To fully prove Lemma 3, let us push further analogy. In accordance with (A4),

$$1-p\left(x_i,\theta,c\right)^{S_2}=\frac{1+c\exp(\theta^T x_i)}{1+\exp(\theta^T x_i)}\,, \tag{A7}$$

$$\frac{p\left(x_i,\theta,c\right)^{S_2}}{1-p\left(x_i,\theta,c\right)^{S_2}}=\frac{\left(1-c\right)\exp(\theta^T x_i)}{1+c\exp(\theta^T x_i)}\,. \tag{A8}$$

Applying (A7) and (A8) to (A5) gives

$$\ln L\left(\theta,c\right)^{S_1^{II}}=-\sum_i y_i \ln\left(\left(1-c\right)\exp\left(\theta^T x_i\right)\right)+\left(1-y_i\right)\ln\left(1+c\exp\left(\theta^T x_i\right)\right)-\ln\left(1+\exp\left(\theta^T x_i\right)\right).$$

This proves **Lemma 3.**

Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation, grant 074-U01.

References

[1] N.A. Zaidi, G.I. Webb, M.J. Carman, F. Petitjean, J. Cerquides, ALR^n: accelerated higher-order logistic regression, Machine Learning. 104 (2016) 151–194.

[2] A.V. Kuznetsov, V.V. Myasnikov, A comparison of algorithms for supervised classification using hyperspectral data, Computer Optics. 38 (2014) 494–502.

[3] D.A. Zherdev, N.L. Kazanskiy, V.A. Fursov, Object recognition by the radar signatures of electromagnetic field scattering on base of support subspaces method, Computer Optics. 38 (2014) 503–510.

[4] N.Yu. Ilyasova, A.V. Kupriyanov, R.A. Paringer, Formation of features for improving the quality of medical doagnosis based on discriminant analysis methods, Computer Optics. 38 (2014) 851–855.

[5] I.M. Kulikovskikh, Anomaly detection in an ecological feature space to improve the accuracy of human activity identification in buildings, Computer Optics. 41 (2017) 126–133. DOI: 10.18287/2412-6179-2017-41-1-126-133.

[6] A. Agresti, Foundations of linear and generalized linear models, Wiley Series in Probability and Statistics, 2015.

[7] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: Data mining, inference, and prediction: 2nd ed. Springer Series in Statistics, 2013.

[8] C.E. McCulloch, S.R. Searle, J.M. Neuhaus, Generalized, linear, and mixed models: 2nd ed. John Wiley, New York, 2009.

[9] G. Fort, S. Lambert-Lacroix, Classification using partial least squares with penalized logistic regression, Bioinformatics. 21, 7 (2005) 1104–1111.

[10] G. Heinze, M. Schemper, A solution to the problem of separation in logistic regression, Statistics in Medicine. 21 (2002) 2409–2419.

[11] S. Donnelly, J. Verkuilen, Empirical logit analysis is not logistic regression, Journal of Memory and Language. 94 (2017) 28–42.

[12] A. Gelman, A. Jakulin, M.G. Pittau, Y.-S. Su, A weakly informative default prior distribution for logistic and other regression models, The Annals of Applied Statistics. 2, 4 (2008) 1360–1383.

[13] B. Ding, C.M. Gentleman, Classification using generalized partial least squares, Graphical Statistics. 14, 2 (2005) 280–298.

[14] D. Firth, Bias reduction, the Jeffreys prior and GLIM, Advances in GLIM and Statistical Modelling, Springer-Verlag, New York, 1992, pp. 91–100.

[15] D. Firth, Generalized linear models and Jeffreys priors: An iterative weighted least-squares approach, Computational Statistics, Physica-Verlag, Vienna, 1992, pp. 553–557.

[16] Y. Fan, J. Lv, Asymptotic equivalence of regularization methods in thresholded parameter space, Journal of the American Statistical Association. 108, 503 (2013) 1044–1061.

[17] M.Y. Park, T. Hastie, L_1-regularization path algorithm for generalized linear models, Journal of the Royal Statistical Society, Series B. 69, 4 (2007) 659–677.

[18] UCI Machine Learning Repository, Haberman's Survival, https://archive.ics.uci.edu/ml/datasets/Haberman%27s +Survival (30.05.2017).

[19] UCI Machine Learning Repository, Liver Disorders, https://archive.ics.uci.edu/ml/datasets/liver+disorders (30.05.2017).

[20] UCI Machine Learning Repository, Pima Indians Diabetes, https://archive.ics.uci.edu/ml/datasets/pima+indians+ diabetes (30.05.2017).

[21] UCI Machine Learning Repository, Breast Cancer Wisconsin, https://archive.ics.uci.edu/ml/datasets/breast+cancer+ wisconsin+(original) (30.05.2017).

[22] UCI Machine Learning Repository, Statlog (Heart), http://archive.ics.uci.edu/ml/datasets/Statlog+(Heart) (29.01.2017).