Pattern Recognition Letters 32 (2011) 101–106

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# A sparse version of the ridge logistic regression for large-scale text categorization

Sujeevan Aseervatham [a,*], Anestis Antoniadis [b], Eric Gaussier [a], Michel Burlet [c], Yves Denneulin [d]

[a] LIG – Université Joseph Fourier, 385, rue de la Bibliothèque, BP 53, F-38041 Grenoble Cedex 9, France
[b] LJK – Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France
[c] Lab. Leibniz-Université Joseph Fourier, 46 Avenue Félix Viallet, F-38031 Grenoble Cedex 1, France
[d] LIG – ENSIMAG, 51 avenue Jean Kuntzmann, F-38330 Montbonnot Saint Martin, France

ARTICLE INFO

*Article history:*
Received 16 January 2010
Available online 1 October 2010
Communicated by R.C. Guido

*Keywords:*
Logistic regression
Model selection
Text categorization
Large scale categorization

ABSTRACT

The ridge logistic regression has successfully been used in text categorization problems and it has been shown to reach the same performance as the Support Vector Machine but with the main advantage of computing a probability value rather than a score. However, the dense solution of the ridge makes its use unpractical for large scale categorization. On the other side, LASSO regularization is able to produce sparse solutions but its performance is dominated by the ridge when the number of features is larger than the number of observations and/or when the features are highly correlated. In this paper, we propose a new model selection method which tries to approach the ridge solution by a sparse solution. The method first computes the ridge solution and then performs feature selection. The experimental evaluations show that our method gives a solution which is a good trade-off between the ridge and LASSO solutions.



## 1. Introduction

The automatic text categorization problem consists in assigning, according to its content, a textual document to one or more relevant predefined categories. Given a training dataset, where the documents have been manually labeled, the problem lies in inducing a function $f$, from the training data, which can then be used to classify documents. Machine learning algorithms are used to find the optimal $f$ by solving a minimization problem which can be stated as the minimization of the cost of misclassification over the training dataset (Empirical Risk Minimization).

In order to use numerical machine learning algorithm, the Vector Space Model is commonly used to represent a textual documents by a simple term-frequency vector (Salton et al., 1975). This representation produces datasets in which (1) the number of features is often larger than the number of documents, (2) the vectors are very sparse, i.e., a lot of features are set to zero and (3) the features are highly correlated (due to the nature of natural languages). Moreover, real-life datasets tend to be larger and larger which makes the automatic categorization process complicated and leads to scalability problems. As long as the datasets only grow in terms of the number of observations, the problem can be tackled by distributing the computation over a network of processors (Chu et al., 2006). However, when the number of features becomes larger than the number of observations, machine learning techniques tend to perform poorly due to overfitting, i.e., the model performs well on the training set but poorly on any other set. To prevent overfitting, the complexity of the model must be controlled during the training process, through model selection techniques. In the Support Vector Machine (SVM) algorithm (Vapnik, 1995), the model complexity is given by the VC-dimension, which is the maximum number of vectors, for any combination of labels, that can be shattered by the model. SVMs rely on the Structural Risk Minimization (SRM) principle, which not only aims at minimizing the empirical risk (Empirical Risk Minimization – ERM) but also the VC-dimension. SVMs have been used for text categorization and their performance is among the best ones obtained so far (Joachims, 1998).

The VC-dimension remaining unknown for many functions, the SRM is difficult to implement. Another model selection, widely used, is to minimize both the ERM and a regularization term: $\lambda\Omega[f]$ where $\lambda$ is a penalty factor, $\Omega[f]$ a convex non-negative regularization term and $f$ the model. For linear functions: $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$, the regularization term is often defined as $\Omega[f] = \|\boldsymbol{w}\|_p$ where $\|\cdot\|_p$ is the $L_p$-norm (Hoerl and Kennard, 1970; Tibshirani, 1994; Zou and Hastie, 2005). This has the effect of smoothing $f$ and reducing its generalization error. The use of the $L_2$-norm is known as the ridge penalization, whereas the use of the $L_1$-norm as the LASSO penalization, which has the property of simultaneously doing shrinkage and feature selection.

In this paper, we focus on penalized logistic regression. Logistic regression has the main advantage of computing a probability value rather than a score, as for the SVM. Furthermore, the ridge

* Corresponding author. Tel.: +33 (0) 476514515; fax: +33 (0) 476446675.
*E-mail addresses:* Sujeevan.Aseervatham@imag.fr (S. Aseervatham), Anestis.Antoniadis@imag.fr (A. Antoniadis), Eric.Gaussier@imag.fr (E. Gaussier), Michel.Burlet@imag.fr (M. Burlet), Yves.Denneulin@imag.fr (Y. Denneulin).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.09.023

logistic regression has been shown to reach the same performance as the SVM on standard text categorization problems (Zhang and Oles, 2001). Nevertheless, it produces a dense solution which cannot be used for large scale categorization. In (Genkin et al., 2007), the LASSO logistic regression was used to obtain a sparse solution. However, when the number of features is larger than the number of observations and/or when the features are correlated, the ridge penalization performance dominates the LASSO one (Zou and Hastie, 2005). Taking into account these observations, we propose a new model selection which produces a sparse solution by approaching the ridge solution.

The rest of the paper is organized as follows: in the next section we discuss related works; we then describe, in Section 3, our model selection approach before reporting, in Section 4, our experimental results; Section 5 concludes the paper.

## 2. Related work

In (le Cessie and van Houwelingen, 1992), the authors have shown how ridge penalization can be used to improve the logistic regression parameter estimates in the cases where the number of features is larger than the number of observations or when the variables are highly correlated. They have applied ridge logistic regression on DNA data and have obtained good results with stable parameters. More recently, the ridge logistic regression was used in (Zhang and Oles, 2001) on the text categorization problem where the data are sparse and the number of features is larger than the number of observations. The authors have proposed several algorithms, which take advantage of the sparsity of the data, to solve efficiently the ridge optimization problem. The experimental results show that the $L_2$ logistic regression reaches the same performance as the SVM. Although the ridge method allows to select a more stable model by doing continuous shrinkage, the produced solution is dense and thus not appropriate for large and sparse data such as textual data.

The LASSO regularization ($L_1$-norm) has been introduced in (Tibshirani, 1994). The author shows, for linear regression, that the $L_1$ penalization can not only do continuous shrinkage but has also the property of doing automatic variable selection simultaneously which means that the $L_1$ solution is sparse. In (Genkin et al., 2007), an optimization algorithm based on Zhang and Oles (2001) is presented for Ridge and LASSO logistic regressions in the context of text categorization. According to their experiments, the LASSO penalization gives slightly better results than the ridge penalization in terms of the macro-averaged-$F_1$ measure (the micro-averaged results are not given). It has been shown in (Efron et al., 2004; Tibshirani, 1994; Zou and Hastie, 2005) that the performance of the LASSO is dominated by the ridge in the following cases (we denote by $p$ the number of features and by $n$ the number of observations):

- $p > n$: the LASSO will only select at most $n$ features,
- the features are highly correlated: the LASSO will select only one feature among the correlated features.

To tackle the limitations of the LASSO, the Elastic net method has been proposed in (Zou and Hastie, 2005) which tries to capture the best of both $L_1$ and $L_2$ penalizations. The Elastic net uses both $L_1$ and $L_2$ regularization in the linear regression problem. The authors show that the $L_2$ regularization term can be reformulated by adding $p$ artificial input data such that each artificial data $i$ has only the $i$th component non-null set to $\sqrt{\lambda_2}$ where $\lambda_2$ is the $L_2$ regularization hyperparameter. This reformulation, which leads to a LASSO problem, relies on the particular form of the least square term, and cannot be extended to the logistic regression problem. Furthermore, as the $L_1$ and $L_2$ regularizations are done simultaneously, it is unclear how the solution of the Elastic net approaches the $L_2$ solution. In (Zhao and Yu, 2006), the model consistency of LASSO is studied for linear regression and it is shown that the consistency of LASSO depends on the regularization parameter. In (Bach, 2008), the author proves that for a regularization parameter decay factor of $\frac{1}{\sqrt{n}}$, a consistent model can be obtained by applying LASSO on bootstrap samples and by selecting only the intersecting features. Nevertheless, using LASSO on bootstrap samples is a time consuming process. Moreover, since this method is based on LASSO, it also fails to induce a good model when the variables are correlated.

## 3. Selected Ridge Logistic Regression

The logistic regression model is part of the Generalized Linear Model (GLM) family (Hastie and Tibshirani, 1990; Mccullagh and Nelder, 1989). The GLM is a family of models, parametrized by $\boldsymbol{\beta}$, which associate a target variable $y$ to an input data $\boldsymbol{x}$ ($x \in \mathbb{R}^p$) according to the relation $\boldsymbol{\beta} \cdot \boldsymbol{x} = g(y)$ where $g$ is a link function and $\beta \in \mathbb{R}^p$. For simplicity, we represent any linear function $\boldsymbol{\beta}' \cdot \boldsymbol{x}' + \beta'_0$ by $\boldsymbol{\beta} \cdot \boldsymbol{x}$, where $\boldsymbol{x}$ is $\boldsymbol{x}'$ with an extra dimension set to 1, and $\boldsymbol{\beta}$ is $\boldsymbol{\beta}'$ with an extra dimension set to $\beta'_0$. The logistic regression model is obtained by using a logit function $g(y) = \frac{P(y|\boldsymbol{\beta},\boldsymbol{x})}{1-P(y|\boldsymbol{\beta},\boldsymbol{x})}$. When $y \in \{-1, 1\}$, the logistic regression model can be written as:

$$P(y = 1|\boldsymbol{\beta}, \boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{\beta} \cdot \boldsymbol{x})} \tag{1}$$

$\boldsymbol{\beta}$ can be obtained by maximizing the log-likelihood over the training set $\mathcal{D} = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n)\}$. However, in order to obtain a strictly convex optimization problem and to avoid overfitting, a Tikhonov regularization term (Hoerl and Kennard, 1970) is added, leading to the following ridge logistic regression problem:

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{\beta} \cdot \boldsymbol{x_i})) + \lambda \|\boldsymbol{\beta}\|_2^2}_{l(\boldsymbol{\beta})} \tag{2}$$

where $\lambda$ is a strictly positive scalar. Adding a ridge regularization term is equivalent, in a Bayesian framework, to using a Gaussian prior on each component of $\boldsymbol{\beta}$, under the assumption that the components are independent, i.e. $P(\boldsymbol{\beta}) = \prod_j P(\beta_j)$ with $P(\beta_j) \sim \mathcal{N}(0, \frac{1}{2\lambda})$.

Several algorithms have been proposed in the literature to solve the optimization problem in (2) (Friedman et al., 2008; Minka, 2003). In (Genkin et al., 2007), an efficient algorithm, based on the one presented in (Zhang and Oles, 2001), is proposed to solve problems with sparse data, such as text documents. However, the ridge regression solution is a dense vector which can hardly be used in large scale categorization where hundreds of thousand features are used. The problem we face is thus the one of finding $\hat{\boldsymbol{\beta}}$ such that:

1. $\hat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}^*$ and thus behaves well, i.e. $l(\hat{\boldsymbol{\beta}}) \simeq l(\boldsymbol{\beta}^*)$;
2. $\hat{\boldsymbol{\beta}}$ is a sparse solution and thus can be used on large datasets.

The second order Taylor series expansion on $l(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^*$ leads to:

$$\begin{aligned} l(\boldsymbol{\beta}) &\simeq l(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla l(\boldsymbol{\beta}^*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T H_l(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ &= l(\boldsymbol{\beta}^*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T H_l(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \end{aligned} \tag{3}$$

where $\nabla l(\boldsymbol{\beta}^*)$ and $H_l(\boldsymbol{\beta}^*)$ are respectively the gradient and the Hessian of $l(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^*$ and where the equality derives from the fact that for $\boldsymbol{\beta}^*$, the ridge solution, the gradient vanishes. Hence, obtaining a $\hat{\boldsymbol{\beta}}$ yielding a value for $l$ close to the one of $\boldsymbol{\beta}^*$ while being sparse