Pattern Recognition 48 (2015) 3739–3748

Contents lists available at ScienceDirect

Pattern Recognition

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

Model selection for linear classifiers using Bayesian error estimation

Heikki Huttunen [a,*], Jussi Tohka [b,c]

[a] Department of Signal Processing, Tampere University of Technology, Finland
[b] Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Spain
[c] Instituto de Investigación Sanitaria Gregorio Marañon, Madrid, Spain

ARTICLE INFO

Article history:
Received 20 November 2014
Received in revised form
6 March 2015
Accepted 2 May 2015
Available online 14 May 2015

Keywords:
Logistic regression
Support vector machine
Regularization
Bayesian error estimator
Linear classifier

ABSTRACT

Regularized linear models are important classification methods for high dimensional problems, where regularized linear classifiers are often preferred due to their ability to avoid overfitting. The degree of freedom of the model dis determined by a regularization parameter, which is typically selected using counting based approaches, such as *K*-fold cross-validation. For large data, this can be very time consuming, and, for small sample sizes, the accuracy of the model selection is limited by the large variance of CV error estimates. In this paper, we study the applicability of a recently proposed Bayesian error estimator for the selection of the best model along the regularization path. We also propose an extension of the estimator that allows model selection in multiclass cases and study its efficiency with L_1 regularized logistic regression and L_2 regularized linear support vector machine. The model selection by the new Bayesian error estimator is experimentally shown to improve the classification accuracy, especially in small sample-size situations, and is able to avoid the excess variability inherent to traditional cross-validation approaches. Moreover, the method has significantly smaller computational complexity than cross-validation.

1. Introduction

The task in supervised classification is to learn to make predictions about the class of an unknown object given a training set of *P*-dimensional feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ with known class memberships. An important special case of supervised classification problems arises when the number of features *P* is larger or nearly as large than the number of training samples *N*. These classification problems are increasingly important, for example, in genomics and neuroimaging [1,2]. Due to a small number of training samples (compared to the data dimensionality), linear classifiers are preferred in such cases. Also, some form of regularization is necessary to cope with small *N*.

In this paper, we concentrate on two widely used regularized linear classifiers: L_1 or LASSO regularized logistic regression [3–5] and support vector machine (SVM) [6,7]. These classifiers are trained by minimizing a cost function that is a weighted sum of data term and a regularization term. The usual strategy for selecting the weights (or the value for the regularization parameter) is to train classifiers for various values of regularization parameter producing a set of models and then select the best model according to some model selection criteria. The most widely used approach is to select the best model based on a (non-parametric) estimate of classification error, such as cross-validation (CV), bootstrap and resubstitution error estimators.

The randomness of the cross-validation has certain drawbacks: the model selection depends on the particular split of the data, the approach is time consuming, and the resulting error estimate may have a large variance [8]. In particular, the latter problem has been documented already almost four decades ago [9], but it is still often dismissed [8]. Thus, we are interested in finding a *deterministic*, *accurate* and *fast* approach for choosing the regularization parameter of a regularized linear classification model.

Other approaches for model selection include information theoretic tools, such as the Akaike Information Criterion (AIC) [10], the Bayesian Information Criterion (BIC) [11]; including its use for logistic regression models [12] and SVMs [13], and extended BIC (EBIC) [12], which corrects drawbacks of the BIC when $P > N$. However, all of the above are based on the likelihood of the model, not on the prediction error. In predictive modeling, the actual model is often of secondary importance, and the critical issue is the prediction ability and the minimal error. For the SVM model selection, various techniques based on different error bounds and concepts from algorithmic information theory have been suggested [14,15]. However, these are either complicated and expensive to compute or do not yield satisfactory results [16]. Probably for this reason, the *K*-fold CV is still the most popular model selection criterion also in small sample settings.

* Corresponding author.
E-mail address: heikki.huttunen@tut.fi (H. Huttunen).

http://dx.doi.org/10.1016/j.patcog.2015.05.005
0031-3203/

Recently, a few alternatives to the CV type methods for the estimation of classification error have been proposed. One of them is the *Bolstered error estimation* [17], which attempts to smooth the empirical distribution of the available data by placing bolstering kernels at each data point location. A more recent approach, the *Bayesian minimum mean-square estimator for classification error* describes the error in a Bayesian framework [18,19]. Moreover, a closed form expression can be derived for the posterior expectation of the classification error in the binary classification case under mild assumptions about the covariance structure. The method is attractive, because the errors are estimated directly from the training data, and no iterative resampling or splitting operations are required. This results also in a significant speedup, since the classifier training is done only once. For example, the 5-fold CV (CV-5) includes five training iterations on partial data and one on all training data, while the Bayesian error estimator requires only the last training step with all data.

Experimental data suggests that the Bayesian error estimator (BEE) can be more accurate in absolute terms than the CV-based classification error estimates, in particular with small sample sizes [19]. In our earlier work, we have shown that the BEE is accurate for model selection as well [20]. More specifically, we compared the BEE with CV and BIC criteria when used for selecting the regularization parameter λ for the binary logistic regression model with LASSO penalty. This paper extends the earlier study by (1) proposing a Bayesian model selection rule for multinomial classification problems, (2) considering BEE model selection under more general priors than in [20] and (3) studying the rule for selection of the regularization parameter for both SVM and logistic regression classifiers. Moreover, extensive experiments show that the BEE criterion is significantly faster than the CV, and also more accurate unless the model assumptions are severely violated. The implementation of the proposed error estimator in Matlab and Python is available for download.[1]

The rest of this paper is organized as follows: In Section 2 we will briefly review the regularized logistic regression and SVM classifiers; Section 3 defines the Bayesian error estimator for binary and multiclass cases; and Section 4 compares the accuracy of the BEE to CV and BIC based model selection in various experimental cases. Finally, Section 5 discusses the applicability and the limits of the proposed method.

2. Linear classifiers

In the following, we denote the observation matrix as $\mathbf{X} \in \mathbb{R}^{N\times P}$, whose rows $\mathbf{x}_i$ are the samples with corresponding class labels $\mathbf{y} = (y_1, \ldots, y_N)^T$ with $y_i = \{-1, 1\}$ in the 2-class case and $y_i \in \{1, 2, \ldots, C\}$ in the multiclass case. The predicted class label $\hat{y}$ for the feature vector $\mathbf{x}$ is given by $\hat{y} = \mathrm{sign}(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}) \doteq \mathrm{g}(\mathbf{x})$ in the binary case and $\hat{y} = \arg\max_c(\beta_{c,0} + \boldsymbol{\beta}_c^T\mathbf{x})$ in the multiclass case, where the classifier parameters $\beta_0, \beta_{c,0} \in \mathbb{R}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^T$, $\boldsymbol{\beta}_c = (\beta_{c,1}, \beta_{c,2}, \ldots, \beta_{c,P})^T \in \mathbb{R}^P$ are learned from training data.

2.1. Regularized logistic regression

Logistic regression (LR) is a statistical classification method modeling the class conditional probability densities by the logistic function. Binary logistic regression models the class probabilities of the sample $\mathbf{x} = (x_1, x_2, \ldots, x_P)^T \in \mathbb{R}^P$ belonging to class $c \in \{-1, 1\}$ as

$$\Pr(c|\mathbf{x}) = \frac{1}{1+\exp\left[c(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})\right]}.$$

[1] https://sites.google.com/site/bayesianerrorestimate/

A slightly different model is adopted for multinomial logistic regression for C classes [5], which models the probability $\Pr(c|\mathbf{x})$ of the sample $\mathbf{x}$ belonging to class $c \in \{1, 2, \ldots, C\}$ as

$$\Pr(c|\mathbf{x}) = \frac{\exp(\beta_{c,0} + \boldsymbol{\beta}_c^T\mathbf{x})}{\sum_{k=1}^{C}\exp(\beta_{k,0} + \boldsymbol{\beta}_k^T\mathbf{x})}.$$

Adopting the notation $\boldsymbol{\beta}_1 \doteq \boldsymbol{\beta}$ for the 2-class case, the model parameters are learned from the training data by maximizing the ℓ_1-penalized log-likelihood

$$\sum_{i=1}^{N}\log\ \Pr(y_i|\mathbf{x}_i) - \lambda\sum_{c=1}^{L}\|\boldsymbol{\beta}_c\|_1$$

where $L = C$ for the multiclass case and $L = 1$ for the 2-class case.

Although the penalized log-likelihood function is not differentiable everywhere, several approximate algorithms exist for the minimization task [4,5,21] and the implementation of this paper uses the GLMNET algorithm [5].

2.2. The support vector machine

Support vector machines (SVM) are widely used due to their maximum margin property. The binary SVM with the linear kernel solves the following problem:

$$\min_{\boldsymbol{\beta},\beta_0,\xi}\left(\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + C^*\sum_{i=1}^{l}\xi_i\right)\text{such that}\begin{cases} y_i(\boldsymbol{\beta}^T\mathbf{x}_i + \beta_0) \geq 1 - \xi_i, \\ \xi_i \geq 0 \quad \text{for } i = 1, \ldots, N. \end{cases}$$

where $C^* \in \mathbb{R}$ is the upper bound. Alternatively, the above constrained minimization problem can written in a form emphasizing the regularization [7]

$$\min_{\beta_0,\boldsymbol{\beta}}\sum_{i=1}^{N}[1 - y_i(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i)]_+ + \lambda\|\boldsymbol{\beta}\|_2^2,$$

where $[x]_+ = \max(0, x)$ and $\lambda = 1/(2C^*)$. In our work, we use the LIBSVM implementation of the SVM and extend it into the multiclass case using a one-against-one strategy [22].

2.3. Model selection

With both logistic regression and SVM classifiers, the parameter $\lambda > 0$ controls the strength of the regularization. Different choices of $\lambda \in \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ produce different classifiers and we denote the parameters as $\beta_0(\lambda)$, $\boldsymbol{\beta}(\lambda)$, $\beta_{c,0}(\lambda)$ and $\boldsymbol{\beta}_c(\lambda)$ to make this fact apparent when needed.

The best value of λ is traditionally selected by cross-validation: either as the minimum of the cross-validation error curve [22] or as the largest λ whose error is within one standard deviation from the minimum [5]. The latter rule favors slightly sparser solutions and tends to decrease the generalization error. However, in our earlier work [20] choosing the minimum CV error solution resulted in more accurate prediction, so from now on we will focus on the minimum of the CV error as the selection rule.

Fig. 1 illustrates the model selection using different error estimators. Examples of error curves for different values of the regularization parameter λ with the logistic regression model are shown in Fig. 1. In this example, a 20-dimensional toy dataset with altogether 500 normally distributed samples drawn from two classes was generated. The errors for models with $\log_{10}(\lambda) \in \{-0.5, -0.6, \ldots, -3.9, -4.0\}$ were estimated using 5-fold CV (top left) and the BEE (top right). There is a significant variation between resulting error curves as shown in Fig. 1 (top left) and there is even more significant variation between the location of the minima of the curves, as seen in Fig. 1 (bottom). In particular, note that there are two isolated cases where the minima are far from the majority of cases, with $\lambda = 10^{-2.8}$ and $\lambda = 10^{-2.9}$.