International Journal of Forecasting 28 (2012) 722–738

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

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

Fast sparse regression and classification

Jerome H. Friedman

Department of Statistics, Stanford University, Stanford, CA 94305, United States

ARTICLE INFO

Keywords:
Regression
Classification
Regularization
Sparsity
Variable selection
Bridge-regression
Lasso
Elastic net
l_p-norm penalization

ABSTRACT

Many present day applications of statistical learning involve large numbers of predictor variables. Often, that number is much larger than the number of cases or observations available for training the learning algorithm. In such situations, traditional methods fail. Recently, new techniques have been developed, based on regularization, which can often produce accurate models in these settings. This paper describes the basic principles underlying the method of regularization, then focuses on those methods which exploit the sparsity of the predicting model. The potential merits of these methods are then explored by example.

1. Introduction

Linear structural models are among the most popular for fitting data. One is given N observations of the form

$$\{y_i, \boldsymbol{x}_i\}_1^N = \{y_i, x_{i1}, \ldots, x_{in}\}_1^N, \tag{1}$$

which is considered to be a random sample from some joint (population) distribution with probability density $p(\boldsymbol{x}, y)$. The random variable y is the "outcome" or "response" and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ are the predictor variables. These predictors may be the original measured variables and/or selected functions constructed from them. The goal is to estimate the joint values for the parameters $\boldsymbol{a} = \{a_0, a_1, \ldots, a_n\}$ of the linear model

$$F(\boldsymbol{x}; \boldsymbol{a}) = a_0 + \sum_{j=1}^{n} a_j x_j \tag{2}$$

for predicting y given $\boldsymbol{x}$, that minimize the expected loss ("risk")

$$R(\boldsymbol{a}) = E_{\boldsymbol{x},y} L(y, F(\boldsymbol{x}; \boldsymbol{a})) \tag{3}$$

over future predictions $\boldsymbol{x}, y \backsim p(\boldsymbol{x}, y)$. Here, $L(y, F)$ is a loss criterion that specifies the cost of predicting the value F when the actual value is y. Popular loss criteria include the squared error

$$L(y, F) = (y - F)^2, \tag{4}$$

and the Bernoulli negative log-likelihood

$$L(y, F) = \log(1 + e^{-yF}), \quad y \in \{-1, 1\}, \tag{5}$$

associated with logistic regression. The negative log-likelihood representing any probability model can be characterized by a corresponding loss criterion.

For a specified loss criterion, the optimal parameter values are from Eq. (3)

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}} R(\boldsymbol{a}). \tag{6}$$

Since the population probability density $p(\boldsymbol{x}, y)$ is unknown, a common practice is to substitute an empirical estimate of the expected value in Eq. (3) based on the available data (Eq. (1)), yielding

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \hat{R}(\boldsymbol{a}) \tag{7}$$

as an estimate for $\boldsymbol{a}^*$, where

$$\hat{R}(\boldsymbol{a}) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, a_0 + \sum_{j=1}^{n} a_j x_{ij}\right). \tag{8}$$

E-mail address: jhf@stanford.edu.

0169-2070/$ – see front matter © 2012 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2012.05.001

2. Regularization

It is well known that $\hat{\boldsymbol{a}}$ in Eqs. (7) and (8) often provides a poor estimate of $\boldsymbol{a}^*$; that is, $R(\hat{\boldsymbol{a}}) \gg R(\boldsymbol{a}^*)$ (Eq. (3)). This is especially the case when the sample size N is not large compared to the number of parameters $(n + 1)$. This is caused by the high variability of the estimates in Eq. (7) when Eq. (8) is evaluated on different random samples drawn from the population distribution. A common remedy is to modify Eq. (7) in order to stabilize the estimates by placing a restriction on the joint solution values. That is,

$$\hat{\boldsymbol{a}}(t) = \arg\min_{\boldsymbol{a}} \hat{R}(\boldsymbol{a}) \quad \text{s.t.}\, P(\boldsymbol{a}) \le t. \tag{9}$$

Here, $P(\boldsymbol{a})$ is a non-negative function of the parameters specifying the form of the constraint, and $t \ge 0$ regulates its strength. For a given data set (Eq. (1)), the loss criterion $L(y, F)$ in Eqs. (3) and (8), and the constraint function $P(\boldsymbol{a})$, the solution to Eq. (9) depends only on the value chosen for t. Varying its value induces a family of solutions, with each member being indexed by a particular value of $t \in [0, P(\hat{\boldsymbol{a}})]$ (Eq. (7)). This same family of solutions can be obtained through the equivalent (penalized) formulation of Eq. (9):

$$\hat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a}} [\hat{R}(\boldsymbol{a}) + \lambda \cdot P(\boldsymbol{a})], \tag{10}$$

where $P(\boldsymbol{a})$ is the constraining function in Eq. (9), here called a penalty, and $\lambda > 0$ regulates its strength. Setting $\lambda = \infty$ produces the totally constrained solution $(t = 0)$, whereas $\lambda = 0$ yields the unrestricted solution $(t \ge P(\hat{\boldsymbol{a}}))$. Each value of $0 \le \lambda \le \infty$ in Eq. (10) produces one of the solutions $0 \le t \le P(\hat{\boldsymbol{a}})$ in Eq. (9), with smaller values of λ corresponding to larger values of t. Thus, Eq. (10) produces a family of estimates in which each member of the family is indexed by a particular value for the strength parameter λ. This family lies on a one-dimensional path of finite length in the $(n + 1)$-dimensional space of all joint parameter values.

2.1. Model selection

The optimal parameter values $\boldsymbol{a}^*$ (Eq. (6)) also represent a point in the parameter space. For a given penalty, the goal is to find a point λ^* on its path such that the corresponding solution $\hat{\boldsymbol{a}}(\lambda^*)$ is closest to $\boldsymbol{a}^*$, where the distance is characterized by the prediction risk in Eq. (3)

$$D(\boldsymbol{a}, \boldsymbol{a}^*) = R(\boldsymbol{a}) - R(\boldsymbol{a}^*). \tag{11}$$

This is a classic model selection problem where one attempts to obtain an estimate $\hat{\lambda}$ of the optimal value of the strength parameter

$$\lambda^* = \arg\min_{0 \le \lambda \le \infty} R(\hat{\boldsymbol{a}}(\lambda)) \tag{12}$$

through

$$\hat{\lambda} = \arg\min_{0 \le \lambda \le \infty} \tilde{R}(\hat{\boldsymbol{a}}(\lambda)), \tag{13}$$

where $\tilde{R}(\boldsymbol{a})$ is a surrogate model selection criterion computed from the training data in Eq. (1), whose minimum is intended to approximate that of the actual risk (Eq. (3)).

There are a wide variety of model selection criteria available, each developed for a particular combination of loss (Eq. (3)) and penalty $P(\boldsymbol{a})$. Among the most general, being applicable to any loss-penalty combination, is cross-validation. The data are partitioned randomly into two subsets (learning and test). The path is constructed using only the learning sample. The test sample is then used as an empirical surrogate for the population density $p(\boldsymbol{x}, y)$ to compute the corresponding (estimated) risk in Eq. (3). These estimates are then used in Eq. (13) to obtain the estimate $\hat{\lambda}$. Sometimes the risk used in Eq. (13) is estimated by averaging over several (K) such partitions ("K-fold" cross-validation).

2.2. Penalty selection

Given a model selection procedure, the goal is to construct a path $\hat{\boldsymbol{a}}(\lambda)$ in the parameter space such that some of the points on that path are close to the point $\boldsymbol{a}^*$ (Eq. (6)) representing the optimal solution. If no points on the path are close to $\boldsymbol{a}^*$, as measured by Eq. (11), then no model selection procedure can produce accurate estimates $\hat{\boldsymbol{a}}(\hat{\lambda})$. Since the path produced by Eq. (10) depends on the data, different randomly drawn data sets (Eq. (1)) will produce different paths for the same penalty. Thus, the paths are themselves random, and one seeks a penalty $P(\boldsymbol{a})$ that produces paths $\hat{\boldsymbol{a}}(\lambda)$ such that

$$\left[E_T R(\hat{\boldsymbol{a}}(\lambda^*)) - R(\boldsymbol{a}^*)\right] / R(\boldsymbol{a}^*) = \text{ small}, \tag{14}$$

with T being repeated data samples (Eq. (1)) drawn randomly from the joint density $p(\boldsymbol{x}, y)$, and λ^* is given by Eq. (12). This will depend on the particular $\boldsymbol{a}^*$ (Eq. (6)) associated with the application. Therefore, penalty choice is governed by whatever is known about the properties of $\boldsymbol{a}^*$.

2.3. Sparsity

One property of $\boldsymbol{a}^*$ that is often suspected is sparsity. That is, only a small fraction of the input variables $\{x_j\}_1^n$ actually influence predictions, with the identities of those influential variables being unknown. The degree of sparsity $S(\boldsymbol{a})$ of a parameter vector $\boldsymbol{a}$ can be defined as

$$S(\boldsymbol{a}) = \frac{1}{n} \sum_{k=1}^{n} I\left(|a_k| \le \eta \cdot \max_j |a_j|\right), \tag{15}$$

with $\eta \ll 1$. If the predictor variables are all standardized to have similar scales, then $S(\boldsymbol{a}^*)$ represents the fraction of non-influential variables characterizing the problem.

If $\hat{\boldsymbol{a}}(\lambda^*) \simeq \boldsymbol{a}^*$ (Eq. (14)) then $S(\hat{\boldsymbol{a}}(\lambda^*)) \simeq S(\boldsymbol{a}^*)$, and in the absence of other information it is reasonable to choose a penalty that produces solutions $\hat{\boldsymbol{a}}(\lambda)$ with a sparsity similar to that of $\boldsymbol{a}^*$ at $\lambda = \lambda^*$. Since the actual sparsity of $\boldsymbol{a}^*$ is generally unknown, one can define a family of penalties $P_\gamma(\boldsymbol{a})$, where γ indexes particular penalties in the family that produce solutions of differing sparseness, and then use model selection (Section 2.1) to jointly estimate good values for γ and λ. That is,

$$\hat{\boldsymbol{a}}_\gamma(\lambda) = \arg\min_{\boldsymbol{a}} [\hat{R}(\boldsymbol{a}) + \lambda \cdot P_\gamma(\boldsymbol{a})] \tag{16}$$

$$(\hat{\gamma}, \hat{\lambda}) = \arg\min_{\gamma, \lambda} \tilde{R}(\hat{\boldsymbol{a}}_\gamma(\lambda)). \tag{17}$$