Adaptive Sparseness Using Jeffreys Prior

Mário A. T. Figueiredo
Institute of Telecommunications,
and Department of Electrical and Computer Engineering.
Instituto Superior Técnico
1049-001 Lisboa, Portugal
mtf@lx.it.pt

Abstract

In this paper we introduce a new sparseness inducing prior which does not involve any (hyper-)parameters that need to be adjusted or estimated. Although other applications are possible, we focus here on supervised learning problems: regression and classification. Experiments with several publicly available benchmark data sets show that the proposed approach yields state-of-the-art performance. In particular, our method outperforms support vector machines and performs competitively with the best alternative techniques, both in terms of error rates and sparseness, although it involves no tuning or adjusting of sparseness-controlling hyper-parameters.

1 Introduction

The goal of supervised learning is to infer a functional relation $y = f(x)$, based on a set of (maybe noisy) training examples $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Usually, the inputs are vectors, $x_i = [x_{i,1}, \ldots, x_{i,d}]^T \in \mathbb{R}^d$. When $y$ is continuous (typically $y \in \mathbb{R}$), we are in the context of regression, whereas in classification $y$ is of categorical nature (e.g., $y \in \{-1, 1\}$). Usually, the structure of $f(\cdot)$ is assumed fixed and the objective is to estimate a vector of parameters $\beta$ defining it; accordingly we write $y = f(x, \beta)$.

To achieve good generalization (i.e., to perform well on yet unseen data) it is necessary to control the complexity of the learned function (see [1] - [4], and the many references therein). In Bayesian approaches, complexity is controlled by placing a prior on the function to be learned, i.e., on $\beta$. This should not be confused with a generative (informative) Bayesian approach, since it involves no explicit modelling of the joint probability $p(x, y)$.

A common choice is a zero-mean Gaussian prior, which appears under different names, like ridge regression [5], or weight decay, in the neural learning literature [6]. Gaussian priors are also used in non-parametric contexts, like the Gaussian processes (GP) approach [2], [7], [8], [9], which has roots in earlier spline models [10] and regularized radial basis functions [11]. Very good performance has been reported for methods based on Gaussian priors [8], [9]. Their main disadvantage is that they do not control the structural complexity of the resulting functions. That is, if one of the components of $\beta$ (say, a weight in a neural network) happens to be irrelevant, a Gaussian prior will not set it exactly to zero, thus...

---

This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Ministry of Science and Technology, under project POSI/33143/SRI/2000.
pruning that parameter, but to some small value.

Sparse estimates (i.e., in which irrelevant parameters are set exactly to zero) are desirable because (in addition to other learning-theoretic reasons [4]) they correspond to a structural simplification of the estimated function. Using Laplacian priors (equivalently, $l_1$-penalized regularization) is known to promote sparseness [12] - [15]. Support vector machines (SVM) take a non-Bayesian approach to the goal of sparseness [2], [4]. Interestingly, however, it can be shown that the SVM and $l_1$-penalized regression are closely related [13].

Both in approaches based on Laplacian priors and in SVMs, there are hyper-parameters which control the degree of sparseness of the obtained estimates. These are commonly adjusted using cross-validation methods which do not optimally utilize the available data, and are time consuming. We propose an alternative approach which involves no hyper-parameters. The key steps of our proposal are: (i) a hierarchical Bayes interpretation of the Laplacian prior as a normal/independent distribution (as used in robust regression [16]); (ii) a Jeffreys’ non-informative second-level hyper-prior (in the same spirit as [17]) which expresses scale-invariance and, more importantly, is parameter-free [18]; (iii) a simple expectation-maximization (EM) algorithm which yields a maximum a posteriori (MAP) estimate of $\beta$ (and of the observation noise variance, in the case of regression).

Our method is related to the automatic relevance determination (ARD) concept [7], [19], which underlies the recently proposed relevance vector machine (RVM) [20], [21]. The RVM exhibits state-of-the-art performance, beating SVMs both in terms of accuracy and sparseness [20], [21]. However, we do not resort to a type-II maximum likelihood approximation [18] (as in ARD and RVM); rather, our modelling assumptions lead to a marginal a posteriori probability function on $\beta$ whose mode is located by a very simple EM algorithm. Like the RVM, but unlike the SVM, our classifier produces probabilistic outputs.

Experimental evaluation of the proposed method, both with synthetic and real data, shows that it performs competitively with (often better than) GP-based methods, RVM, and SVM.

2 Regression

We consider functions of the type $f(x, \beta) = \beta^T h(x)$, i.e., that are linear with respect to $\beta$ (whose dimensionality we will denote by $k$). This includes: (i) classical linear regression, $h(x) = [1, x_1, ..., x_d]^T$; (ii) nonlinear regression via a set of $k$ basis functions, $h(x) = [\phi_1(x), ..., \phi_k(x)]^T$; (iii) kernel regression, $h(x) = [1, K(x, x_1), ..., K(x, x_n)]^T$, where $K(x, y)$ is some (symmetric) kernel function [2] (as in SVM and RVM regression), not necessarily verifying Mercer’s condition.

We follow the standard assumption that $y_i = f(x_i, \beta) + \epsilon_i$, for $i = 1, ..., n$, where $[\epsilon_1, ..., \epsilon_n]$ is a set of independent zero-mean Gaussian variables with variance $\sigma^2$. With $y \equiv [y_1, ..., y_n]^T$, the likelihood function is then $p(y | \beta) = N(y | \beta^T H \beta, \sigma^2 I)$, where $H$ is the $(n \times k)$ design matrix which depends on the $x_i$s and on the adopted function representation, and $N(v | \mu, \Sigma)$ denotes a Gaussian density of mean $\mu$ and covariance $\Sigma$, evaluated at $v$.

With a zero-mean Gaussian prior with covariance $A$, $p(\beta | A) = N(\beta | 0, A)$, the posterior $p(\beta | y)$ is still Gaussian with mean and mode at

$$\hat{\beta} = (\sigma^2 A^{-1} + H^T H)^{-1} H^T y.$$

When $A$ is proportional to identity, say $A = \mu^2 I$, this is called ridge regression [5].

With a Laplacian prior for $\beta$, $p(\beta | \alpha) = \prod_i p(\beta_i | \alpha)$, with $p(\beta_i | \alpha) = \frac{1}{2} \exp\{-\alpha | \beta_i |\}$, the posterior $p(\beta | y)$ is not Gaussian. The maximum a posteriori (MAP) estimate is given by

$$\hat{\beta} = \arg\min\{ ||H \beta - y||_2^2 + 2 \sigma^2 \alpha ||\beta||_1 \},$$

- (1)
where \( \|v\|_2 \) is the Euclidean \((\ell_2)\) norm, and \( \|v\|_1 = \sum |v_i| \) is the \( \ell_1 \) norm. In linear regression this is called the LASSO (least absolute shrinkage and selection operator) [14].

The main effect of the \( \ell_1 \) penalty is that some of the components of \( \beta \) may be exactly zero. If \( H \) is an orthogonal matrix, \((1)\) can be solved separately for each \( \beta_i \), leading to the soft threshold estimation rule, widely used in wavelet-based signal/image denoising [22].

Let us consider an alternative model: let each \( \beta \) have a zero-mean Gaussian prior \( p(\beta|\tau_i) = N(\beta|0, \tau_i) \), with its own variance \( \tau_i \) (like in ARD and RVM). Now, rather than adopting a type-II maximum likelihood criterion (as in ARD and RVM), let us consider hyper-priors for the \( \tau_i \)s and integrate them out. Assuming exponential hyper-priors for the \( \tau_i \)s and integrate them out. Assuming exponential hyper-priors for the \( \tau_i \)s and integrate them out.

\[
p(\beta|\gamma) = (\gamma/2) \exp\{-\gamma \tau_i/2\} \quad \text{(for } \tau_i \geq 0, \text{ because these are variances) we obtain}
\]

\[
p(\beta|\gamma) = \int_0^{\infty} p(\beta|\tau_i)p(\tau_i|\gamma) d\tau_i = \frac{\sqrt{\pi}}{2} \exp\{-\gamma |\beta_i|\}.
\]

This shows that the Laplacian prior is equivalent to a 2-level hierarchical-Bayes model: zero-mean Gaussian priors with independent exponentially distributed variances. This decomposition has been exploited in robust least absolute deviation (LAD) regression [16].

The hierarchical decomposition of the Laplacian prior allows using the EM algorithm to implement the LASSO criterion in \((1)\) by simply regarding \( \tau \equiv \tau_1, \ldots, \tau_h \) as hidden/missing data. In fact, the complete log-posterior (with a flat prior for \( \sigma^2 \), and where \( \mathbf{Y}(\tau) \equiv \text{diag}(\tau_1^{-1}, \ldots, \tau_h^{-1}) \),

\[
\log p(\beta, \sigma^2 | \mathbf{y}, \tau) \propto -n \log \sigma^2 - \frac{\|y - H \beta\|_2^2}{\sigma^2} - \beta^T \mathbf{Y}(\tau) \beta,
\]

is easy to maximize with respect to \( \beta \) and \( \sigma^2 \). The E-step reduces to the computation of the conditional expectation of \( \mathbf{Y}(\tau) \), given current (at iteration \( t \)) estimates \( \hat{\tau}^{(t)} \) and \( \hat{\beta}_{(t)} \). This leads to \( \mathbf{V}_{(t)} = \mathbb{E}[\mathbf{Y}(\tau)|\mathbf{y}, \hat{\tau}_{(t)}] = \gamma \text{diag}(|\hat{\beta}_1|^{-1}, \ldots, |\hat{\beta}_h|^{-1}) \).

The M-step is then defined by the two following update equations:

\[
\hat{\sigma}_{(t+1)}^2 = \frac{1}{n} \|\mathbf{y} - H \hat{\beta}_{(t)}\|_2^2
\]

and

\[
\hat{\beta}_{(t+1)} = \left( \hat{\sigma}_{(t+1)}^2 \mathbf{V}_{(t)} + H^T H \right)^{-1} H^T \mathbf{y}.
\]

This EM algorithm is not the most efficient way to solve \((1)\); see, e.g., the methods proposed in [23], [14]. Our main goal is to open the way to the adoption of different hyper-priors.

One question remains: how to adjust \( \gamma \), which controls the degree of sparseness of the estimates? Our proposal is to remove \( \gamma \) from the model, by replacing the exponential hyper-prior by a non-informative Jeffreys hyper-prior: \( p(\tau_i) \propto \tau_i^{-1} \). This prior expresses ignorance with respect to scale (see [17], [18]) and, most importantly, it is parameter-free. Of course this is no longer equivalent to a Laplacian prior on \( \beta \), but to some other prior. As will be shown experimentally, this prior strongly induces sparseness and yields state-of-the-art performance. Computationally, this choice leads to a minor modification of the EM algorithm described above: matrix \( \mathbf{V}_{(t)} \) is now given by \( \mathbf{V}_{(t)} = \text{diag}(|\hat{\beta}_1|^{-2}, \ldots, |\hat{\beta}_h|^{-2}) \).

Since several of the \( \hat{\beta}_i \)s may go to zero, it is not convenient to deal with \( \mathbf{V}_{(t)} \). However, we can re-write the M-step as

\[
\hat{\beta}_{(t+1)} = \mathbf{U}_{(t)} \hat{\sigma}_{(t+1)}^2 \mathbf{I} + \mathbf{U}_{(t)} H^T H \mathbf{U}_{(t)}^{-1} \mathbf{U}_{(t)}^{-1} H^T \mathbf{y},
\]

where \( \mathbf{U}_{(t)} \equiv \text{diag}(|\hat{\beta}_1|^{-2}, \ldots, |\hat{\beta}_h|^{-2}) \), thus avoiding the inversion of the elements of \( \hat{\beta}_{(t)} \). Moreover, it is not necessary to invert the matrix, but simply to solve the corresponding linear system, whose dimension is only the number of non-zero elements in \( \mathbf{U}_{(t)} \).