Journal of Machine Learning Research 8 (2007) 841-861 Submitted 11/06; Published 4/07

Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters

Gavin C. Cawley GCC@CMP.UEA.AC.UK
Nicola L. C. Talbot NLCT@CMP.UEA.AC.UK
School of Computing Sciences
University of East Anglia
Norwich, United Kingdom NR4 7TJ

Editors: Isabelle Guyon and Amir Saffari

Abstract

While the model parameters of a kernel machine are typically given by the solution of a convex optimisation problem, with a single global optimum, the selection of good values for the regularisation and kernel parameters is much less straightforward. Fortunately the leave-one-out cross-validation procedure can be performed or a least approximated very efficiently in closed form for a wide variety of kernel learning methods, providing a convenient means for model selection. Leave-one-out cross-validation based estimates of performance, however, generally exhibit a relatively high variance and are therefore prone to over-fitting. In this paper, we investigate the novel use of Bayesian regularisation at the second level of inference, adding a regularisation term to the model selection criterion corresponding to a prior over the hyper-parameter values, where the additional regularisation parameters are integrated out analytically. Results obtained on a suite of thirteen real-world and synthetic benchmark data sets clearly demonstrate the benefit of this approach.

Keywords: model selection, kernel methods, Bayesian regularisation

1. Introduction

Leave-one-out cross-validation (Lachenbruch and Mickey, 1968; Luntz and Brailovsky, 1969; Stone, 1974) provides the basis for computationally efficient model selection strategies for a variety of kernel learning methods, including the Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Chapelle et al., 2002), Gaussian Process (GP) (Rasmussen and Williams, 2006; Sundararajan and Keerthi, 2001), Least-Squares Support Vector Machine (LS-SVM) (Suykens and Vandewalle, 1999; Cawley and Talbot, 2004), Kernel Fisher Discriminant (KFD) analysis (Mika et al., 1999; Cawley and Talbot, 2003; Saadi et al., 2004; Bo et al., 2006) and Kernel Logistic Regression (KLR) (Keerthi et al., 2005; Cawley and Talbot, 2007). These methods have proved highly successful for kernel machines having only a small number of hyper-parameters to optimise, as demonstrated by the set of models achieving the best average score in the WCCI-2006 performance prediction challenge[1] (Cawley, 2006; Guyon et al., 2006). Unfortunately, while leave-one-out cross-validation estimators have been shown to be almost unbiased (Luntz and Brailovsky, 1969), they are known to exhibit a relatively high variance (e.g., Kohavi, 1995). A kernel with many hyper-parameters, for instance those used in Automatic Relevance Determination (ARD) (e.g., Rasmussen and Williams, 2006) or feature scaling methods (Chapelle et al., 2002; Bo et al., 2006), may provide sufficient

1. See `http://www.modelselect.inf.ethz.ch/index.php`.

degrees of freedom to over-fit leave-one-out cross-validation based model selection criteria, resulting in performance inferior to that obtained using a less flexible kernel function. In this paper, we investigate the novel use of regularisation (Tikhonov and Arsenin, 1977) of the hyper-parameters in model selection in order to ameliorate the effects of the high variance of leave-one-out cross-validation based selection criteria, and so improve predictive performance. The regularisation term corresponds to a zero-mean Gaussian prior over the values of the kernel parameters, representing a preference for smooth kernel functions, and hence a relatively simple classifier. The regularisation parameters introduced in this step are integrated out analytically in the style of Buntine and Weigend (1991), to provide a Bayesian model selection criterion that can be optimised in a straightforward manner via, for example, scaled conjugate gradient descent (Williams, 1991).

The paper is structured as follows: The remainder of this section provides a brief overview of the least-squares support vector machine, including the use of leave-one-out cross-validation based model selection procedures, given in sufficient detail to ensure the reproducibility of the results. Section 2 describes the use of Bayesian regularisation to prevent over-fitting at the second level of inference, that is, model selection. Section 3 presents results obtained over a suite of thirteen benchmark data sets, which demonstrate the utility of this approach. Section 4 provides discussion of the results and suggests directions for further research. Finally, the work is summarised and directions for further work are outlined in Section 5.

1.1 Least Squares Support Vector Machine

In the remainder of this section, we provide a brief overview of the least-squares support vector machine (Suykens and Vandewalle, 1999) used as the testbed for the investigation of the role of regularisation in the model selection process described in this study. Given training data,

$$\mathcal{D} = \{(x_i,\, y_i)\}_{i=1}^{\ell}, \quad \text{where} \quad x_i \in \mathcal{X} \subset \mathbb{R}^d \quad \text{and} \quad y_i \in \{-1, +1\},$$

we seek to construct a linear discriminant, $f(x) = \phi(x) \cdot w + b$, in a *feature* space, $\mathcal{F}$, defined by a fixed transformation of the input space, $\phi : \mathcal{X} \to \mathcal{F}$. The parameters of the linear discriminant, $(w,\ b)$, are given by the minimiser of a *regularised* (Tikhonov and Arsenin, 1977) least-squares training criterion,

$$L = \frac{1}{2}\|w\|^2 + \frac{1}{2\mu}\sum_{i=1}^{\ell}[y_i - \phi(x_i)\cdot w - b]^2, \tag{1}$$

where μ is a regularisation parameter controlling the bias-variance trade-off (Geman et al., 1992). Rather than specify the feature space directly, it is instead induced by a kernel function, $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which evaluates the inner-product between the projections of the data onto the feature space, $\mathcal{F}$, that is, $\mathcal{K}(x, x') = \phi(x) \cdot \phi(x')$. The interpretation of an inner-product in a fixed feature space is valid for any Mercer kernel (Mercer, 1909), for which the Gram matrix, $K = [k_{ij} = \mathcal{K}(x_i, x_j)]_{i,j=1}^{\ell}$ is positive semi-definite, that is,

$$a^T K a \geq 0, \qquad \forall\, a \in \mathbb{R}^{\ell}, \qquad a \neq 0.$$

The Gram matrix effectively encodes the spatial relationships between the projections of the data in the feature space, $\mathcal{F}$. A linear model can thus be implicitly constructed in the feature space using only information contained in the Gram matrix, without explicitly evaluating the positions of the data in the feature space via the transformation $\phi(\cdot)$. Indeed, the *representer* theorem (Kimeldorf

and Wahba, 1971) shows that the solution of the optimisation problem (1) can be written as an expansion over the training patterns,

$$w = \sum_{i=1}^{\ell} \alpha_i \phi(x_i) \qquad \Longrightarrow \qquad f(x) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(x_i, x) + b.$$

The advantage of the "kernel trick" then becomes apparent; a linear model can be constructed in an extremely rich, high- (possibly infinite-) dimensional feature space, using only finite-dimensional quantities, such as the Gram matrix, K. The "kernel trick" also allows the construction of statistical models that operate directly on structured data, for instance strings, trees and graphs, leading to the current interest in kernel learning methods in computational biology (Schölkopf et al., 2004) and text-processing (Joachims, 2002). The Radial Basis Function (RBF) kernel,

$$\mathcal{K}(x, x') = \exp\left\{-\eta \|x - x'\|^2\right\}$$

is commonly encountered in practical applications of kernel learning methods, here η is a *kernel parameter*, controlling the sensitivity of the kernel function. The feature space for the radial basis function kernel consists of the positive orthant of an infinite-dimensional unit hyper-sphere (e.g., Shawe-Taylor and Cristianini, 2004). The Gram matrix for the radial basis function kernel is thus of full rank (Micchelli, 1986), and so the kernel model is able to form an arbitrary shattering of the data.

1.1.1 A DUAL TRAINING ALGORITHM

The basic training algorithm for the least-squares support vector machine (Suykens and Vandewalle, 1999) views the regularised loss function (1) as a constrained minimisation problem:

$$\min_{w,b,\varepsilon_i} \ \frac{1}{2}\|w\|^2 + \frac{1}{2\mu}\sum_{i=1}^{\ell} \varepsilon_i^2 \qquad \text{subject to} \qquad \varepsilon_i = y_i - w \cdot \phi(x_i) - b.$$

The primal Lagrangian for this constrained optimisation problem gives the unconstrained minimisation problem defined by the following regularised loss function,

$$\mathcal{L} = \frac{1}{2}\|w\|^2 + \frac{1}{2\mu}\sum_{i=1}^{\ell} \varepsilon_i^2 - \sum_{i=1}^{\ell} \alpha_i \left\{w \cdot \phi(x_i) + b + \varepsilon_i - y_i\right\},$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_\ell) \in \mathbb{R}^\ell$ is a vector of Lagrange multipliers. The optimality conditions for this problem can be expressed as follows:

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \Longrightarrow w = \sum_{i=1}^{\ell} \alpha_i \phi(x_i), \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{\ell} \alpha_i = 0, \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \varepsilon_i} = 0 \Longrightarrow \alpha_i = \frac{\varepsilon_i}{\mu}, \quad \forall_i \in \{1, 2, \ldots, \ell\}, \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \Longrightarrow w \cdot \phi(x_i) + b + \varepsilon_i - y_i = 0, \quad \forall\, i \in \{1, 2, \ldots, \ell\}. \tag{5}$$

Using (2) and (4) to eliminate w and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_\ell)$, from (5), we find that

$$\sum_{j=1}^{\ell} \alpha_j \phi(x_j) \cdot \phi(x_i) + b + \mu\alpha_i = y_i \qquad \forall\, i \in \{1, 2, \ldots, \ell\}. \tag{6}$$

Noting that $\mathcal{K}(x, x') = \phi(x) \cdot \phi(x')$, the system of linear equations, (6) and (3), can be written more concisely in matrix form as

$$\begin{bmatrix} K + \mu I & 1 \\ 1^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix},$$

where $K = [k_{ij} = \mathcal{K}(x_i, x_j)]_{i,j=1}^{\ell}$, I is the $\ell \times \ell$ identity matrix and 1 is a column vector of ℓ ones. The optimal parameters for the model of the conditional mean can then be obtained with a computational complexity of $O(\ell^3)$ operations, using direct methods, such as Cholesky decomposition (Golub and Van Loan, 1996).

1.1.2 Efficient Implementation Via Cholesky Decomposition

A more efficient training algorithm can be obtained, taking advantage of the special structure of the system of linear equations (Suykens et al., 2002). The system of linear equations to be solved in fitting a least-squares support vector machine is given by,

$$\begin{bmatrix} M & 1 \\ 1^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}, \tag{7}$$

where $M = K + \mu I$. Unfortunately the matrix on the left-hand side is not positive definite, and so we cannot solve this system of linear equations directly using the Cholesky decomposition. However, the first row of (7) can be re-written as

$$M\left(\alpha + M^{-1} 1 b\right) = y. \tag{8}$$

Rearranging (8), we see that $\alpha = M^{-1}(y - 1b)$, using this result to eliminate α, the second row of (7) can be written as

$$1^T M^{-1} 1 b = 1^T M^{-1} y.$$

The system of linear equations can then be re-written as

$$\begin{bmatrix} M & 0 \\ 0^T & 1^T M^{-1} 1 \end{bmatrix} \begin{bmatrix} \alpha + M^{-1} 1 b \\ b \end{bmatrix} = \begin{bmatrix} y \\ 1^T M^{-1} y \end{bmatrix}. \tag{9}$$

In this case, the matrix on the left hand side is positive-definite, as $M = K + \lambda I$ is positive-definite and $1^T M^{-1} 1$ is positive since the inverse of a positive definite matrix is also positive definite. The revised system of linear equations (9) can be solved as follows: First solve

$$M\rho = 1 \qquad \text{and} \qquad M\nu = y, \tag{10}$$

which may be performed efficiently using the Cholesky factorisation of M. The model parameters of the least-squares support vector machine are then given by

$$b = \frac{1^T \nu}{1^T \rho} \qquad \text{and} \qquad \alpha = \nu - \rho b.$$

The two systems of linear equations (10) can be solved efficiently using the Cholesky decomposition of $M = R^T R$, where R is the upper triangular Cholesky factor of M.

1.2 Leave-One-Out Cross-Validation

Cross-validation (Stone, 1974) is commonly used to obtain a reliable estimate of the test error for performance estimation or for use as a model selection criterion. The most common form, *k*-fold cross-validation, partitions the available data into *k* disjoint subsets. In each iteration a classifier is trained on a different combination of $k-1$ subsets and the unused subset is used to estimate the test error rate. The *k*-fold cross-validation estimate of the test error rate is then simply the average of the test error rate observed in each of the *k* iterations, or folds. The most extreme form of cross-validation, where $k = \ell$ such that the test partition in each fold consists of only a single pattern, is known as leave-one-out cross-validation (Lachenbruch and Mickey, 1968) and has been shown to provide an almost unbiased estimate of the test error rate (Luntz and Brailovsky, 1969). Leave-one-out cross-validation is however computationally expensive, in the case of the least-squares support vector machine a naïve implementation having a complexity of $O(\ell^4)$ operations. Leave-one-out cross-validation is therefore normally only used in circumstances where the available data are extremely scarce such that the computational expense is no longer prohibitive. In this case the inherently high variance of the leave-one-out estimator (Kohavi, 1995) is offset by the minimal decrease in the size of the training set in each fold, and so may provide a more reliable estimate of generalisation performance than conventional *k*-fold cross-validation. Fortunately leave-one-out cross-validation of least-squares support vector machines can be performed in closed form with a computational complexity of only $O(\ell^3)$ operations (Cawley and Talbot, 2004). Leave-one-out cross-validation can then be used in medium to large scale applications, where there may be a few thousand data-points, although the relatively high variance of this estimator remains potentially problematic.

1.2.1 VIRTUAL LEAVE-ONE-OUT CROSS-VALIDATION

The optimal values of the parameters of a Least-Squares Support Vector Machine are given by the solution of a system of linear equations:

$$\begin{bmatrix} K+\mu I & 1 \\ 1^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}. \tag{11}$$

The matrix on the left-hand side of (11) can be decomposed into block-matrix representation, as follows:

$$\begin{bmatrix} K+\mu I & 1 \\ 1^T & 0 \end{bmatrix} = \begin{bmatrix} c_{11} & c_1^T \\ c_1 & C_1 \end{bmatrix} = C.$$

Let $[\alpha^{(-i)}; b^{(-i)}]$ represent the parameters of the least-squares support vector machine during the i^{th} iteration of the leave-one-out cross-validation procedure, then in the first iteration, in which the first training pattern is excluded,

$$\begin{bmatrix} \alpha^{(-1)} \\ b^{(-1)} \end{bmatrix} = C_1^{-1} [y_2, \ldots, y_\ell, 0]^T .$$

The leave-one-out prediction for the first training pattern is then given by

$$\hat{y}_1^{(-1)} = c_1^T \begin{bmatrix} \alpha^{(-1)} \\ b^{(-1)} \end{bmatrix} = c_1^T C_1^{-1} [y_2, \ldots, y_\ell, 0]^T .$$

Considering the last ℓ equations in the system of linear equations (11), it is clear that $[c_1\, C_1]$ $[\alpha_1,\ldots,\alpha_\ell,b]^T = [y_2,\ldots,y_\ell,0]^T$, and so

$$\hat{y}_1^{(-1)} = c_1^T C_1^{-1}\,[c_1\, C_1]\,\left[\alpha^T, b\right]^T = c_1^T C_1^{-1} c_1 \alpha_1 + c_1\,[\alpha_2,\ldots,\alpha_\ell,b]^T\,.$$

Noting, from the first equation in the system of linear equations (11), that $y_1 = c_{11}\alpha_1 + c_1^T\,[\alpha_2,\ldots,\alpha_\ell,b]^T$, thus

$$\hat{y}_1^{(-1)} = y_1 - \alpha_1\left(c_{11} - c_1^T C_1^{-1} c_1\right).$$

Finally, via the block matrix inversion lemma,

$$\left[\begin{array}{cc} c_{11} & c_1^T \\ c_1 & C_1 \end{array}\right]^{-1} = \left[\begin{array}{cc} \kappa^{-1} & -\kappa^{-1} c_1 C_1^{-1} \\ C_1^{-1} + \kappa^{-1} C_1^{-1} c_1^T c_1 C_1^{-1} & -\kappa^{-1} C_1^{-1} c_1^T \end{array}\right],$$

where $\kappa = c_{11} - c_1^T C_1^{-1} c$, and noting that the system of linear equations (11) is insensitive to permutations of the ordering of the equations and of the unknowns, we have that,

$$y_i - \hat{y}_i^{(-i)} = \frac{\alpha_i}{C_{ii}^{-1}}. \tag{12}$$

This means that, assuming the system of linear equations (11) is solved via explicit inversion of C, a leave-one-out cross-validation estimate of an appropriate model selection criterion can be evaluated using information already available a by-product of training the least-squares support vector machine on the entire data set (cf., Sundararajan and Keerthi, 2001).

1.2.2 EFFICIENT IMPLEMENTATION VIA CHOLESKY FACTORISATION

The leave-one-out cross-validation behaviour of the least-squares support vector machine is described by (12). The coefficients of the kernel expansion, α, can be found efficiently, via Cholesky factorisation, as described in Section 1.1.2. However we must also determine the diagonal elements of C^{-1} in an efficient manner. Using the block matrix inversion formula, we obtain

$$C^{-1} = \left[\begin{array}{cc} M & 1 \\ 1^T & 0 \end{array}\right]^{-1} = \left[\begin{array}{cc} M^{-1} + M^{-1} 1 S_M^{-1} 1^T M^{-1} & -M^{-1} 1 S_M^{-1} \\ -S_M^{-1} 1^T M^{-1} & S_M^{-1} \end{array}\right]$$

where $M = K + \mu I$ and $S_M = -1^T M^{-1} 1 = -1^T \eta$ is the Schur complement of M. The inverse of the positive definite matrix, M, can be computed efficiently from its Cholesky factorisation, via the `SYMINV` algorithm (Seaks, 1972), for example using the LAPACK routine `DTRTRI` (Anderson et al., 1999). Let $R = [r_{ij}]_{i,j=1}^{\ell}$ be the lower triangular Cholesky factor of the positive definite matrix M, such that $M = RR^T$. Furthermore, let

$$S = [s_{ij}]_{i,j=1}^{\ell} = R^{-1}, \qquad \text{where} \qquad s_{ii} = \frac{1}{r_{ii}} \qquad \text{and} \qquad s_{ij} = -s_{ii}\sum_{k=1}^{i-1} r_{ik} s_{kj},$$

represent the (lower triangular) inverse of the Cholesky factor. The inverse of M is then given by $M^{-1} = S^T S$. In the case of efficient leave-one-out cross-validation of least-squares support vector machines, we are principally concerned only with the diagonal elements of M^{-1}, given by

$$M_{ii}^{-1} = \sum_{j=1}^{i} s_{ij}^2 \qquad \Longrightarrow \qquad C_{ii}^{-1} = \sum_{j=1}^{i} s_{ij}^2 + \frac{\rho_i^2}{S_M} \qquad \forall\, i \in \{1,2,\ldots,\ell\}.$$

The computational complexity of the basic training algorithm is $O(\ell^3)$ operations, being dominated by the evaluation of the Cholesky factor. However, the computational complexity of the analytic leave-one-out cross-validation procedure, when performed as a by-product of the training algorithm, is only $O(\ell)$ operations. The computational expense of the leave-one-out cross-validation procedure therefore becomes increasingly negligible as the training set becomes larger.

1.3 Model Selection

The virtual leave-one-out cross-validation procedure described in the previous section provides the basis for a simple automated model selection strategy for the least-squares support vector machine. Perhaps the most basic model selection criterion is provided by the Predicted REsidual Sum of Squares (PRESS) criterion (Allen, 1974), which is simply the leave-one-out estimate of the sum-of-squares error,

$$Q(\theta) = \frac{1}{2}\sum_{i=1}^{\ell}\left[y_i - \hat{y}_i^{(-i)}\right]^2.$$

A minimum of the model selection criterion is often found via a simple grid-search procedure in the majority of practical applications of kernel learning methods. However, this is rarely necessary and often highly inefficient as a grid-search spends a large amount of time investigating hyper-parameter values outside the neighbourhood of the global optimum. A more efficient approach uses the Nelder-Mead simplex algorithm (Nelder and Mead, 1965), as implemented by the `fminsearch` function of the MATLAB optimisation toolbox. An alternative easily implemented approach uses conjugate gradient methods, with the required gradient information estimated by the method of finite differences, and implemented by the `fminunc` function from the MATLAB optimisation toolbox. In this study however, we use scaled conjugate gradient descent (Williams, 1991), with the required gradient information evaluated analytically, as this is approximately twice as efficient.

1.3.1 PARTIAL DERIVATIVES OF THE PRESS MODEL SELECTION CRITERION

Let $\theta = \{\theta_1, \ldots, \theta_n\} = \{\lambda, \eta_1, \ldots, \eta_d\}$ represent the vector of hyper-parameters for a least-squares support vector machine, where $\eta_1, \ldots, \eta_d$ represent the kernel parameters. The PRESS statistic (Allen, 1974) can be written as

$$Q(\theta) = \frac{1}{2}\sum_{i=1}^{\ell}\left[r_i^{(-i)}\right]^2, \qquad \text{where} \qquad r_i^{(-i)} = y_i - \hat{y}_i^{(-i)} = \frac{\alpha_i}{C_{ii}^{-1}}.$$

Using the chain rule, the partial derivative of the PRESS statistic, with respect to an individual hyper-parameter, θ_j, is given by,

$$\frac{\partial Q(\theta)}{\partial \theta_j} = \sum_{i=1}^{\ell}\frac{\partial Q(\theta)}{\partial r_i^{(-i)}}\frac{\partial r_i^{(-i)}}{\partial \theta_j},$$

where

$$\frac{\partial Q(\theta)}{\partial r_i^{(-i)}} = r_i^{(-i)} = \frac{\alpha_i}{C_{ii}^{-1}} \qquad \text{and} \qquad \frac{\partial r_i^{(-i)}}{\partial \theta_j} = \frac{\partial \alpha_i}{\partial \theta_j}\frac{1}{C_{ii}^{-1}} - \frac{\alpha_i}{\left[C_{ii}^{-1}\right]^2}\frac{\partial C_{ii}^{-1}}{\partial \theta_j},$$

such that

$$\frac{\partial Q(\theta)}{\partial \theta_j} = \sum_{i=1}^{\ell}\frac{\alpha_i}{C_{ii}^{-1}}\left\{\frac{\partial \alpha_i}{\partial \theta_j}\frac{1}{C_{ii}^{-1}} - \frac{\alpha_i}{\left[C_{ii}^{-1}\right]^2}\frac{\partial C_{ii}^{-1}}{\partial \theta_j}\right\}.$$

We begin by deriving the partial derivatives of the model parameters, $\left[\alpha^T\ b\right]^T$, with respect to the hyper-parameter θ_j. The model parameters are given by the solution of a system of linear equations, such that

$$\left[\alpha^T\ b\right]^T = C^{-1}\left[y^T\ 0\right]^T.$$

Using the following identity for the partial derivatives of the inverse of a matrix,

$$\frac{\partial C^{-1}}{\partial \theta_j} = -C^{-1}\frac{\partial C}{\partial \theta_j}C^{-1}, \tag{13}$$

we obtain,

$$\frac{\partial\left[\alpha^T\ b\right]^T}{\partial \theta_j} = -C^{-1}\frac{\partial C}{\partial \theta_j}C^{-1}\left[y^T\ 0\right] = -C^{-1}\frac{\partial C}{\partial \theta_j}\left[\alpha^T\ b\right]^T.$$

Note the computational complexity of evaluating the partial derivatives of the model parameters is $O(\ell^2)$, as only two successive matrix-vector products are required. The partial derivatives of the diagonal elements of C^{-1} can be found using the inverse matrix derivative identity (13). For a kernel parameter, $\partial C/\partial \eta_j$ will generally be fully dense, and so the computational complexity of evaluating the diagonal elements of $\partial C^{-1}/\partial \eta_j$ will be $O(\ell^3)$ operations. If, on the other hand, we consider the regularisation parameter, μ, we have that

$$\frac{\partial C}{\partial \mu} = \begin{bmatrix} I & 0 \\ 0^T & 0 \end{bmatrix},$$

and so the computation of the partial derivatives of the model parameters, with respect to the regularisation parameter, is slightly simplified,

$$\frac{\partial\left[\alpha^T\ b\right]^T}{\partial \mu} = -C^{-1}\left[\alpha^T\ b\right]^T.$$

More importantly, as $\partial C/\partial \mu$ is diagonal, the diagonal elements of (13) can be evaluated with a computational complexity of only $O(\ell^2)$ operations. This suggests that it may be more efficient to adopt different strategies for optimising the regularisation parameter, μ, and the vector of kernel parameters, η, (cf., Saadi et al., 2004). For a kernel parameter, η_j, the partial derivatives of C with respect to η_j are given by the partial derivatives of the kernel matrix, that is,

$$\frac{\partial C}{\partial \eta_j} = \begin{bmatrix} \partial K/\partial \eta_j & 0 \\ 0^T & 0 \end{bmatrix}.$$

For the spherical radial basis function kernel, used in this study, the partial derivative with respect to the kernel parameter is given by

$$\frac{\partial \mathcal{K}(x,x')}{\partial \eta} = -\mathcal{K}(x,x')\|x-x'\|^2.$$

Finally, since the regularisation parameter, μ, and the scale parameter of the radial basis function kernel are strictly positive quantities, in order to permit the use of an unconstrained optimisation procedure, we adopt the parameterisation $\widetilde{\theta}_j = \log_2 \theta_j$, such that

$$\frac{\partial Q(\theta)}{\partial \widetilde{\theta}_j} = \frac{\partial Q(\theta)}{\partial \theta_j}\frac{\partial \theta_j}{\partial \widetilde{\theta}_j} \qquad \text{where} \qquad \frac{\partial \theta_j}{\partial \widetilde{\theta}_j} = \theta_j \log 2.$$

1.3.2 AUTOMATIC RELEVANCE DETERMINATION

Automatic Relevance Determination (ARD) (e.g., Rasmussen and Williams, 2006), also known as feature scaling (Chapelle et al., 2002; Bo et al., 2006), aims to identify informative input features as a natural consequence of optimising the model selection criterion. This can be most easily achieved using an *elliptical* radial basis function kernel,

$$\mathcal{K}(x,x') = \exp\left\{-\sum_{i=1}^{d}\eta_i[x_i - x_i']^2\right\},$$

that incorporates individual scaling factors for each input dimension. The partial derivatives with respect to the kernel parameters are then given by,

$$\frac{\partial\mathcal{K}(x,x')}{\partial\eta_i} = -\mathcal{K}(x,x')[x_i - x_i']^2.$$

Generalisation performance is likely to be enhanced if irrelevant features are down-weighted. It is therefore hoped that minimising the model selection criterion will lead to very small values for the scaling factors associated with redundant input features, allowing them to be identified and pruned from the model.

2. Bayesian Regularisation in Model Selection

In order to overcome the observed over-fitting in model selection using leave-one-out cross-validation based methods, we propose to add a regularisation term (Tikhonov and Arsenin, 1977) to the model selection criterion, which penalises solutions where the kernel parameters take on unduly large values. The regularised model selection criterion is then given by

$$M(\theta) = \zeta Q(\theta) + \xi\Omega(\theta), \tag{14}$$

where ξ and ζ are additional regularisation parameters, $Q(\theta)$ is the model selection criterion, in this case the PRESS statistic and $\Omega(\theta)$ is a regularisation term,

$$Q(\theta) = \frac{1}{2}\sum_{i=1}^{\ell}\left[y_i - \hat{y}_i^{(-i)}\right]^2 \qquad \text{and} \qquad \Omega(\theta) = \frac{1}{2}\sum_{i=1}^{d}\eta_i^2.$$

In this study we have left the regularisation parameter, μ, unregularised. However, we have now introduced two further regularisation parameters ξ and ζ for which good values must also be found. This problem may be solved by taking a Bayesian approach and adopting an ignorance prior and integrating out the additional regularisation parameters analytically in the style of Buntine and Weigend (1991). Adapting the approach taken by Williams (1995), the regularised model selection criterion (14) can be interpreted as the posterior density in the space of the hyper-parameters,

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta),$$

by taking the negative logarithm and neglecting additive constants. Here $P(\mathcal{D}|\theta)$ represents the *likelihood* with respect to the hyper-parameters and $P(\theta)$ represents our prior beliefs regarding the

hyper-parameters, in this case that they should have a small magnitude, corresponding to a relatively simple model. These quantities can be expressed as

$$P(\mathcal{D}|\theta) = Z_Q^{-1} \exp\{-\zeta Q(\theta)\} \qquad \text{and} \qquad P(\theta) = Z_\Omega^{-1} \exp\{-\xi\Omega(\theta)\}$$

where Z_Q and Z_Ω are the appropriate normalising constants. Assuming the data represent an i.i.d. sample, the *likelihood* in this case is Gaussian,

$$P(\mathcal{D}|\theta) = \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left[y_i - \hat{y}_i^{(-i)}\right]^2}{2\sigma^2} \right\} \quad \text{where} \quad \zeta = \frac{1}{\sigma^2} \implies Z_Q = \left(\frac{2\pi}{\zeta}\right)^{\ell/2}.$$

Likewise, the *prior* is a Gaussian, centred on the origin,

$$P(\theta) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi/\xi}} \exp\left\{ -\frac{\xi}{2}\eta_i^2 \right\} \qquad \text{such that} \qquad Z_\Omega = \left(\frac{2\pi}{\xi}\right)^{d/2}.$$

Minimising (14) is thus equivalent to maximising the posterior density with respect to the hyper-parameters. Note that the use of a prior over the hyper-parameters is in accordance with normal Bayesian practice and has been investigated in the case of Gaussian Process classifiers by Williams and Barber (1998). The combination of frequentist and Bayesian approaches at the first and second levels of inference is however somewhat unusual. The marginal likelihood is dependent on the assumptions of the model, which may not be completely appropriate. Cross-validation based procedures may therefore be more robust in the case of model mis-specification (Wahba, 1990). It seems reasonable for the model to be less sensitive to assumptions at the second level of inference than the first, and so the proposed approach represents a pragmatic combination of techniques.

2.1 Elimination of Second Level Regularisation Parameters ξ and ζ

Under the *evidence framework* proposed by MacKay (1992a,b,c) the hyper-parameters ξ and ζ are determined by maximising the marginal likelihood, also known as the Bayesian *evidence* for the model. In this study, however we opt to integrate out the hyper-parameters analytically, extending the work by Buntine and Weigend (1991) and Williams (1995) to consider Bayesian regularisation at the second level of inference, namely the selection of good values for the hyper-parameters. We begin with the prior over the hyper-parameters, which depends on ξ,

$$P(\theta|\xi) = Z_\Omega(\xi)^{-1} \exp\{-\xi\Omega\}.$$

The regularisation parameter ξ may then be integrated out analytically using a suitable prior, $P(\xi)$,

$$P(\theta) = \int P(\theta|\xi)P(\xi)d\xi.$$

The improper Jeffreys' prior, $P(\xi) \propto 1/\xi$ is an appropriate ignorance prior in this case as ξ is a scale parameter, noting that ξ is strictly positive,

$$p(\theta) = \frac{1}{(2\pi)^{d/2}} \int_0^\infty \xi^{d/2-1} \exp\{-\xi\Omega\}\, d\xi.$$

Using the Gamma integral $\int_0^\infty x^{\nu-1}e^{-\mu x}dx = \Gamma(\nu)/\mu^\nu$ (Gradshteyn and Ryzhic, 1994, equation 3.384), we obtain

$$p(\theta) = \frac{1}{(2\pi)^{d/2}} \frac{\Gamma(d/2)}{\Omega^{d/2}} \implies -\log p(\theta) \propto \frac{d}{2}\log\Omega.$$

Finally, adopting a similar procedure to eliminate ζ, we obtain a revised model selection criterion with Bayesian regularisation,

$$L = \frac{\ell}{2}\log Q(\theta) + \frac{d}{2}\log\Omega(\theta). \tag{15}$$

in which the regularisation parameters have been eliminated. As before, this criterion can be optimised via standard methods, such as the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) or scaled conjugate gradient descent (Williams, 1991). The partial derivatives of the proposed Bayesian model selection criterion are given by

$$\frac{\partial L}{\partial \theta_i} = \frac{\ell}{2Q(\theta)}\frac{\partial Q(\theta)}{\partial \theta_i} + \frac{d}{2\Omega(\theta)}\frac{\partial \Omega(\theta)}{\partial \theta_i} \qquad \text{and} \qquad \frac{\partial \Omega(\theta)}{\partial \eta_i} = \eta_i.$$

The additional computational expense involved in Bayesian regularisation of the model selection criterion is only $O(d)$ operations, and is extremely small in comparison with the $O(\ell^3)$ operations involved in obtaining the leave-one-out error (including the cost of training the model on the entire data set). Per iteration of the model selection process, the cost of the Bayesian regularisation is therefore minimal. There seems little reason to suppose that the regularisation will have an adverse effect on convergence, and this seems to be the case in practice.

2.2 Relationship with the Evidence Framework

Under the evidence framework of MacKay (1992a,b,c) the regularisation parameters, ξ and ζ, are selected so as to maximise the marginal likelihood, also known as the Bayesian evidence, for the model. The log-evidence is given by

$$\log P(\mathcal{D}) = -\xi\Omega(\theta) - \zeta Q(\theta) - \frac{1}{2}\log|A| + \frac{d}{2}\log\xi + \frac{\ell}{2}\log\zeta - \frac{\ell}{2}\log\{2\pi\},$$

where A is the Hessian of the regularised model selection criterion (14) with respect to the hyper-parameters, θ. Setting the partial derivatives of the log evidence with respect to the regularisation parameters, ξ and ζ, equal to zero, we obtain the familiar update formulae,

$$\xi^{\text{new}} = \frac{\gamma}{2\Omega(\theta)} \qquad \text{and} \qquad \zeta^{\text{new}} = \frac{\ell - \gamma}{2Q(\theta)},$$

where γ is the number of well defined hyper-parameters, that is, the hyper-parameters for which the optimal value is primarily determined by the log-likelihood term, $Q(\theta)$ rather than by the regulariser, $\Omega(\theta)$. In the case of the L2 regularisation term, corresponding to a Gaussian prior, the number of well determined hyper-parameters is given by

$$\gamma = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \xi}$$

where, $\lambda_i, \ldots, \lambda_n$ represent the eigenvalues of the Hessian of the unregularised model selection criterion, $Q(\theta)$ with respect to the kernel parameters. Comparing the partial derivatives of the regularised model selection criterion (14) with those of the Bayesian criterion (15), reveals that the

Bayesian regularisation scheme is equivalent to optimising the regularised model selection criterion (14) assuming that the regularisation parameters, ξ and ζ, are continuously updated according to the following update rules,

$$\xi^{\text{eff}} = \frac{d}{2\Omega(\theta)} \qquad \text{and} \qquad \zeta^{\text{eff}} = \frac{\ell}{2Q(\theta)}.$$

This exactly corresponds to the "cheap and cheerful" approximation of the evidence framework suggested by MacKay (1994), which assumes that all of the hyper-parameters are well-determined and that the number of hyper-parameters is small in relation to the number of training patterns. Since $\gamma \leq d$, it seems self evident that the proposed Bayesian regularisation scheme will be prone to a degree of under-fitting, especially in the case of a feature scaling kernel with many redundant features. The theoretical and practical pros and cons of the integrate-out approach and the evidence framework are discussed in some detail by MacKay (1994) and Bishop (1995) and references therein. However, the integrate-out approach does not require the evaluation of the Hessian matrix of the original selection criterion, $Q(\theta)$, which is likely to prove computationally prohibitive.

3. Results

In this section, we present experimental results demonstrating the benefits of the proposed model selection strategy incorporating Bayesian regularisation to overcome the inherent high variance of leave-one-out cross-validation based selection criteria. Table 2 shows a comparison of the error rates of least-squares support vector machines, using model selection procedures with, and without, Bayesian regularisation, (LS-SVM and LS-SVM-BR respectively) over the suite of thirteen public domain benchmark data sets used in the study by Mika et al. (2000). Results obtained using a Gaussian process classifier (Rasmussen and Williams, 2006), based on the expectation propagation method, are also provided for comparison (EP-GPC). The same set of 100 random partitions of the data (20 in the case of the larger `image` and `splice` benchmarks) to form training and test sets used in that study are also used here. In each case, model selection is performed independently for each realisation of the data set, such that the standard errors reflect the variability of both the training algorithm and the model selection procedure with changes in the sampling of the data. Both conventional spherical and elliptical radial basis kernels are used for all kernel learning methods, so that the effectiveness of each algorithm for automatic relevance determination can be assessed. The use of multiple training/test partitions allows an estimate of the statistical significance of differences in performance between algorithms to be computed. Let $\hat{x}$ and $\hat{y}$ represent the means of the performance statistic for a pair of competing algorithms, and e_x and e_y the corresponding standard errors, then the z statistic is computed as

$$z = \frac{\hat{y} - \hat{x}}{\sqrt{e_x^2 + e_y^2}}.$$

The z-score can then be converted to a significance level via the normal cumulative distribution function, such that $z = 1.64$ corresponds to a 95% significance level. All statements of statistical significance in the remainder of this section refer to a 95% level of significance.

3.1 Performance of Models Based on the Spherical RBF Kernel

The results shown in the first three data columns of Table 2 show the performance of LS-SVM, LS-SVM-BR and EP-ARD models based on the spherical Gaussian kernel. The performance of

Data Set	Training Patterns	Testing Patterns	Number of Replications	Input Features
Banana	400	4900	100	2
Breast cancer	200	77	100	9
Diabetis	468	300	100	8
Flare solar	666	400	100	9
German	700	300	100	20
Heart	170	100	100	13
Image	1300	1010	20	18
Ringnorm	400	7000	100	20
Splice	1000	2175	20	60
Thyroid	140	75	100	5
Titanic	150	2051	100	3
Twonorm	400	7000	100	20
Waveform	400	4600	100	21

Table 1: Details of data sets used in empirical comparison.

LS-SVM models with and without Bayesian regularisation are very similar, with neither model proving significantly better than the other on any of the data sets. This seems reasonable given that only two hyper-parameters are optimised during model selection and so there is little scope for over-fitting the PRESS model selection criterion and the regularisation term will have little effect. The LS-SVM model with Bayesian regularisation is significantly out-performed by the Gaussian Process classifier on one benchmark `banana`, but performs significantly better on a further four (`ringnorm`, `splice`, `twonorm`, `waveform`). Demšar (2006) recommends the use of the Wilcoxon signed rank test for assessing the statistical significance of differences in performance over multiple data sets. According to this test, neither the LSSVM-BR nor the EP-PGC is statistically superior at the 95% level of significance.

3.2 Performance of Models Based on the Elliptical RBF Kernel

The performances of LS-SVM, LS-SVM-BR and EP-GPC models based on the elliptical Gaussian kernel, which includes a separate scale parameter for each input feature, are shown in the last three columns of Table 2. Before evaluating the effects of Bayesian regularisation in model selection it is worth noting that the use of elliptical RBF kernels does not generally improve performance. For the LS-SVM, the elliptical kernel produces significantly better results on only two benchmarks (`image` and `splice`) and significantly worse results on a further eight (`banana`, `breast cancer`, `diabetis`, `german`, `heart`, `ringnorm`, `twonorm`, `waveform`), with the degradation in performance being very large in some cases (e.g., `heart`). This seems likely to be a result of the additional degrees of freedom involved in the model selection process, allowing over-fitting of the PRESS model selection criterion as a result of its inherently high variance. Note that fully Bayesian approaches, such as the Gaussian Process Classifier, are also unable to reliably select kernel parameters for the elliptical RBF kernel. The elliptical kernel is significantly better on only three benchmarks (`flare solar`,

Data Set	Radial Basis Function			Automatic Relevance Determination		
	LSSVM	**LSSVM-BR**	**EP-GPC**	**LSSVM**	**LSSVM-BR**	**EP-GPC**
Banana	10.60 ± 0.052	10.59 ± 0.050	$\mathbf{10.41 \pm 0.046}$	10.79 ± 0.072	10.73 ± 0.070	*10.46 ± 0.049*
Breast cancer	*26.73 ± 0.466*	27.08 ± 0.494	$\mathbf{26.52 \pm 0.489}$	29.08 ± 0.415	27.81 ± 0.432	27.97 ± 0.493
Diabetes	23.34 ± 0.166	$\mathbf{23.14 \pm 0.166}$	*23.28 ± 0.182*	24.35 ± 0.194	23.42 ± 0.177	23.86 ± 0.193
Flare solar	34.22 ± 0.169	34.07 ± 0.171	34.20 ± 0.175	34.39 ± 0.194	*33.61 ± 0.151*	$\mathbf{33.58 \pm 0.182}$
German	*23.55 ± 0.216*	23.59 ± 0.216	$\mathbf{23.36 \pm 0.211}$	26.10 ± 0.261	23.88 ± 0.217	23.77 ± 0.221
Heart	*16.64 ± 0.358*	$\mathbf{16.19 \pm 0.348}$	16.65 ± 0.287	23.65 ± 0.355	17.68 ± 0.623	19.68 ± 0.366
Image	3.00 ± 0.158	2.90 ± 0.154	2.80 ± 0.123	$\mathbf{1.96 \pm 0.115}$	*2.00 ± 0.113*	2.16 ± 0.068
Ringnorm	$\mathbf{1.61 \pm 0.015}$	$\mathbf{1.61 \pm 0.015}$	*4.41 ± 0.064*	2.11 ± 0.040	1.98 ± 0.026	8.58 ± 0.096
Splice	10.97 ± 0.158	10.91 ± 0.154	11.61 ± 0.181	*5.86 ± 0.179*	$\mathbf{5.14 \pm 0.145}$	7.07 ± 0.765
Thyroid	4.68 ± 0.232	4.63 ± 0.218	*4.36 ± 0.217*	4.68 ± 0.199	4.71 ± 0.214	$\mathbf{4.24 \pm 0.218}$
Titanic	$\mathbf{22.47 \pm 0.085}$	22.59 ± 0.120	22.64 ± 0.134	*22.58 ± 0.108*	22.86 ± 0.199	22.73 ± 0.134
Twonorm	$\mathbf{2.84 \pm 0.021}$	$\mathbf{2.84 \pm 0.021}$	*3.06 ± 0.034*	5.18 ± 0.072	4.53 ± 0.077	4.02 ± 0.068
Waveform	*9.79 ± 0.045*	$\mathbf{9.78 \pm 0.044}$	10.10 ± 0.047	13.56 ± 0.141	11.48 ± 0.177	11.34 ± 0.195

Table 2: Error rates of least-squares support vector machine, with and without Bayesian regularisation of the model selection criterion, in this case the PRESS statistic (Allen, 1974), and Gaussian process classifiers over thirteen benchmark data sets (Rätsch et al., 2001), using both standard radial basis function and automatic relevance determination kernels. The results for the EP-GPC were obtained using the MATLAB software accompanying the book by Rasmussen and Williams (2006). The results for each method are presented in the form of the mean error rate over test data for 100 realisations of each data set (20 in the case of the image and splice data sets), along with the associated standard error. The best results are shown in boldface and the second best in italics (without implication of statistical significance).

image and splice), while being significantly worse on six (breast cancer, diabetis, heart, ringnorm, twonorm and waveform).

In the case of the elliptical RBF kernel, the use of Bayesian regularisation in model selection has a dramatic effect on the performance of LS-SVM models, with the LS-SVM-BR model proving significantly better than the conventional LS-SVM on nine of the thirteen benchmarks (breast cancer, diabetis, flare solar, german, heart, ringnorm, splice, twonorm and waveform) without being significantly worse on any of the remaining four data sets. This demonstrates that over-fitting in model selection, due to the larger number of kernel parameters, is likely to be the significant factor causing the relatively poor performance of models with the elliptical RBF kernel. Again, the Gaussian Process classifier is significantly better than the LS-SVM with Bayesian regularisation on the banana and twonorm data sets, but is significantly worse on four of the remaining eleven (diabetis, heart, ringnorm and splice). Again, according to the Wilcoxon signed rank test, neither the LSSVM-BR nor the EP-PGC is statistically superior at the 95% level of significance. However the magnitude of the difference in performance between LSSVM-BR and EP-GPC approaches tends to be greatest when the LSSVM-BR out-performs EP-GPC, most notably on the heart, splice and ringnorm data sets. This provides some support for the observation of Wahba (1990) that cross-validation based model selection procedures should be more robust against model mis-specification (see also Rasmussen and Williams, 2006).

4. Discussion

The experimental evaluation presented in the previous section demonstrates that over-fitting can occur in model selection, due to the variance of the model selection criterion. In many cases the minimum of the selection criterion using the elliptical RBF kernel is lower than that achievable using the spherical RBF kernel, however this results in a degradation in generalisation performance. Using the PRESS statistic, the over-fitting is likely to be most severe in cases with a small number of training patterns, as the variance of the leave-one-out estimator decreases as the sample size becomes larger. Using the standard LSSVM, the elliptical RBF kernel only out-performs the spherical RBF kernel on two of the thirteen data sets, image and splice, which also happen to be the two largest data sets in terms of the number of training patterns. The greatest degradation in performance is obtained on the heart benchmark, the third smallest. The heart data set also has a relatively large number of input features (13). A large number of input features introduces a many additional degrees of freedom to optimise the model selection criterion, and so will generally tend to encourage over-fitting. However, there may be a compact subset of highly relevant features with the remainder being almost entirely uninformative. In this case the advantage of suppressing the noisy inputs is so great that it overcomes the predisposition towards over-fitting, and so results in improved generalisation (as observed in the case of the image and splice benchmarks). Whether the use of an elliptical RBF kernel will improve or degrade generalisation largely depends on such characteristics of the data that are not known *a-priori*, and so it seems prudent to consider a range of kernel functions and select the best via cross-validation.

The experimental results indicate that Bayesian regularisation of the hyper-parameters is generally beneficial, without at this stage providing a complete solution to the problem of over-fitting the model selection criterion. The effectiveness of the Bayesian regularisation scheme is to a large extent dependent on the appropriateness of the prior imposed on the hyper-parameters. There is no reason to assume that the simple Gaussian prior used here is in any sense optimal, and this is an

issue where further research is necessary (see Section 4.2). The comparison of the integrate-out approach and the evidence framework highlights a deficiency of the simple Gaussian prior. It suggests that the integrate-out approach is likely to result in mild over-regularisation of the hyper-parameters in the presence of a large number of irrelevant features, as the corresponding hyper-parameters will be ill-determined.

The LSSVM with Bayesian regularisation of the hyper-parameters does not significantly outperform the expectation propagation based Gaussian process classifier over the suite of thirteen benchmark data sets considered. This is not wholly surprising as the EP-GPC is at least very close to the state-of-the-art, indeed it is interesting that the EP-GPC does not out-perform such a comparatively simple model. The EP-GPC uses the marginal likelihood as the model selection criterion, which gives the probability of the data, *given the assumptions of the model* (Rasmussen and Williams, 2006). Cross-validation based approaches, on the other hand, provide an estimate of generalisation performance that does not depend on the model assumptions, and so may be more robust against model mis-specification (Wahba, 1990). The no free lunch theorems suggest that, at least in terms of generalisation performance, there is a lack of inherent superiority of one classification method over another, in the absence of *a-priori* assumptions regarding the data. This implies that if one classifier performs better than another on a particular data set it is because the inductive biases of that classifier provide a better fit to the particular pattern recognition task, rather than to its superiority in a more general sense. A model with strong inductive biases is likely to benefit when these biases are well suited to the data, but will perform badly when they do not. While a model with weak inductive biases will be more robust, it is less likely to perform conspicuously well on any given data set. This means there are complementary advantages and disadvantages to both approaches.

4.1 Relationship to Existing Work

The use of a prior over the hyper-parameters is in accordance with normal Bayesian practice and has been used in Gaussian Process classification (Williams and Barber, 1998). The problem of over-fitting in model selection has also been addressed by Qi et al. (2004), in the case of selecting informative features for a logistic regression model using an Automatic Relevance Determination (ARD) prior (cf., Tipping, 2000). In this case, the Expectation Propagation method (Minka, 2001) is used to obtain a deterministic approximation of the posterior, and also (as a by-product) a leave-one-out performance estimate. The latter is then used to implement a form of early-stopping (e.g., Sarle, 1995) to prevent over-fitting resulting from the direct optimization of the marginal likelihood until convergence. It seems likely that this approach would be also be beneficial in the case of tuning the hyper-parameters of the covariance function of Gaussian process model, using either the leave-one-out estimate arising from the EP approximation, or an approximate leave-one-out estimate from the Laplace approximation (cf., Cawley and Talbot, 2007).

4.2 Directions for Further Research

In this paper, the regularisation term corresponds to a simple spherical Gaussian prior over the kernel parameters. One direction of research would be to investigate alternative regularisation terms. The

first possibility would be to use a regularisation term corresponding to a separable Laplace prior,

$$\Omega(\theta) = \frac{1}{2}\sum_{i=1}^{d} |\eta_i| \qquad \Longrightarrow \qquad p(\theta) = \prod_{i=1}^{d} \frac{\xi}{2} \exp\{-\xi|\eta_i|\}.$$

Setting the derivative of the regularised model selection criterion (14) to zero, we obtain

$$\left|\frac{\partial Q}{\partial \eta_i}\right| = \frac{\xi}{\zeta} \qquad \text{if } |\eta_i| > 0 \qquad \text{and} \qquad \left|\frac{\partial Q}{\partial \eta_i}\right| < \frac{\xi}{\zeta} \qquad \text{if } |\eta_i| = 0,$$

which implies that if the sensitivity of the leave-one-out error, $Q(\theta)$, falls below ξ/ζ, the value of the hyper-parameter, η_i will be set exactly to zero, effectively pruning that input from the model. In this way explicit feature selection may be obtained as a consequence of (regularised) model selection. The model selection criterion with Bayesian regularisation then becomes

$$L = \frac{\ell}{2}\log Q(\theta) + N \log \Omega(\theta)$$

where N is the number of input features with non-zero scale factors. This potentially overcomes the propensity towards under-fitting the data that might be expected when using the Gaussian prior, as the pruning action of the Laplace prior means that the values of all remaining hyper-parameters are well-determined by the data. In the case of the Laplace prior, the integrate-out approach is exactly equivalent to continuous updates of the hyper-parameters according to the update formulae under the evidence framework (Williams, 1995). Alternatively, defining a prior over the *function* of a model seems more in accordance with Bayesian ideals than choosing a prior over the parameters of the model. The use of a prior over the hyper-parameters based on the smoothness of the resulting model also provides a potential direction for future research. In this case, the regularisation term might take the form,

$$\Omega(\theta) = \frac{1}{2\ell}\sum_{i=1}^{\ell}\sum_{j=1}^{d}\left[\frac{\partial^2 \hat{y}_i}{\partial x_{ij}^2}\right]^2,$$

directly penalising models with excess curvature. This regularisation term corresponds to curvature driven smoothing in multi-layer perceptron networks (Bishop, 1993), except that the model output $\hat{y}_i$ is viewed as a function of the hyper-parameters, rather than of the weights. A penalty term based on the first-order partial derivatives is also feasible (cf., Drucker and Le Cun, 1992).

5. Conclusion

Leave-one-out cross-validation has proved to be an effective means of model selection for a variety of kernel learning methods, provided the number of hyper-parameters to be tuned is relatively small. The use of kernel functions with large numbers of parameters often provides sufficient degrees of freedom to over-fit the model selection criterion, leading to poor generalisation. In this paper, we have proposed the use of regularisation at the second level of inference, that is, model selection. The use of Bayesian regularisation is shown to be effective in reducing over-fitting, by ensuring the values of the kernel parameters remain small, giving a smoother kernel and hence a less complex classifier. This is achieved with only a minimal computational expense as the additional regularisation parameters are integrated out analytically using a reference prior. While a fully Bayesian

model selection strategy is conceptually more elegant, it may also be less robust to model mis-specification. The use of leave-one-out cross-validation in model selection and Bayesian methods at the next level seems to be a pragmatic compromise. The effectiveness of this approach is clearly demonstrated in the experimental evaluation where, on average, the LS-SVM with Bayesian regularisation out-performs the expectation-propagation based Gaussian process classifier, using both spherical and elliptical RBF kernels.

Acknowledgments

The authors would like to thank the organisers of the WCCI model selection workshop and performance prediction challenge and the NIPS multi-level inference workshop and model selection game, and fellow participants for the stimulating discussions that have helped to shape this work. We also thank Carl Rasmussen and Chris Williams for their advice regarding the EP-GPC and the anonymous reviewers for their detailed and constructive comments that have significantly improved this paper.

References

D. M. Allen. The relationship between variable selection and prediction. *Technometrics*, 16:125–127, 1974.

E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorenson. *LAPACK Users' Guide*. SIAM Press, third edition, 1999.

C. M. Bishop. Curvature-driven smoothing: a learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 4(5):882–884, September 1993.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

L. Bo, L. Wang, and L. Jiao. Feature scaling for kernel Fisher discriminant analysis using leave-one-out cross-validation. *Neural Computation*, 18:961–978, April 2006.

W. L. Buntine and A. S. Weigend. Bayesian back-propagation. *Complex Systems*, 5:603–643, 1991.

G. C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN-2006)*, pages 2970–2977, Vancouver, BC, Canada, July 16–21 2006.

G. C. Cawley and N. L. C. Talbot. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592, November 2003.

G. C. Cawley and N. L. C. Talbot. Fast leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17(10):1467–1475, December 2004.

G. C. Cawley and N. L. C. Talbot. Approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning* (submitted), 2007.

C. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

H. Drucker and Y. Le Cun. Improving generalization performance using double back-propagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilema. *Neural Computation*, 4(1):1–58, 1992.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, third edition edition, 1996.

I. S. Gradshteyn and I. M. Ryzhic. *Table of Integrals, Series and Products*. Academic Press, fifth edition, 1994.

I. Guyon, A. R. Saffari Azar Alamdari, G. Dror, and J. Buhmann. Performance prediction challenge. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN-2006)*, pages 1649–1656, Vancouver, BC, Canada, July 16–21 2006.

T. Joachims. *Learning to Classify Text using Support Vector Machines - Methods, Theory and Algorithms*. Kluwer Academic Publishers, 2002.

S. S. Keerthi, K. B. Duan, S. K. Shevade, and A. N. Poo. A fast dual algorithm for kernel logistic regression. *Machine Learning*, 61(1–3):151–165, November 2005.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence (IJCAI)*, pages 1137–1143, San Mateo, CA, 1995. Morgan Kaufmann.

P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, February 1968.

A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition (in Russian). *Techicheskaya Kibernetica*, 3, 1969.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992a.

D. J. C. MacKay. A practical Bayesian framework for backprop networks. *Neural Computation*, 4 (3):448–472, 1992b.

D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992c.

D. J. C. MacKay. Hyperparameters: Optimise or integrate out? In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*. Kluwer, 1994.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London, A*, 209:415–446, 1909.

C. A. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, volume IX, pages 41–48. IEEE Press, New York, 1999.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. J. Smola, and K.-R. Müller. Invariant feature extraction and classification in feature spaces. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 526–532. MIT Press, 2000.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kauffmann, 2001.

J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7: 308–313, 1965.

Y. Qi, T. P. Minka, R. W. Picard, and Z. Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of the 21st International Conference on Machine Learning*, pages 671–678, Banff, Alberta, Canada, July 4–8 2004.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006.

G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. *Machine Learning*, 42(3): 287–320, 2001.

K. Saadi, N. L. C. Talbot, and G. C. Cawley. Optimally regularised kernel Fisher discriminant analysis. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR-2004)*, volume 2, pages 427–430, Cambridge, United Kingdom, August 23–26 2004.

W. S. Sarle. Stopped training and other remidies for overfitting. In *Proceedings of the 27*th *Symposium on the Interface of Computer Science and Statistics*, pages 352–360, Pittsburgh, PA, USA, June 21–24 1995.

B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel Methods in Computational Biology*. MIT Press, 2004.

T. Seaks. SYMINV : An algorithm for the inversion of a positive definite matrix by the Cholesky decomposition. *Econometrica*, 40(5):961–962, September 1972.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, B 36(1):111–147, 1974.

S. Sundararajan and S. S. Keerthi. Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Computation*, 13(5):1103–1118, May 2001.

J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, June 1999.

J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. John Wiley, New York, 1977.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, June 2000.

G. Wahba. *Spline Models for Observational Data*. SIAM Press, Philadelphia, PA, 1990.

C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, December 1998.

P. M. Williams. A Marquardt algorithm for choosing the step size in backpropagation learning with conjugate gradients. Technical Report CSRP-229, University of Sussex, February 1991.

P. M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7 (1):117–143, 1995.