Kernel Learning at the First Level of Inference

Gavin C. Cawley[a,*], Nicola L. C. Talbot[a]

[a]*School of Computing Sciences, University of East Anglia, Norwich. NR4 7TJ, U.K.*

Abstract

Kernel learning methods, whether Bayesian or frequentist, typically involve multiple levels of inference, with the coefficients of the kernel expansion being determined at the first level and the kernel and regularisation parameters carefully tuned at the second level, a process known as *model selection*. Model selection for kernel machines is commonly performed via optimisation of a suitable model selection criterion, often based on cross-validation or theoretical performance bounds. However, if there are a large number of kernel parameters, as for instance in the case of automatic relevance determination (ARD), there is a substantial risk of over-fitting the model selection criterion, resulting in poor generalisation performance. In this paper we investigate the possibility of learning the kernel, for the Least-Squares Support Vector Machine (LS-SVM) classifier, at the first level of inference, i.e. parameter optimisation. The kernel parameters and the coefficients of the kernel expansion are jointly optimised at the first level of inference, minimising a training criterion with an additional regularisation term acting on the kernel parameters. The key advantage of this approach is that the values of only two regularisation parameters need be determined in model selection, substantially alleviating the problem of over-fitting the model selection criterion. The benefits of this approach are demonstrated using a suite of synthetic and real-world binary classification benchmark problems, where kernel learning at the first level of inference is shown to be statistically superior to the conventional approach, improves on our previous work (Cawley and Talbot, 2007) and is competitive with Multiple Kernel Learning approaches, but with reduced computational expense.

Keywords: Kernel methods, model selection, regularisation, over-fitting, automatic relevance determination.

[*] Corresponding author

Email addresses: gcc@cmp.uea.ac.uk (Gavin C. Cawley), nlct@cmp.uea.ac.uk (Nicola L. C. Talbot)

1. Introduction

The training procedures for artificial neural networks (Bishop, 1995; MacKay, 1992), kernel learning methods (Schölkopf and Smola, 2002) and Gaussian process classifiers (Williams and Barber, 1998; MacKay, 1998; Rasmussen and Williams, 2006), can be viewed as multi-level optimisation problems (Guyon et al., 2009). The model parameters are optimised at the first level of inference, for instance the weights of an artificial neural network, or the coefficients of the kernel expansion of a kernel machine. However, there are normally a number of *hyper-parameters* that must be determined, for example the number of hidden layer units in a multi-layer perceptron network, the choice of kernel and the values of any associated kernel parameters for a kernel machine, or regularisation parameters controlling the complexity of the model. These hyper-parameters are normally optimised at a second level of inference, a process known as *model selection* (Guyon, 2009). The division between parameters and hyper-parameters typically arises due to computational considerations. The dual parameters of a kernel machine, for example, are generally given by the solution of a convex optimisation problem, for which computationally efficient algorithms are available (Boyd and Vandenberghe, 2004). It is therefore computationally convenient to alternate between optimising the coefficients of the kernel expansion at the first level of inference and optimising the kernel and regularisation parameters at the second level of inference, taking advantage of the simple mathematical structure of the problem at the first level of inference.

In the case of kernel learning methods, the convex nature of the optimisation problem at the first level of inference implies a single, global optimum, thus avoiding the potential pitfall of multiple local minima that complicates the application of multi-layer perceptron networks. However, in order to maximise generalisation performance in practical applications, the values of a small number of regularisation and kernel parameters must also be carefully tuned during model selection (Chapelle et al., 2002). This is most often achieved via minimisation of a cross-validation estimate of generalisation performance, using grid search, Nelder-Mead simplex (Nelder and Mead, 1965) or gradient descent-based methods (Chapelle et al., 2002). This approach has been shown to be highly effective for kernel machines with a small number of hyper-parameters (e.g. Cawley, 2006). However, as the number of hyper-parameters becomes large, there is an increasing risk of over-fitting the model selection criterion, resulting in poor performance (Cawley and Talbot, 2007, 2010). Chapelle (2002) suggests the additional estimation error might reasonably be expected to grow with the square root of the number of hyper-parameters. This danger has been observed previously (Bengio, 2000), and is especially evident in studies involving Automatic Relevance Determination (ARD), where the kernel includes separate scaling parameters for each feature. It is also well understood that the model selection criterion should not be also used for performance estimation as its direct optimisation during model selection will introduce an optimistic bias, and hence procedures such as nested cross-validation are necessary (Cherkassky and Mulier, 1998; Hastie et al., 2001; Cawley and Talbot, 2010). While over-fitting of the model selection criterion is clearly a significant problem, research towards a potential solution appears to have received relatively little attention. Cawley and Talbot (2007) propose the addition of a regularisation term to the model selection criterion penalising large values of the kernel parameters, and thus promoting a relatively smooth model. Regularisation of the kernel parameters is shown to be effective in some cases, however the problem of over-fitting in model selection is far from solved. The use of automatic relevance determination has several distinct benefits, including (c.f. Chapelle et al., 2002):

- The potential for improved generalisation performance — it is intuitively reasonable to expect that surpressing irrelevant attributes should result in improvements in accuracy.
- Explanation of the data — determination of which attributes have useful explanatory power, and which do not, is often a useful scientific finding.
- Reduced cost of data collection — if redundant attributes can be identified and eliminated, there is no need to determine the values of that attribute in operation. In some applications (such as medical diagnosis, where some screening tests are more expensive to conduct than others), the cost of evaluating the attributes may be an important practical consideration.

Thus, even if the use of automatic relevance determination does not give a performance advantage over the more basic RBF kernel, it is worth developing methods to avoid over-fitting in model selection so that the second and third benefits of ARD can be obtained more fully and reliably. In many applications, especially where data are in limited supply, a simple but incorrect model will out-perform a more correct, but more complex model because the parameters of the model can be estimated more reliably. A common example is the use of naive Bayes in text classification, where the assumption of independence is clearly not justified. If explaining the data is an important concern, the correct model should be used, and methods developed to allow the parameters to be estimated more accurately and reliably.

The approach presented in this paper seeks to minimise the risk of over-fitting in model selection by minimising the number of hyper-parameters to be optimised during model selection, hence minimising the degrees of freedom available to over-fit the model selection criterion. This is achieved by demoting the selection of kernel parameters from the second level of inference to the first, such that they are jointly optimised with the dual model parameters, minimising a single regularised training criterion. An additional regularisation term is used to penalise values of the kernel parameters likely to result in poor generalisation performance. As the values of only two regularisation parameters need then be determined in model selection, it is reasonable to expect the chance of over-fitting the model selection criterion to be substantially reduced, even when many kernel parameters are used. The optimisation of kernel parameters at the first level of inference is similar to the design of radial basis function networks via gradient descent methods (Webb and Shannon, 1998); however the addition of a regularisation term is required to maintain generalisation performance.

The remainder of this paper is structured as follows: Section 2 describes a training algorithm for kernel ridge regression with optimisation of the kernel parameters at the first level of inference. Results obtained on a suite of synthetic and real-world benchmark datasets is presented in Section 3. Section 4 provides discussion, including suggestions for further research and recommendations for practical applications. Finally, the work is summarised and conclusions drawn in Section 5.

2. Kernel Learning at the First Level of Inference

Let $\mathcal{D} = \{(\boldsymbol{x}_i,\ y_i)\}_{i=1}^{\ell}$, represent the training sample, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of explanatory variables describing the i^{th} example, and $y_i \in \{-1, +1\}$, is the corresponding desired response indicating the class to which the example belongs. The Least-Squares Support Vector Machine (LS-SVM) classifier (Suykens et al., 2002) constructs a linear classifier, $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b$, in a feature space, $\mathcal{F}$, defined via a fixed transformation $\phi : \mathcal{X} \to \mathcal{F}$. However, rather than define the feature space directly, it is instead induced by a positive definite kernel function, $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, giving the inner product between points in the feature space, such that $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$. In this study, we adopt the simple Gaussian Radial Basis Function (RBF) kernel,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \exp\left(-\theta_1 \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right), \tag{1}$$

where θ_1 is a kernel parameter controlling the sensitivity of the kernel, and the automatic relevance determination (ARD) or feature scaling variant of the RBF kernel (Chapelle et al., 2002),

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \exp\left(-\sum_{i=1}^{d} \theta_i [\boldsymbol{x}_i - \boldsymbol{x}_i']^2\right), \tag{2}$$

where θ_i are kernel parameters allowing the sensitivity of the kernel with respect to each of the explanatory variables to be tuned independently. Ideally, the kernel parameters associated with irrelevant features will adopt very small values, implementing a form of Automatic Relevance Determination (ARD) (MacKay, 1994; Neal, 1996). For fixed $\boldsymbol{\theta}$, the primal model parameters, $(\boldsymbol{w}, b)$, are given by the minimiser of a convex training criterion

$$\mathcal{L}(\boldsymbol{w}, b) = \sum_{i=1}^{\ell} c\left(y_i, f(\boldsymbol{x}; \boldsymbol{w}, b)\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2,$$

where $c(\cdot, \cdot)$ is a convex loss (in this case, the squared loss $c(y, f) = 0.5(y - f)^2$) representing the data misfit and λ is a regularisation parameter controlling the bias-variance trade-off (Geman et al., 1992). It can be shown that the vector of model parameters, $\boldsymbol{w}$, can be expressed as an expansion over the training examples, such that

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

where $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^{\ell}$ is a vector of *dual* model parameters. For a fixed value of the regularisation parameter, λ, the optimal dual model parameters are given by the solution of a system of linear equations,

$$\left[\begin{array}{cc} \boldsymbol{K} + \lambda \boldsymbol{I} & \boldsymbol{1} \\ \boldsymbol{1}^T & 0 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\alpha} \\ b \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y} \\ 0 \end{array}\right] \tag{3}$$

where

$$\boldsymbol{K} = \left[K_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_{i,j=1}^{\ell}, \tag{4}$$

which may be found efficiently via Cholesky decomposition of $\boldsymbol{K} + \lambda \boldsymbol{I}$ (Suykens et al., 2002).

The use of a squared loss is common in regression tasks, and may seem somewhat incongruous in a statistical pattern recognition setting, however there are theoretical and practical justifications for it's use. Firstly, the use of the least-squares loss in regression on class labels assymptotically provides estimates of Bayesian *a-posteriori* probability of class membership (White, 1989). Furthermore, Saerens et al. (2002) shows that *any* reasonable cost function (including the least-squares loss) can be used for *a posteriori* probability estimation. Statistical learning theory provides similar bounds, based on covering numbers, on the generalisation peformance of regularized least squares and support vector machines (Vapnik, 1998; Bousquet and Elisseeff, 2002; Rifkin, 2004). Secondly, it can be shown (e.g. Webb and Copsey, 2011) that least-squares regression on the class labels is equivalent to Fisher's Linear Discriminant Analysis (LDA), a classic statistical pattern recognition method (Fisher, 1936). Similarly, there is a close correspondence (Xu et al., 2001; Zhang et al., 2010) between the LS-SVM and Kernel Fisher Discriminant (KFD) analysis (Mika et al., 1999). Van Gestel et al. (2002) present a Bayesian treatment of the LS-SVM classifier and motivate the use of a least-squeares loss for binary classification by considering the case where the data belonging to each class are distributed in feature space according to Gaussian distributions with identical covariance matrices (a common motivation for linear classifiers, e.g. Duda et al., 2001). If the linear model constructed in the feature space achieves an output of ±1 at the means of these Gaussian distributions, then the residual errors will also have a Gaussian distribution, matching the implied noise model for a least-squares loss function. Lastly, KRR (Saunders et al., 1998), LS-SVM (Suykens et al., 2002), KFD (Mika et al., 1999), Regularized Least Squares (Rifkin et al., 2003) and Proximal Support Vector Machine (Fung, 2001) classifiers are widely used[1] and have proven highly effective in practical applications and show state-of-the-art performance on benchmark learning tasks (e.g. Mika et al., 1999; Rifkin, 2002; Cawley and Talbot, 2003; Van Gestel et al., 2004). Kernel Logistic Regression, with a cross-entropy loss that is more obviously suited to statistical pattern recognition, does not out-perform kernel Fisher discriminant analysis (Cawley

[1] at the time of writing 370, 5406, 1660, 151 and 592 citations respectively, according to Google Scholar.

and Talbot, 2008). Furthermore, we have used these methods in winning entries in a number of open machine learning challenges and highly competitive baseline methods in others (Cawley, 2006; Guyon et al., 2008; Cawley, 2009, 2011). The principal benefit of the LS-SVM, is that it has a simple and efficient implementation, and the leave-one-out cross-validation error can be evaluated in closed form at essentially negligible computational expense (Cawley, 2006), providing an efficient model selection criterion.

In this study, rather than optimise the kernel parameters of a least-squares support vector machine classifier (Suykens et al., 2002) during model selection, we chose instead to optimise them jointly with the model parameters, at the first level of inference, by minimising a single training criterion,

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\theta}) = \sum_{i=1}^{\ell} c\left(y_i, f(\boldsymbol{x}_i)\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{\mu}{2}\|\boldsymbol{\theta}\|^2, \tag{5}$$

where the squared norm of the kernel parameters is used as a regularisation term penalising large values of the scaling parameters of the radial basis function or automatic relevance determination kernels, thus promoting a relatively smooth model and avoiding over-fitting. With only two regularisation parameters (λ, μ) to be determined at the second level of inference, the scope for over-fitting in model selection should also be reduced.

The additional regularisation term used here, based on the squared norm of the kernel parameters, is intended to express a preference for relatively smooth models, with small magnitude scaling parameters, such that the output of the model is not unduly sensitive to small changes in any of the inputs and thereby decrease the tendency of the model to overfit the training sample. This form of regularisation, also known as weight decay (Krogh and Hertz, 1991), is used to implement automatic relevance determination for multi-layer perceptron neural networks (MacKay, 1994; Neal, 1996; Bishop, 1995), with similar justification. This heuristic is further supported by performance bounds provided by Bartlett (1998) and Anthony and Bartlett (1999), which suggest that generalisation performance is more strongly affected by the magnitude of the weights (both input-to-hidden and hidden-to-output layer) than the size of the network. The scaling parameters of the automatic relevance determination variant of the radial basis function kernel are analogous to the input-to-hidden layer weights of an MLP network, in that they govern the flexibility of the feature space in which a linear discriminant is constructed. It is somewhat surprising that regularisation of the scaling parameters of radial basis function neural networks is not as commonly encountered as in MLP networks, given that over-fitting is an important issue for both types of neural network model. Lastly, the same choice of regularisation term proved beneficial in previous work (Cawley and Talbot, 2007), although in that case the leave-one-out cross-validation based model selection criterion was regularised, rather than the training criterion, providing further empirical justification.

2.1. Training Algorithm

It would be feasible to optimise both model and kernel parameters jointly via gradient descent methods; however for a fixed set of kernel parameters, the dual parameters are given by a system of linear equations (3), which may be solved efficiently via Cholesky decomposition. It seems sensible therefore to alternate between exact updates of the dual model parameters and gradient descent updates of the kernel parameters. Let $z_i = f(\boldsymbol{x}_i)$, then the necessary gradient information is given by

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta_r} &= \sum_{i=1}^{\ell}\left[\frac{\partial c_i}{\partial z_i} + \lambda\alpha_i\right]\sum_{j=1}^{\ell}\frac{\partial \alpha_j}{\partial \theta_r}K_{ij} \\ &+ \sum_{i=1}^{\ell}\left[\frac{\partial c_i}{\partial z_i} + \frac{\lambda\alpha_i}{2}\right]\sum_{j=1}^{\ell}\alpha_j\frac{\partial K_{ji}}{\partial \theta_r} \\ &+ \frac{\partial b}{\partial \theta_r}\sum_{i=1}^{\ell}\frac{\partial c_i}{\partial z_i} + \mu\theta_r,\end{aligned}$$

where $c_i = c\{y_i, z_i\}$. However, noting that at the minimum of $\mathcal{L}$, we have that

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \quad \Longrightarrow \quad \frac{\partial c_i}{\partial z_i} + \lambda\alpha_i = 0$$

and

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell}\frac{\partial c_i}{\partial z_i} = 0,$$

and so

$$\frac{\partial \mathcal{L}}{\partial \theta_r} = \sum_{i=1}^{\ell}\alpha_i\sum_{j=1}^{\ell}\frac{\partial K_{ij}}{\partial \theta_r}\left[\frac{\partial c_j}{\partial z_j} + \frac{\lambda\alpha_j}{2}\right] + \mu\theta_r.$$

Assuming the radial basis function kernel for automatic relevance determination, the partial derivatives of the kernel function with respect to a kernel parameter are

$$\frac{\partial K_{ij}}{\partial \theta_r} = -K_{ij}(x_{ir} - x_{jr})^2$$

Let $\boldsymbol{D}_r = \left[(\boldsymbol{x}_{ir} - \boldsymbol{x}_{jr})^2\right]_{i,j=1}^{\ell}$ and $\boldsymbol{\Delta} = [\partial c_i/\partial z_i]_{i=1}^{\ell}$ then the required partial derivatives can be expressed in matrix form as

$$\frac{\partial \mathcal{L}}{\partial \theta_r} = -\boldsymbol{\alpha}^T(\boldsymbol{K} \circ \boldsymbol{D}_r)\left(\boldsymbol{\Delta} + \frac{\lambda}{2}\boldsymbol{\alpha}\right) + \mu\theta_r,$$

where $\circ$ represents the Hadamard (element-wise) matrix product. It is then straightforward to optimise the kernel parameters via, for example, scaled conjugate gradient methods (as implemented by the `fminunc` routine of the MATLAB Optimisation Toolbox), where the solution of (3) is implicit in the evaluation of the cost function.

2.2. Model Selection

The regularisation parameters, λ and μ, are determined in model selection at the second level of inference. In this study, we simply minimise a k-fold cross-validation estimate of the loss function $c(\cdot, \cdot)$. A logarithmic transformation of the (strictly positive) regularisation parameters is used to obtain an unconstrained optimisation problem. While gradient descent optimisation of the model selection criterion would be feasible, we adopt the Nelder-Mead simplex procedure (as implemented by the `fminsearch` routine of the MATLAB Optimisation Toolbox) for ease of implementation. For the ARD kernel, the model selection criterion for the LSSVM and training criterion for kernel learning at the first level of inference is likely to exhibit multiple local minima. A sensible heuristic approach is to initialise these models with the kernel parameter obtained after model selection for an LSSVM model with an RBF kernel (which is a special case of the ARD kernel and less susceptible to local minima). This is computationally inexpensive and generally provides a good starting point for the optimisation procedure.

2.3. Theoretical Perspective

The algorithm presented in this paper is intended to be an essentially heuristic approach to the problem of over-fitting in model selection for kernel machines involving a non-trivial number of kernel parameters. However, it is nevertheless interesting to consider theoretical perspectives on the proposed method.

The training criterion (5) could be viewed as the negative logarithm of the Bayesian posterior distribution (neglecting additive constants), for the parameters of the model:

$$p(\boldsymbol{w}, b, \boldsymbol{\theta})|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w}, b, \boldsymbol{\theta}) \times p(\boldsymbol{w}, b, \boldsymbol{\theta}),$$

where

$$p(\mathcal{D}|\boldsymbol{w}, b, \boldsymbol{\theta}) \propto \prod_{i=1}^{\ell} \exp\left\{-c(y_i, f(\boldsymbol{x}_i))\right\},$$

and

$$p(\boldsymbol{w}, b, \boldsymbol{\theta}) = p(\boldsymbol{w}) \times p(\boldsymbol{\theta}) \propto \exp\left\{-\frac{\lambda}{2}\|\boldsymbol{w}\|^2\right\} \times \exp\left\{-\frac{\mu}{2}\|\boldsymbol{\theta}\|^2\right\}.$$

Minimisation of the regularised training criterion (5) then corresponds to maximum a-posteriori (MAP) estimation of the model within a Bayesian framework, where the regularisation terms represent prior beliefs regarding the two sets of model parameters $\boldsymbol{w}$ and $\boldsymbol{\theta}$. The model itself bears resemblance to several earlier learning systems, in particular multi-layer perceptron neural networks and radial basis function neural networks. Like an MLP neural network, the parameters of both layers of the model are tuned by optimisation of a regularised loss function. Like the RBF neural network, the output layer parameters are determined by solving a system of linear equations. However, like other kernel machines, the basis functions need not be radial basis functions, and can be used to establish known invariances, or incorporate expert domain knowledge (e.g. Shawe-Taylor and Cristianini, 2004). All of these learning systems have Bayesian interpretations (MacKay, 1994; Neal, 1996; Barber and Schottky, 1998; Van Gestel et al., 2002); a fully Bayesian treatment of the proposed method would be an interesting direction for further work.

For the conventional Support Vector Machine (Cortes and Vapnik, 1995), there are two types of hyper-parameters (Cherkassky and Mulier, 2007):

1. hyper-parameters controlling the size of the margin.
2. hyper-parameters controlling the complexity of the kernel.

There is substantial interaction between the two types of hyper-parameters, which means that neither can be tuned independently (e.g. Cherkassky and Mulier, 2007; Murphy, 2012). There is, however, a conceptual distinction between the two types of parameters: For a fixed kernel, the hyper-parameters of the first type, controlling the size of the margin (or equivalently regularisation parameters for other kernel machines), implement the structural risk minimisation (SRM) principle, defining a nested set of hypothesis classes of increasing complexity. The aim is to find a model that minimises the empirical error, whilst at the same time limiting the complexity of the hypothesis class from which the model is drawn. This is achieved by determining the optimal value of the hyper-parameters of the first type. However, this is only valid for a fixed kernel, if we tune the kernel function to suit the particular sample of data, via adaption of the hyper-parameters of the second kind, margin-based bounds on generalisation are no longer fully valid. This means that the structural risk minimisation provided by regularisation can be circumvented by tuning of the kernel-parameters.

Kernel learning at the first level of inference addresses this problem by also bringing the tuning of the kernel parameters within the framework of structural risk minimisation. The regularised loss function (5) could also viewed as the Lagrangian for the constrained optimisation problem:

$$\begin{aligned} \underset{\boldsymbol{w},b,\boldsymbol{\theta}}{\text{minimize}} \quad & \sum_{i=1}^{\ell} c\left(y_i, f(\boldsymbol{x}_i)\right), \\ \text{subject to} \quad & \|\boldsymbol{w}\|^2 < L, \\ & \|\boldsymbol{\theta}\|^2 < M. \end{aligned}$$

For a fixed value of M, L defines a nested set of models of increasing complexity as L increases. Similarly, for a fixed value of L, M defines a nested set of models of increasing complexity with increasing values of M. Thus the proposed learning system embodies the Structural Risk Minimisation (SRM) principle, as the aim is to find a model that minimises the empirical error, whilst at the same time limiting the complexity of the hypothesis class from which the model is drawn, with respect to both the margin of the linear model in the feature space and the complexity of the kernel. Note that, like kernel learning at the first level of inference, the regularisation parameter of the conventional support vector machine (which controls the complexity of the hypothesis class) is selected via cross-validation (e.g. Chang and Lin, 2011).

3. Results

In this section, we provide an empirical evaluation of kernel learning at the first level of inference, using the conventional Least-Squares Support Vector Machine classifier as the baseline. We begin with an illustrative example that shows the model selection problem for kernel learning at the first level of inference is relatively straightforward. The performance of the proposed learning system is then compared against the baseline method, using both RBF and ARD kernels, over a suite of fourteen benchmarks datasets. An analysis of the kernel parameters demonstrates that the proposed approach reduces the variance of parameter estimates, which explains the improvement in performance using the ARD kernel. The section concludes by assessing whether the proposed method fully solves the problem of over-fitting in model selection and puts the work into context via comparisons with previous work and multiple kernel learning.

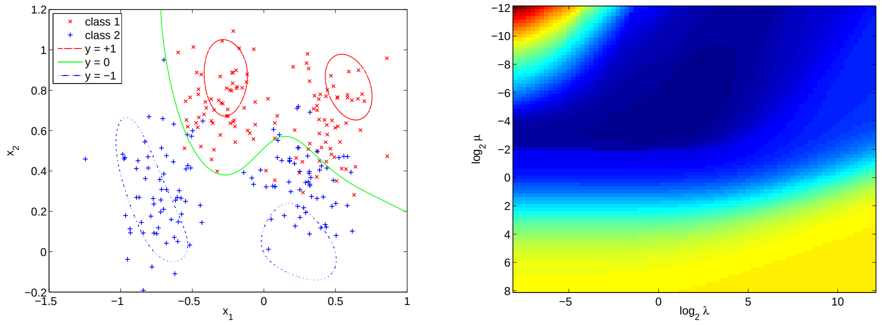
Table 1 shows the synthetic and real-world binary classification benchmark datasets used in the empirical evaluation of the proposed learning system. These comprise the binary classification datasets used in a previous study by Rätsch et al. (2001), and many others, augmented by Ripley's `synthetic` (Ripley, 1996) benchmark. For each dataset there are 100 random partitions of the data to form training and test sets (20 in the case of the larger `image` and `splice` benchmarks). The attributes are standardised, as in previous studies using this suite of benchmarks (e.g. Rätsch et al., 2001), for compatibility with the spherical RBF kernel (e.g. Murphy, 2012, section 14.5.3). Model selection is performed independently for each replicate, as described in section 2.2, in order to avoid biased performance evaluation due to over-fitting that occurs in model selection (Cawley and Talbot, 2010).

3.1. An Illustrative Example

Figure 1(a) shows the output of a least-squares support vector machine classifier for Ripley's synthetic benchmark (Ripley, 1996) with kernel learning at the first level of inference, using the automatic relevance determination variant of the radial basis function kernel (2). Clearly, the regularisation terms over the dual model parameters and over the kernel parameters are effective in avoiding over-fitting, and a good model is obtained. Figure 1(b) shows the five-fold cross-validation estimate of the test sum of squared error, used as the model selection criterion, as a function of the regularisation parameters, λ and μ. While the model selection criterion is non-convex, it is smooth and unimodal, which suggests that the numerical optimisation problem involved in model selection is likely to be relatively straightforward.

Table 1: Details of datasets used in empirical comparison.

Dataset	**Training Patterns**	**Testing Patterns**	**Number of Replicates**	**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
Synthetic	250	1000	100	2
Thyroid	140	75	100	5
Titanic	150	2051	100	3
Twonorm	400	7000	100	20
Waveform	400	4600	100	21



(a) (b)

Figure 1: (a) Least-squares support vector machine classifier for Ripley's synthetic benchmark (Ripley, 1996) with kernel learning at the first level of inference. (b) Five-fold cross-validation estimate of the sum of squared errors as a function of the two regularisation parameters, λ and μ.

3.2. Analysis of Results using the RBF Kernel

We begin by evaluating the performance of kernel learning at the first level of inference using a simple spherical RBF kernel, where over-fitting in model selection is unlikely to be a substantial problem as there is only one kernel parameter. Table 2 shows the mean test error rates and their standard errors for the conventional least-squares support vector machine classifier with leave-one-out (RBF-LOO-LSSVM) and five-fold cross-validation (RBF-XVAL-LSSVM) based model selection, and LSSVM with kernel learning at the first level of inference (RBF-FLKL-LSSVM, c.f. Equation (5)). In each case, a simple radial basis function kernel (1), with a single scale parameter, is used. The results for RBF-LOO-LSSVM are representative of current best practice for this family of kernel machines, the results for RBF-XVAL-LSSVM are included to illustrate the difference in performance that might potentially be explained by the difference in cross-validation based model selection criterion. Conventional LSSVM with leave-one-out cross-validation based model selection achieves the highest average rank. The RBF kernel has only a single kernel parameter, so the RBF-FLKL-LSSVM approach is unable to reduce the number of hyper-parameters to be tuned in model selection, and so it is unsurprising that it does not outperform the conventional approach in this case. In fact the RBF-FLKL-LSSVM performs slightly worse, although the difference is not statistically significant (see below), which *suggests* that the regularisation of the kernel parameters *may* not be beneficial, at least for the RBF kernel.

Following the recommendation of Demšar (2006), we use Friedman's test (Friedman, 1937, 1940) to determine whether there is a statistically significant difference in the performance of the three classifiers. The null hypothesis assumes that all k algorithms are equivalent, and so their average ranks over all N benchmarks, R_i, should be equal. The statistic,

$$\chi_F^2 = \frac{12N}{(k+1)}\left[\sum_{i=1}^{k} R_i^2 - \frac{k(k+1)^2}{4}\right],$$

is then distributed according to χ^2 with $k-1$ degrees of freedom. A less conservative test (Iman and Davenport, 1980) adopts the statistic

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2},$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. If the value of this statistic exceeds the appropriate tabulated critical value, the null hypothesis is rejected. In this case, the Nemenyi post-hoc test (Nemenyi, 1963) is used to test the significance of pairwise differences in average ranks. The performance of two classifiers is considered significantly different, at the $p = 0.1$ level, if their average ranks differ by at least the critical difference (CD),

$$CD = 2.052\sqrt{\frac{k(k+1)}{6N}}.$$

Table 2: Error rates of conventional least-squares support vector machine, with the spherical RBF kernel Equation (1) and leave-one-out (RBF-LOO-LSSVM) and 5-fold cross-validation (RBF-XVAL-LSSVM) -based model selection, and LSSVM with first-level kernel learning (RBF-FLKL-LSSVM). Error rates are given for fourteen benchmark datasets (Rätsch et al., 2001; Ripley, 1996), presented in the form of the mean error rate over test data for 100 realisations of each dataset (20 in the case of the `image` and `splice` benchmarks), along with their associated standard errors. The best result for each benchmark is shown in bold, results that are statistically indistinguishable from the best, according to the Wilcoxon signed ranks test (Wilcoxon, 1945) ($\alpha = 0.95$), are shown in italics.

Dataset	**Radial Basis Function**		
	RBF-LOO-LSSVM	**RBF-XVAL-LSSVM**	**RBF-FLKL-LSSVM**
banana	*10.605 ± 0.052*	**10.587 ± 0.051**	10.614 ± 0.053
breast cancer	*27.000 ± 0.476*	*26.753 ± 0.470*	**26.610 ± 0.506**
diabetis	**23.320 ± 0.166**	*23.390 ± 0.170*	23.933 ± 0.189
flare solar	**34.230 ± 0.168**	*34.325 ± 0.176*	*34.295 ± 0.182*
german	**23.543 ± 0.217**	*23.620 ± 0.240*	24.987 ± 0.253
heart	**16.550 ± 0.354**	*16.590 ± 0.366*	17.140 ± 0.337
image	*2.995 ± 0.159*	**2.861 ± 0.172**	*2.960 ± 0.166*
ringnorm	*1.610 ± 0.015*	1.650 ± 0.015	**1.608 ± 0.011**
splice	*10.828 ± 0.138*	**10.811 ± 0.148**	*10.926 ± 0.150*
synthetic	**9.642 ± 0.059**	*9.657 ± 0.060*	*9.658 ± 0.057*
thyroid	*4.707 ± 0.229*	**4.627 ± 0.233**	*4.720 ± 0.235*
titanic	*22.581 ± 0.102*	*22.568 ± 0.095*	**22.531 ± 0.089**
twonorm	2.845 ± 0.021	2.802 ± 0.019	**2.578 ± 0.015**
waveform	**9.786 ± 0.045**	*9.805 ± 0.044*	9.970 ± 0.044
Mean Rank	**1.8571**	*1.8571*	*2.2857*

In this case, there is insufficient evidence for the null hypothesis to be rejected, and so all three classifiers demonstrate statistically equivalent performance, as illustrated by Figure 2.

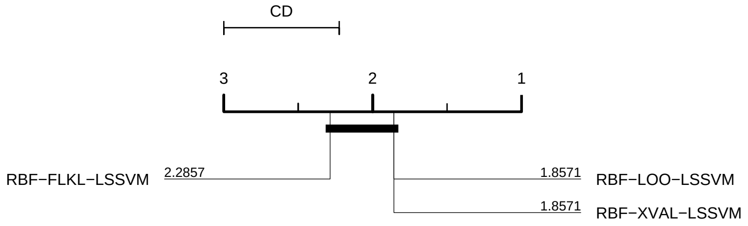


Figure 2: Critical difference diagram depicting the mean ranks of the classifiers using the spherical RBF kernel. Groups of classifiers that are not significantly different (at $p = 0.1$) are shown connected by a heavy bar.

3.3. Analysis of Results using the ARD Kernel

We now move on to results obtained using an automatic relevance determination kernel, where there are many more kernel parameters, and hence over-fitting in model selection is a much greater risk for conventional approaches. In order to ensure that poor performance of a model with an ARD kernel could not be attributed to local minima in the model selection criterion, the regularisation and kernel parameters were initialised using the corresponding model based on the RBF kernel (which is a special case of the ARD kernel). The final value of the model selection criterion for the ARD kernel can not then be higher than that of the corresponding model based on the RBF kernel, so if the generalisation performance of the model with the ARD kernel is worse than that of the RBF based model it can only be due to over-fitting the model selection criterion. Table 3 shows the corresponding error rates for classifiers based on the automatic relevance determination (ARD) kernel. This illustrates the susceptibility of each approach to over-fitting of the model selection criterion, due to the increase in the degrees of freedom introduced by the additional kernel parameters. The ARD-FLKL-LSSVM achieves the highest average rank. In this case, the Friedman test shows that there is a statistically significant difference in the rankings of the classifiers (i.e. the null hypothesis is rejected), and the Nemenyi post-hoc tests reveal that the ARD-FLKL-LSSVM is statistically superior to the ARD-LOO-LSSVM and ARD-XVAL-LSSVM models, as shown in Figure 3. This unequivocally demonstrates that first level kernel learning is able to successfully address the problem of over-fitting in model selection.

3.4. Analysis of Kernel Parameters

Figure 4 (a) and (b) show box plots of the kernel parameter for RBF-XVAL-LSSVM and RBF-FLKL-LSSVM respectively, for the simple RBF kernel function (1). It is immediately apparent that the kernel parameter values for RBF-

Table 3: Error rates of conventional least-squares support vector machine, with the eliptical ARD kernel Equation (2) and leave-one-out (RBF-LOO-LSSVM) and 5-fold cross-validation (RBF-XVAL-LSSVM) -based model selection, and LSSVM with first-level kernel learning (RBF-FLKL-LSSVM). Error rates are given for fourteen benchmark datasets (Rätsch et al., 2001; Ripley, 1996), presented in the form of the mean error rate over test data for 100 realisations of each dataset (20 in the case of the `image` and `splice` benchmarks), along with their associated standard errors. The best result for each benchmark is shown in bold, results that are statistically indistinguishable from the best, according to the Wilcoxon signed ranks test (Wilcoxon, 1945) ($\alpha = 0.95$), are shown in italics.

Dataset	**Automatic Relevance Determination**		
	ARD-LOO-LSSVM	**ARD-XVAL-LSSVM**	**ARD-FLKL-LSSVM**
banana	10.670 ± 0.058	*10.656 ± 0.059*	**10.600 ± 0.054**
breast cancer	29.208 ± 0.479	28.377 ± 0.421	**26.584 ± 0.454**
diabetis	23.933 ± 0.202	24.277 ± 0.199	**23.560 ± 0.170**
flare solar	**33.953 ± 0.188**	*34.237 ± 0.181*	*34.160 ± 0.160*
german	*24.787 ± 0.244*	*24.777 ± 0.246*	**24.523 ± 0.234**
heart	20.540 ± 0.443	20.230 ± 0.425	**17.650 ± 0.338**
image	**2.129 ± 0.145**	*2.158 ± 0.129*	*2.198 ± 0.096*
ringnorm	2.062 ± 0.038	2.056 ± 0.027	**1.942 ± 0.021**
splice	4.887 ± 0.105	4.621 ± 0.108	**4.306 ± 0.094**
synthetic	*9.743 ± 0.065*	*9.741 ± 0.063*	**9.685 ± 0.060**
thyroid	*4.907 ± 0.203*	*4.760 ± 0.213*	**4.640 ± 0.209**
titanic	*22.574 ± 0.115*	*22.551 ± 0.104*	**22.452 ± 0.089**
twonorm	4.477 ± 0.062	4.353 ± 0.061	**3.027 ± 0.024**
waveform	12.674 ± 0.120	12.569 ± 0.109	**10.516 ± 0.052**
Mean Rank	2.6429	2.1429	**1.2143**

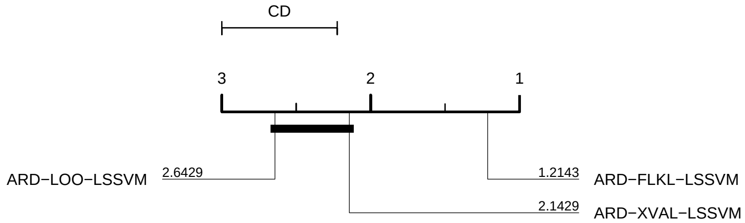
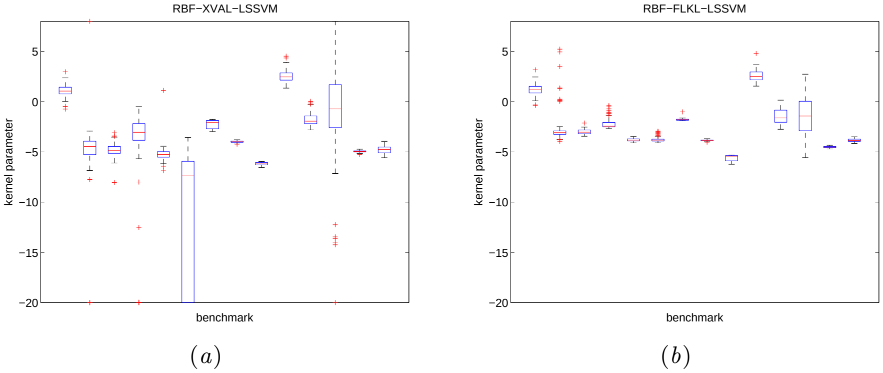


Figure 3: Critical difference diagram depicting the mean ranks of the classifiers using the ARD kernel. Groups of classifiers that are not significantly different (at $p = 0.1$) are shown connected by a heavy bar.



(a) (b)

Figure 4: Box plots of kernel parameter values for least-squares support vector machine classifiers with (a) conventional cross-validation based model selection (RBF-XVAL-LSSVM) and (b) kernel learning at the first level of inference (RBF-FLKL-LSSVM), using the spherical RBF kernel.

FLKL-LSSVM generally show less variability than those of the RBF-XVAL-LSSVM classifier, although this does not appear to have any significant effect on the generalisation performances of the two classifiers. This appears to be due to the regularisation of the kernel parameters in RBF-FLKL-LSSVM, which reduces the variance of the parameter estimates. Figure 5 shows box-plots of the kernel parameters for ARD-XVAL-LSSVM and ARD-FLKL-LSSVM for three illustrative benchmark datasets, `twonorm`, `thyroid` and `splice`.

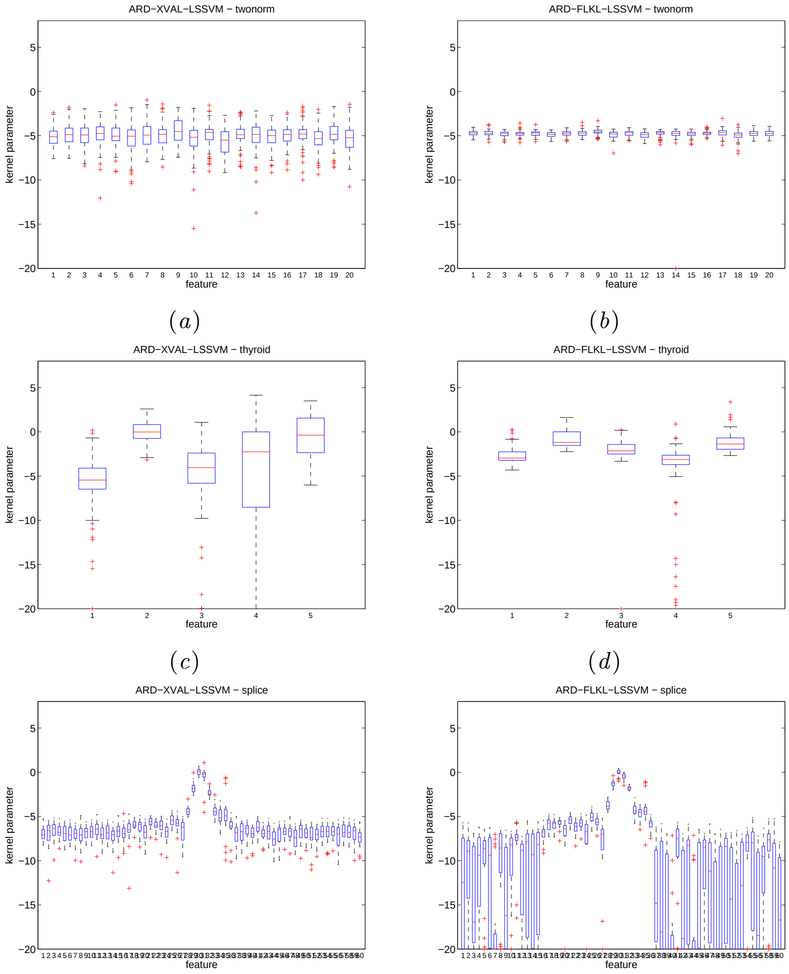
The `twonorm` dataset consists of patterns drawn from two Gaussian distributions representing each class in a twenty dimensional feature space. For this problem, all attributes are of equal importance and have the same characteristic scale, so a classifier with the ARD kernel function should have no advantage over a similar classifier with the basic RBF kernel. However conventional LSSVM with the ARD kernel (ARD-XVAL-LSSVM) exhibits a significantly higher error on this benchmark (4.353% ± 0.061) than the equivalent classifier with the RBF kernel (RBF-XVAL-LSSVM, 2.802% ± 0.019). The reason for this poor performance is apparent in Figure 5(a); while the kernel parameters have similar values for each attribute, as expected, there is a considerable variability in the value of the kernel parameters for different partitions of the data to form training and test sets. This high variability suggests substantial over-fitting of the model selection criterion, which explains the poor generalisation performance of ARD-XVAL-LSSVM on this benchmark. The corresponding kernel parameters for ARD-FLKL-LSSVM, shown in Figure 5(b), display substantially less variability, suggesting that regularisation is effective in avoiding over-fitting in learning the kernel.

The `thyroid` benchmark is representative of datasets where all attributes are useful, but some may be more discriminative than others, or where different attributes benefit from different scalings. A box plot of the kernel parameters for ARD-XVAL-LSSVM are shown in Figure 5(c), and again demonstrate considerable variability, perhaps explaining the poor performance of this model, which achieves a mean error rate of $4.760\% \pm 0.213$. This is slightly greater than the error for the corresponding model with the basic RBF kernel (RBF-XVAL-LSSVM, $4.627\% \pm 0.233$). The variability of the kernel parameters of the ARD-FLKL-LSSVM model, shown in Figure 5(d), display far less variability, and as might be expected a lower error rate of $4.640\% \pm 0.209$, which improves on that of the conventional ARD-XVAL-LSSVM model. We conclude from this that over-fitting the model selection criterion prevents the ARD-XVAL-LSSVM classifier from benefiting from individual scaling of the attributes, but less so the ARD-FLKL-LSSVM classifier as a result of learning the kernel parameters at the first level of inference with regularisation, and simplified model selection requiring only the values of two regularisation parameters to be tuned. Even in the absense of a clear gain in generalisation performance, the ARD-FLKL-LSSVM classifier obtains more reliable estimates of the feature scaling parameters than ARD-XVAL-LSSVM, and so is more useful in helping to understand the data.

The `splice` benchmark is representative of datasets where some of the attributes may be uninformative. Box plots of the kernel parameters for ARD-FLKL-LSSVM and conventional ARD-XVAL-LSSVM are shown in Figure 5(e) and (f). Again the kernel parameters of the informative attributes for the ARD-FLKL-LSSVM classifier exhibit less variability than those of the ARD-XVAL-LSSVM classifier, explaining some of the difference in the mean error rates of the two approaches ($4.306\% \pm 0.094$ and $4.621\% \pm 0.108$ respectively). Note, however, that the ARD-FLKL-LSSVM classifier also surpresses the less informative features much more strongly than the ARD-XVAL-LSSVM classifier, due to the regularisation term penalising large values of the kernel parameters. This implies that the ARD-FLKL-LSSVM classifier is potentially more able to explicitly identify uninformative features than the conventional approach, which may be beneficial, even in the absence of a clear gain in generalisation performance.

3.5. Does First-Level Kernel Learning Solve the Problem?

Figure 6 shows a critical difference diagram for kernel ridge regression with conventional five-fold cross-validation based model selection (RBF-XVAL-LSSVM and ARD-XVAL-LSSVM) and with kernel learning at the first level of inference



(*a*) (*b*)

(*c*) (*d*)

(*e*) (*f*)

Figure 5: Box plots of kernel parameter values for least-squares support vector machine classifiers with the eliptical ARD kernel and (a,c,e) conventional cross-validation based model selection (ARD-XVAL-LSSVM) and (b,d,f) kernel learning at the first level of inference (ARD-FLKL-LSSVM), for three benchmark datasets (a,b) `twonorm`, (c,d) `thyroid` and (e,f) `splice`.

(RBF-FLKL-KRR and ARD-FLKL-LSSVM) with radial basis function and automatic relevance determination kernels. Clearly, first-level kernel learning ameliorates the problem of over-fitting in model selection for the ARD kernel, at least to the extent that the results obtained for the ARD kernel are no longer statistically inferior to those obtained using the RBF kernel. However, as the RBF kernel is a special case of the ARD kernel, if the problem of estimating the optimal values for the kernel parameters had been fully solved, one might expect the results obtained using the ARD kernel to be at least as good, if not superior to those obtained using the RBF kernel. This suggests that kernel learning at the first level of inference, at least in its current form, probably has not fully solved the problem of over-fitting in model selection, but does represent a substantial step in the right direction. However, as ARD-FLKL-LSSVM provides more reliable kernel parameter estimates than ARD-XVAL-LSSVM, and is more able to explicitly identify uninformative features, it would remain a potentially useful algorithm for explaining the data and reducing operational costs, even in the absence of gains in generalisation performance.

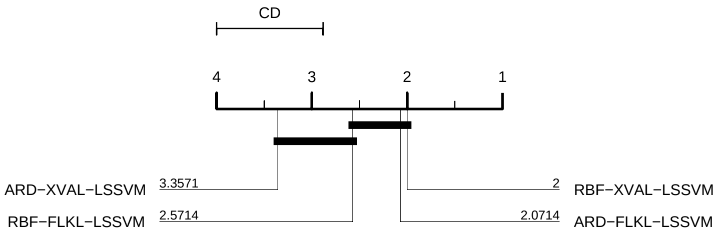


Figure 6: Critical difference diagram depicting the mean ranks of classifiers using the RBF and ARD kernels. Groups of classifiers that are not significantly different (at $p = 0.1$) are shown connected by a heavy bar.

3.6. Comparison with Earlier Work

In previous work (Cawley and Talbot, 2007), we investigated the regularisation of a leave-one-out cross-validation based model selection criterion, penalising kernel parameters with large values, where the additional regularisation parameter is integrated out analytically using an uninformative Jeffrey's prior (Buntine and Weigend, 1991). Figure 7 shows a critical difference diagram depicting the mean ranks of least-squares support vector machine classifiers with Bayesian regularisation of the model selection criterion (ARD-BR-LSSVM) (Cawley and Talbot, 2007), LSSVM with first level kernel learning (ARD-FLKL-LSSVM), and with conventional five-fold cross-validation based model selection (ARD-XVAL-LSSVM) as a baseline for comparison; the automatic relevance determination (ARD) kernel is used in each case. The mean rankings are computed over the thirteen of the fourteen benchmark datasets common to both studies (i.e. excluding the `synthetic` benchmark). It can be seen that both ARD-FLKL-LSSVM and ARD-BR-LSSVM are effective in avoiding the over-fitting in model selection that can occur using the ARD kernel. However, while the difference in mean ranking between these two methods is not statistically significant, ARD-FLKL-LSSVM *is* statistically superior to the conventional approach (ARD-XVAL-LSSVM), while ARD-BR-LSSVM is not. We therefore recommend ARD-FLKL-LSSVM for practical applications using the ARD kernel.

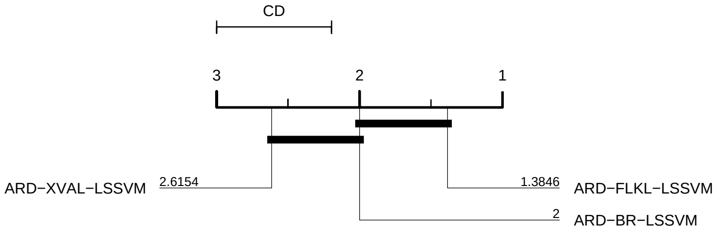


Figure 7: Critical difference diagram depicting the mean ranks of the ARD-XVAL-LSSVM, ARD-BR-LSSVM and ARD-FLKL-LSSVM classifiers using the ARD kernel. Groups of classifiers that are not significantly different (at $p = 0.1$) are shown connected by a heavy bar.

3.7. Comparison with Multiple Kernel Learning

Multiple Kernel Learning (MKL) is a recent approach to defining the kernel that has attracted considerable interest (Gönen and Alpaydin, 2011, and references therein). Rather than optimise the kernel parameters directly, the majority of multiple kernel learning algorithms seek to define the kernel as a non-negatively weighted sum of a set of pre-defined candidate kernel functions, each with a fixed set of parameters, $\boldsymbol{\theta}_i$, i.e.

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\eta}) = \sum_{i=1}^{k} \eta_i \mathcal{K}_i(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}_i) \quad \text{subject to} \quad \eta_i \geq 0, \quad \forall i \in \{1, 2, \ldots, k\}.$$

Gehler and Nowozin (2008) extend multiple kernel learning to the case where the candidate kernel functions may also have tunable hyper-parameters, resulting in the Infinite Kernel Learning (IKL) procedure; they also evaluate the performance of SimpleMKL (Rakotomamonjy et al., 2008) and IKL algorithms on the suite of thirteen benchmark datasets introduced by Rätsch et al. (2001). Three candidate kernel functions are investigated, the first is the *single* kernel (equivalent to the RBF kernel),

$$\mathcal{K}_{\text{single}}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \exp\left\{-\theta^2 \sum_{i=1}^{d} ([\boldsymbol{x}]_i - [\boldsymbol{x}']_i)^2\right\},$$

where $[\boldsymbol{x}]_i$ represents the i^{th} element of the vector $\boldsymbol{x}$. The second is the *separate* kernel, where the candidates consist of the *single* kernel, with additional univariate kernels of the form

$$\mathcal{K}_{\text{separate}}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \exp\left\{-\theta_i^2 ([\boldsymbol{x}]_i - [\boldsymbol{x}']_i)^2\right\},$$

which provides a limited ability to model the data using different characteristic scales for each attribute. The final candidate kernel, corresponding to the ARD kernel, is the *products* kernel,

$$\mathcal{K}_{\text{products}}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{d} -\theta_i^2([\boldsymbol{x}]_i - [\boldsymbol{x}']_i)^2\right\},$$

which provides independent scaling parameters for each attribute. IKL with the products kernel thus provides the multiple kernel learning algorithm providing the most direct competitor for kernel learning at the first level of inference.

A comparison of results obtained by MKL, IKL, LSSVM with cross-validation based model selection and kernel learning at the first level of inference is given in Table 4. It should be noted that the results for RBF-XVAL-LSSVM and ARD-FLKL-LSSVM were obtained using an unbiased procedure, where model selection is performed independently in each fold, but the results for MKL and IKL were obtained using the *median* protocol in which the hyper-parameters were optimised individually in the first five folds, and the median of those values used throughout the performance evaluation. The *median* protocol has been demonstrated to produce optimistically biased performance estimates (Cawley and Talbot, 2010); however, as only a single regularisation parameter is tuned in the cases of MKL and IKL, it is likely that this bias will be relatively small. Clearly the IKL model performs very well on the `image` and `splice` datasets, but very poorly on others, e.g. `twonorm`, `heart` and `waveform`. As a result, neither MKL nor IKL out-perform ARD-FLKL-LSSVM overall, as illustrated by the critical difference diagram, shown in Figure 8. The improved performance of ILK on the `image` and `splice` benchmarks is however obtained at the cost of a substantial increase in computational expense in operation and in a reduction in interpretability, due to the large number of candidate kernels retained by the model (a mean of 27.1 for `image` and 72.8 for `splice`). Gehler and Nowozin (2008) conclude that the practitioner should choose between two models, the SVM (or equivalently LSSVM) or the IKL algorithm, because the enlarged kernel class might lead to significant performance increases for some datasets. We would suggest that the ARD-FLKL-LSSVM algorithm should also be considered as it also provides substantial performance increases on some datasets, but with much lower computational expense and higher interpretability than IKL.

4. Discussion

A number of studies involving automatic relevance determination have noted that the optimisation of a large number of hyper-parameters is likely to degrade performance (e.g. Bengio, 2000; Bo et al., 2006; Keerthi et al., 2006). Cawley and Talbot (2007) demonstrate that regularisation of the kernel parameters is effective in preventing over-fitting. In this study, we show that the kernel parameters can be regarded as parameters rather than hyper-parameters, and optimised during training rather than model selection, and that again regularisation is effective in avoiding over-fitting. The method proposed in this paper is

Table 4: Error rates of the conventional least-squares support vector machine, with 5-fold cross-validation based model selection (RBF-XVAL-KRR), kernel ridge regression with first-level kernel learning (ARD-FLKL-KRR), multiple kernel learning (MKL) and infinite kernel learning (IKL), see text for details. For MKL and IKL, the mean number of canidate kernels used is given as $\#k$. Error rates are given for thirteen benchmark datasets (Rätsch et al., 2001); the results for each method are presented in the form of the mean error rate over test data for 100 realisations of each dataset (20 in the case of the `image` and `splice` benchmarks) and their associated standard deviations. The best result for each benchmark is shown in bold. The results for MKL and IKL are taken from Gehler and Nowozin (2008), which also gives details of the experimental method used.

Dataset	**RBF-XVAL**	**ARD-FLKL**	**MKL separate**		**IKL products**	
	-LSSVM	**-LSSVM**	**err**	**#k**	**err**	**#k**
banana	10.6 ± 0.5	10.6 ± 0.5	**10.5 ± 0.5**	1.0	10.7 ± 0.5	3.7
breast cancer	26.8 ± 4.7	26.6 ± 4.5	26.7 ± 4.2	4.5	**25.7 ± 4.1**	16.1
diabetis	**23.4 ± 1.7**	23.6 ± 1.7	24.5 ± 1.6	4.0	24.3 ± 1.8	22.3
flare solar	34.3 ± 1.8	34.2 ± 1.6	34.3 ± 2.1	2.9	**32.8 ± 1.9**	2.6
german	**23.6 ± 2.4**	24.5 ± 2.3	25.1 ± 2.2	8.3	24.6 ± 2.4	46.1
heart	**16.6 ± 3.7**	17.6 ± 3.4	16.7 ± 4.1	9.0	20.1 ± 3.6	28.2
image	2.9 ± 0.8	2.2 ± 0.4	3.0 ± 0.6	1.6	**1.4 ± 0.3**	27.1
ringnorm	**1.6 ± 0.2**	1.9 ± 0.2	1.7 ± 0.1	2.6	2.1 ± 0.2	16.3
splice	10.8 ± 0.7	4.3 ± 0.4	6.0 ± 0.4	24.1	**3.1 ± 0.3**	72.8
thyroid	4.6 ± 2.3	4.6 ± 2.1	4.7 ± 2.1	1.0	**4.1 ± 2.0**	12.7
titanic	22.6 ± 0.9	22.5 ± 0.9	**22.4 ± 1.0**	1.9	**22.4 ± 1.1**	5.2
twonorm	2.8 ± 0.2	3.0 ± 0.2	**2.5 ± 0.1**	3.8	3.8 ± 0.4	36.2
waveform	**9.8 ± 0.4**	10.5 ± 0.5	10.2 ± 0.4	9.7	11.4 ± 0.6	33.7
Mean Rank	**2.3462**	2.4615	2.5000		2.6923	

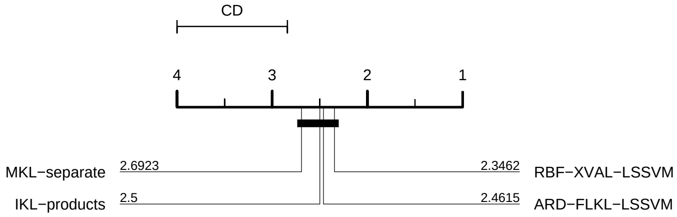


Figure 8: Critical difference diagram depicting the mean ranks of the RBF-XVAL-LSSVM, ARD-FLKL-LSSVM, multiple kernel learning (separate kernel) and infinite kernel learning (products kernel). Groups of classifiers that are not significantly different (at $p = 0.1$) are shown connected by a heavy bar.

not statistically superior to that presented in our previous work; however, unlike our previous method, kernel learning at the first level of inference *is* statistically superior to the conventional approach of learning the kernel parameters using an unregularised cross-validation based model selection criterion. For this reason, we recommend the method proposed in this paper for practical applications when using ARD kernels.

The proposed method is not just applicable to the ARD variant of the radial basis function kernel, provided a regularisation term can be constructed that can be expected to favour more simple models. For simple standard kernel functions, such as the polynomial kernel,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}' + c)^d, \qquad c > 0, \tag{6}$$

FLKL-LSSVM is not recommended as there are too few kernel parameters to justify the exchanging kernel parameters for an additional regularisation parameter to be tuned in model selection. A better approach might be to add a regularisation term to the model selection criterion, as suggested by Cawley and Talbot (2007). In the case of the inhomogeneous polynomial (6), the feature space consists of all monomials of degree up to d, where the kernel parameter c influences the relative scaling of monomials of different degrees. A regularisation term that favoured monomials of low degree, by penalising small values of c might result in improved generalisation. ARD can also be used to extend other kernels, such as the inhomogeneous polynomial kernel, by pre-scaling the attributes, i.e.

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}\boldsymbol{\Sigma}\boldsymbol{x}' + c)^d, \qquad c > 0,$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\eta})$ is a diagonal matrix of attribute scaling parameters, in which case FLKL-LSSVM would again be attractive due to the large number of kernel parameters. FLKL-LSSVM might also be useful in computational biology where the kernel, such as the sequence alignment kernel (Gordon et al., 2003), depends on a matrix of pairwise substitution penalties between nucleotides or even amino acids. The large number of kernel parameters involved would mean that over-fitting the model selection criterion would be highly likely using the conventional approach and FLKL-LSSVM may result in improved performance. In this case the regularisation term should reward substitution matrices with simple structure.

It is interesting to consider whether it is better to regard learning the kernel as being part of fitting the model or as part of model selection. From an optimisation perspective, there is little difference between these two approaches. The coefficients of the kernel expansion are given by the solution of a convex sub-problem (3) for both the FLKL-LSSVM and conventional XVAL-LSSVM classifiers, and share an efficient implementation via Cholesky decomposition. Likewise, for both approaches, the optimisation problems for tuning the kernel and regularisation parameters are non-convex, and so both potentially suffer the problem of local minima. For both approaches, gradient descent and Nelder-Mead simplex methods are feasible for optimising the kernel and regularisation parameters, and both approaches have broadly similar computational expense. From a theoretical perspective, kernel machines are attractive because their interpretation as a linear model constructed in a fixed kernel-induced feature space provides mathematical tractability (Schölkopf and Smola, 2002). However, much of this mathematical tractability is essentially lost if the kernel is not fixed, but is tuned to better suit the data, and so there is also relatively little to choose between approaches from that perspective either.

The k-fold cross-validation procedure used to determine the regularisation parameters of the FLKL-LSSVM is relatively expensive, due to the necessity of re-training the model k times. An alternative would be to iteratively update the regularisation parameters as the model is trained, as under the evidence framework for artificial neural networks (MacKay, 1994). Again the Laplace approximation could be used to estimate the Bayesian evidence for the model, assuming a Gaussian posterior over both the model and kernel parameters. An approximate leave-one-out cross-validation performance estimate (Myles et al., 1997) would also provide a feasible criterion.

Lastly, where predictive performance is the primary concern, for some datasets the RBF kernel will out-perform the ARD kernel, whether the FLKL-LSSVM is used, or the conventional LS-SVM. For some, even the RBF kernel may not be necessary and the linear kernel,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}',$$

will give better performance. In this case, the choice of appropriate kernel cannot be made on the basis of the model selection criterion because of over-fitting of the model selection criterion (c.f. Cherkassky and Mulier, 2007; Cawley and Talbot, 2010). The RBF kernel is a special case of the ARD kernel, and so the ARD kernel will always be able to achieve a value of the model selection criterion that is at least as low as that for the RBF kernel, and very likely lower, due to the additional kernel parameters. As a result, the model selection criterion will always favour the ARD kernel over the RBF kernel. Similarly, as the RBF kernel has one kernel parameter, and the linear kernel none, even for linear tasks the RBF kernel may achieve a lower value of the model selection criterion as the additional degree of freedom introduces a greater opportunity to overfit the model selection criterion. Therefore, the choice of kernel should be based on an independent estimate of generalisation performance, such as an external cross-validation procedure.

5. Conclusions

Kernel learning is typically performed during model selection, commonly by numerical minimisation of the cross-validation error. In this paper, we have demonstrated that kernel learning at the first level of inference is also feasible, jointly optimising the model and kernel parameters using a single training criterion. A key benefit of this approach is that the risk of over-fitting the model selection criterion is greatly reduced, as only the values of two regularisation parameters are then tuned in model selection. As a result, for the automatic relevance determination kernel, the performance of the proposed method is shown to be statistically superior to that of the existing cross-validation based approach over a suite of 14 benchmark datasets, and also improves on our previous method (Cawley and Talbot, 2007). ARD-FLKL-LSSVM is also competitive with multiple kernel learning approaches, but with reduced computational expense and greater interpretability (as there is only a single kernel to evaluate). This is a significant practical advance as automatic relevance determination is useful in applications likely to have irrelevant features, or where identifying a small subset of useful features is of inherent interest.

Acknowledgements

We thank Tony Bagnall, Jon Hills, Isabelle Guyon the anonymous reviewers and the action editor for their helpful and constructive comments, that have substantially improved the paper. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/F010508/1 - Advancing Machine Learning Methodology for New Classes of Prediction Problems. The research presented in this paper was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

References

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 1999.

D. Barber and B. Schottky. Radial basis functions: a Bayesian treatment. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 402–408. Morgan Kauffmann Publishers, 1998.

P. L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):525–536, March 1998. doi: 10.1109/18.661502.

Y. Bengio. Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8):1889–1900, August 2000. doi: 10.1162/089976600300015187.

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(4): 961–978, April 2006. doi: 10.1162/neco.2006.18.4.961.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

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 IEEE/INNS International Joint Conference on Neural Networks (IJCNN-06)*, pages 1661–1668, Vancouver, BC, Canada, July 16–21 2006. doi: 10.1109/IJCNN.2006.246634.

G. C. Cawley. Causal & non-causal feature selection for ridge regression. In *Journal of Machine Learning Research: Workshop and Conference Proceedings*, volume 3: Causation and Prediction Challenge (WCCI-2008), pages 107–128, 2009.

G. C. Cawley. Baseline methods for active learning. In *Proceedings of the AISTATS-2010 Active Learning and Experimental Design Workshop*, volume 16 of *JMLR Workshop and Conference Proceedings*, pages 47–57, 2011.

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. doi: 10.1016/S0031-3203(03)00136-5.

G. C. Cawley and N. L. C. Talbot. Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research*, 8:841–861, April 2007.

G. C. Cawley and N. L. C. Talbot. Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71(2–3):243–264, June 2008. doi: 10.1007/s10994-008-5055-9.

G. C. Cawley and N. L. C. Talbot. Over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, July 2010.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems Technology*, 2(3):27:1–27:27, May 2011. doi: 10.1145/1961189.1961199.

O. Chapelle. *Support Vector Machines: Induction Principles, Adaptive Tuning and Prior Knowledge.* PhD thesis, Université Pierre et Marie Curie, Paris VI, 2002.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1–3):131–159, January 2002. doi: 10.1023/A:1012450327387.

V. Cherkassky and F. Mulier. *Learning from Data - Concepts, Theory and Methods.* Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley and Sons, 1998.

V. Cherkassky and F. Mulier. *Learning from Data - Concepts, Theory and Methods.* John Wiley and Sons, second edition, 2007.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, September 1995. doi: 10.1007/BF00994018.

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

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification.* John Wiley and Sons, second edition, 2001.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Society*, 32(200): 675–701, December 1937.

M. Friedman. A comparison of alternative tests if significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

O. L. Fung, G.and Mangasarian. Proximal support vector machine classifiers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 77–86, 2001.

P. Gehler and S. Nowozin. Infinite kernel learning. Technical Report TR-178, Max Planck Institute For Biological Cybernetics, 2008.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, January 1992. doi: 10.1162/neco.1992.4.1.1.

M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, July 2011.

L. Gordon, A. Y. Chervonenkis, A. J. Gammerman, I. A. Shahmuradov, and V. V. Solovyev. Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, 19(15):1964–1971, October 2003. doi: 10.1093/bioinformatics/btg265.

I. Guyon. A practical guide to model selection. In J. Marie, editor, *Machine Learning Summer School*. Springer, 2009.

I. Guyon, A. Saffari, G. Dror, and G. Cawley. Analysys of the IJCNN 2007 agnostic learning vs. prior knowledge challenge. *Neural Networks*, 21(2–3): 544–550, March–April 2008. doi: 10.1016/j.neunet.2007.12.024.

I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: Beyond the Bayesian/frequentist divide. *Journal of Machine Learning Research*, 11:61–87, 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer, 2001.

R. L. Iman and J. M. Davenport. Approximations of the critical region of the Friedman statistic. *Communications in Statistics*, 9(6):571–595, 1980. doi: 10.1080/03610928008827904.

S. S. Keerthi, V. Sindhwani, and O. Chapelle. An efficient method for gradient-based adaptation of hyperparameters in SVM models. In *Advances in Neural Information Processing Systems 19*. MIT Press, Vancouver, BC, Canada, 2006.

A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 950–957. Morgan Kauffmann Publishers, San Mateo, CA, 1991.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, May 1992. doi: 10.1162/neco.1992.4.3.415.

D. J. C. MacKay. Bayesian methods for backpropagation networks. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks*, volume 3, chapter 6, pages 211–254. Springer, 1994.

D. J. C. MacKay. Introduction to Gaussian processes. In C. M. Bishop, editor, *Neural networks and machine learning*. Springer Verlag, 1998.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48, Maddison, WI, USA, 21–25 August 1999. doi: 10.1109/NNSP.1999.788121.

K. P. Murphy. *Machine Learning - A Probabilistic Perspective.* MIT Press, 2012.

A. J. Myles, A. F. Murray, A. R. Wallace, J. Barnard, and G. Smith. Estimating MLP generalisation ability without a test set using fast, approximate leave-one-out cross-validation. *Neural Computating and Applications*, 5(3):134–151, September 1997. doi: 10.1007/BF01413859.

R. Neal. *Bayesian learning for neural networks*, volume 118 of *Lecture Notes in Statistics.* Springer, 1996.

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

P. B. Nemenyi. *Distribution-free multiple comparisons.* PhD thesis, Princeton University, 1963.

A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

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, March 2001. doi: 10.1023/A:1007618119488.

A. Rifkin, R. amd Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, January 2004.

R. Rifkin, G. Yeo, and T. Poggio. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.

R. M. Rifkin. *Everything old is new again: a fresh look at historical approaches in machine learning.* PhD thesis, Massachusetts Institute of Technology, 2002.

B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

M. Saerens, P. Latinne, and C. Decaestecker. Any reasonable cost function can be used for *a posteriori* probability approximation. *IEEE Transactions on Neural Networks*, 13(5):1204–1271, September 2002. doi: 10.1109/TNN.2002.1031952.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98)*, pages 515–521. Morgan Kaufmann, 1998.

B. Schölkopf and A. J. Smola. *Learning with kernels — support vector machines, regularization, optimization and beyond*. MIT Press, Cambridge, MA, 2002.

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

J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least squares support vector machine*. World Scientific Publishing Company, Singapore, 2002. ISBN 981-238-151-1.

T. Van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. Bayesian framework for least squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. *Neural Computation*, 14(5):1115–1147, 2002.

T. Van Gestel, J. A. K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, and J. Vandewalle. Benchmarking least squares support vector machine classifiers. 54(1):5–32, 2004.

V. N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications and control series. Wiley, 1998.

A. R. Webb and K. D. Copsey. *Statistical pattern recognition*. Wiley, third edition, 2011.

A. R. Webb and S. Shannon. Shape-adaptive radial basis functions. *IEEE Transactions on Neural Networks*, 9(6):1155–1166, November 1998. doi: 10.1109/72.728359.

H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464, Winter 1989. doi: 10.1162/neco.1989.1.4.425.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

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. doi: 10.1109/34.735807.

J. Xu, X. Zhang, and Y. Li. Kernel MSE algorithm: a unified framework for KFD, LS-SVM and KRR. In *Proceedings International Joint Conference on Neural Networks*, volume 2, pages 1486–1491, 2001.

Z. Zhang, G. Dai, C. Xu, and M. I. Jordan. Regularized discriminant analysis, ridge regression and beyond. *Journal of Machine Learning Research*, 11:2199–2228, August 2010.