Hierarchic Bayesian Models for Kernel Learning

Mark Girolami   GIROLAMI@DCS.GLA.AC.UK
Simon Rogers   SROGERS@DCS.GLA.AC.UK
Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, G12 8QQ, UK.

Abstract

The integration of diverse forms of informative data by learning an optimal combination of base kernels in classification or regression problems can provide enhanced performance when compared to that obtained from any single data source. We present a Bayesian hierarchical model which enables kernel learning and present effective variational Bayes estimators for regression and classification. Illustrative experiments demonstrate the utility of the proposed method. Matlab code replicating results reported is available at `http://www.dcs.gla.ac.uk/~srogers/kernel_comb.html`.

1. Introduction

Kernel methods have grown in prominence primarily due to their ability to represent nonlinear problems by models linear in their parameters given a specific kernel function (Shawe-Taylor & Cristianini, 2004). This linearity means that the model parameters, in for example the case of Support Vector Machines (SVM), can be identified by solving a convex optimisation problem conditioned on the chosen kernel function. The overall kernel function chosen is critical to the performance capability of the kernel based machine and in the case of kernels with a parametric form (such as Gaussian kernels) careful selection of the associated parameters defining the kernel is essential. Typically frequentist approaches such as leave-one-out cross-validation are employed in identifying the associated kernel parameters (Shawe-Taylor & Cristianini, 2004).

The Bayesian approaches to inducing kernel machines such as Gaussian Processes (GP)(MacKay, 2003) and

the Relevance Vector Machine (RVM) (Tipping, 2001) require the solution of linear regularised least-squares problems to obtain the posterior mean and covariance of the model parameters. However a further nonlinear optimisation to obtain point estimates of the kernel parameters is required in fully identifying the overall model. A very recent focus of research has been to explore more fully the important problem of *learning* the kernel from the data available and we now review the main work devoted to this problem.

2. Kernel Learning

Kernel learning can range from the estimation of the width parameters of an homogenous Gaussian kernel to obtaining the optimal linear combination of a set of candidate kernels. Employing a candidate set of kernels or kernel matrices provides an elegant way of integrating heterogenous data within a single kernel machine and the practical utility of this has been demonstrated in for example (Lanckriet et al., 2004).

One of the first publications to propose a kernel machine based on a composite kernel of the form $K(x,y) = \sum_j \beta_j K_j(x,y), \ \ \beta_j \geq 0$, was (Gunn & Kandola, 2002). Kernel alignment was proposed in (Cristianini et al., 2002) where it was demonstrated that a re-weighted combination of the rank-1 eigenvector based approximations of the kernel matrix could be obtained in a transductive setting, thus providing a potential means of learning the kernel matrix. More recently (Lanckriet et al., 2004) have provided a general framework for learning the kernel matrix based on semidefinite programming. In that work, composite kernels with a bounded trace are obtained by weighted combinations of positive-semidefinite candidate matrices subject to the constraint that the resulting composite kernel is positive-semidefinite. Restricting the class of possible composite matrices to such conic combinations and then employing SVM's requires the solution of quadratically constrained quadratic programmes (QCQP). More recently (Bach et al., 2004) have focused on providing faster algorithms to solve

the required QCQP within the SDP framework.

In (Ong et al., 2003) the composite kernel function, in an inductive setting, is learned using what is termed hyperkernels. By defining a hyper reproducing kernel Hilbert space and optimising a regularized cost function the composite kernel function can be obtained. Computationally efficient methods based on second-order cone programming are developed by (Tsang & Kwok, 2004) for hyperkernel learning. Other approaches to kernel learning which have been proposed include Boosting (Crammer et al., 2003), computation of the regularisation path (Bach et al., 2005), a gradient descent based algorithm (Bousquet & Herrmann, 2003), and in (Fung et al., 2004) the solution of a biconvex problem defines a multiple kernel based Fisher discriminant. Finally, in (Zhang et al., 2004) the kernel matrix, in a transductive setting and specifically for the classification problem, is learned using a Bayesian hierarchical model of the matrix which consists of an efficient formulation of the Tanner-Wong algorithm for data augmentation. Once the kernel has been learnt it is then employed within a standard classification algorithm, the parameters of which are separately estimated independently of the kernel learning procedure.

What has not been considered in the literature on kernel learning is the definition of a general probabilistic representation of a composite kernel based machine for regression or classification. In the following sections a Bayesian hierarchical model for composite kernel based regression and classification is presented.

3. Hierarchic Probabilistic Model

For data where the target and input samples are $\mathbf{t} = [t_1 \cdots t_N]^{\mathrm{T}}$ and $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_N]^{\mathrm{T}}$ we consider a kernel-based regression model which employs a composite kernel. Each target value t_n is represented by the model response y_n as

$$\sum_{m=0}^{N} \alpha_m K_{\boldsymbol{\beta}}(\mathbf{x}_m, \mathbf{x}_n) = \sum_{m=0}^{N} \alpha_m \sum_{k=1}^{K} \beta_k K_k(\mathbf{x}_m, \mathbf{x}_n) \quad (1)$$

Defining the $N \times (N + 1)$ composite kernel matrix, which includes a bias term, as $\mathbf{K}_{\boldsymbol{\beta}}$ and the $N \times K$ dimensional matrix $\mathbf{Z}_{\boldsymbol{\alpha}}$ whose elements are defined as $Z_{nk} = \sum_{m=0}^{N} \alpha_m K_k(\mathbf{x}_m, \mathbf{x}_n)$ then the $N \times 1$ vector of model responses follows as $\mathbf{y} = \mathbf{Z}_{\boldsymbol{\alpha}} \boldsymbol{\beta} = \mathbf{K}_{\boldsymbol{\beta}} \boldsymbol{\alpha}$ where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are $K \times 1$ and $(N + 1) \times 1$ vectors respectively.

The form of the priors to be placed on the model parameters now has to be considered. A naive assumption on the priors for $\boldsymbol{\beta}$ would be to consider a series of independent *sparsity inducing* hierarchic priors on each β_k. However, if there is no natural constraint on the

norm of $\boldsymbol{\beta}$ then unconstrained growth or reduction in the norm of $\boldsymbol{\alpha}$ will follow due to the inherent coupling of the two coefficient sets in (1). In (Lanckriet et al., 2004) the set of feasible composite kernel matrices is restricted to the set of positive-semidefinite matrices with bounded trace which are a linear and non-negative combination of candidate kernel matrices. Therefore to ensure that $trace(\mathbf{K}_{\boldsymbol{\beta}})$ is maintained at a constant value then provided each candidate kernel matrix $\mathbf{K}_k$ is normalised (Shawe-Taylor & Cristianini, 2004) such that for example, $trace(\mathbf{K}_k) = N$, then the non-negative components of $\boldsymbol{\beta}$ require to be constrained such that $\sum_k \beta_k = 1$ and $\beta_k \geq 0 \quad \forall k$. So each $\boldsymbol{\beta}$ is a point on a $K - 1$-dimensional simplex and this suggests that the prior can be defined by a Dirichlet density.

The overall graphical representation of the probabilistic model for regression is given in Figure (1). For the kernel weighting coefficients ($\boldsymbol{\beta}$), right-hand plate, a hierarchical Dirichlet prior is employed where a product of K Gamma distributions with shared parameters define the distribution over the parameters of the Dirichlet. The distributions of the regression weights $\boldsymbol{\alpha}$ are defined as a Scale Mixture of Gaussians (Andrews & Mallows, 1974) as employed in the Relevance Vector Machine (RVM) (Tipping, 2001) and the error distribution is defined as an isotropic Gaussian with precision γ. The conditional dependency structure of distributions over the model parameters can be read directly from the graphical representation (Figure 1). Thus the corresponding model parameter distributions are

$$\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \mathbf{X} \sim \mathcal{N}_{\mathbf{t}}(\mathbf{K}_{\boldsymbol{\beta}}\boldsymbol{\alpha}, \mathbf{I}\gamma^{-1}) \; ; \; \gamma|\rho, \varrho \sim \Gamma_{\gamma}(\rho, \varrho)$$

$$\boldsymbol{\alpha}|\boldsymbol{\phi} \sim \mathcal{N}_{\boldsymbol{\alpha}}(\mathbf{0}, \boldsymbol{\Phi}^{-1}) \; ; \; \boldsymbol{\phi}|\sigma, \varsigma \sim \prod_{m=0}^{N} \Gamma_{\phi_m}(\sigma, \varsigma)$$

$$\boldsymbol{\beta}|\boldsymbol{\varphi} \sim \mathcal{D}_{\boldsymbol{\beta}}(\boldsymbol{\varphi}) \; ; \; \boldsymbol{\varphi}|\tau, \nu \sim \prod_{k=1}^{K} \Gamma_{\varphi_k}(\tau, \nu)$$

where $\boldsymbol{\Phi} = \mathrm{diag}(\phi_0, \cdots, \phi_N)$, $\mathcal{N}_a(b, c)$ defines a Gaussian distribution computed at a with parameters b and c, $\Gamma_a(b, c)$ is a Gamma distribution over a with shape and inverse scale parameters b and c. The Dirichlet distribution for $\mathbf{a}$ with mean value $\mathbf{b}$ is denoted as $\mathcal{D}_{\mathbf{a}}(\mathbf{b})$.

4. Variational Bayes

From the definition of the model, a Gibbs sampler can be developed in a straightforward manner with Metropolis sampling interleaved to obtain samples of the Dirichlet variables and associated parameters. However in this paper we develop a variational Bayes method to obtain an estimate of the required posterior distribution over the model parameters, see for example (Beal, 2003; Jordan et al.,
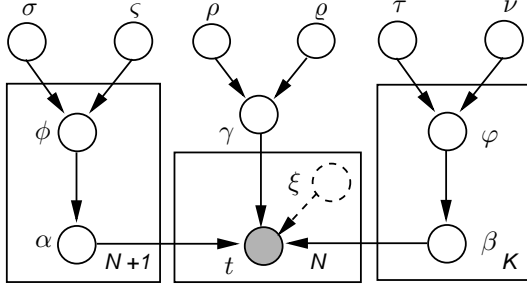
Figure 1. Plates diagram showing the hierarchic Bayesian regression model for combining kernels. For classification the nodes representing γ and its parents are removed and replaced by the node representing the variational parameter ξ shown dashed.

1999; Bishop & Tipping, 2000). By defining the sets of model parameters and associated hyper-parameters as $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\phi}, \boldsymbol{\varphi}\}$ and $\boldsymbol{\psi} = \{\rho, \varrho, \sigma, \varsigma, \tau, \nu\}$ the posterior over all parameters is $P(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{t}, \mathbf{X})$. We assume point estimates for the set of hyper-parameters $\boldsymbol{\psi}$ which can be obtained by type-II Maximum Likelihood estimation and so seek an approximation for the posterior $P(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{t}, \mathbf{X})$. The required posterior can be approximated by a factorable ensemble such that $P(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{t}, \mathbf{X}) \approx \mathcal{Q}(\boldsymbol{\theta}) = Q(\boldsymbol{\alpha})Q(\boldsymbol{\beta})Q(\gamma)Q(\boldsymbol{\phi})Q(\boldsymbol{\varphi})$. Dropping the explicit conditioning on the input data $\mathbf{X}$ and denoting $E_p\{a\}$ as the expectation of the random variable a with respect to the distribution or density p then the well known bound on the model evidence $\log P(\mathbf{t}) \geq E_{\mathcal{Q}(\boldsymbol{\theta})}\{\log P(\mathbf{t}, \boldsymbol{\theta}|\boldsymbol{\psi})\} - E_{\mathcal{Q}(\boldsymbol{\theta})}\{\log \mathcal{Q}(\boldsymbol{\theta})\}$ is minimised by distributions of the form $Q(\boldsymbol{\theta}_i) \propto \exp\left(E_{\mathcal{Q}(\boldsymbol{\theta}_{-i})}\{\log P(\mathbf{t}, \boldsymbol{\theta}|\boldsymbol{\psi})\}\right)$ where $\mathcal{Q}(\boldsymbol{\theta}_{-i})$ denotes the ensemble with the ith component of $\boldsymbol{\theta}$ removed.

4.1. Regression

The optimal distribution associated with the $\boldsymbol{\alpha}$ parameters in Figure (1) are $Q(\boldsymbol{\alpha}) = \mathcal{N}_{\boldsymbol{\alpha}}(\mathbf{m}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$. Using the following notation for the posterior expectations i.e. $\widetilde{a} = E_{Q(a)}\{a\}$ and $\mathbf{k}_{in}$ denotes the $(N+1) \times 1$ vector of kernel values from the ith kernel for data point n, then the associated parameters are defined as

$$\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \left(\widetilde{\gamma} \sum_{i=1}^{K} \sum_{j=1}^{K} \widetilde{\beta_i \beta_j} \sum_{n=1}^{N} \mathbf{k}_{in} \mathbf{k}_{jn}^{\mathrm{T}} + \widetilde{\boldsymbol{\Phi}}\right)^{-1}$$

$$\mathbf{m}_{\boldsymbol{\alpha}} = \widetilde{\gamma} \boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \sum_{n=1}^{N} t_n \sum_{k=1}^{K} \widetilde{\beta_k} \mathbf{k}_{kn}$$

from which the corresponding posterior moments can be obtained. The posterior distribution over $\boldsymbol{\phi}$ and the

required posterior moments are

$$Q(\boldsymbol{\phi}) = \prod_{m=0}^{N} \Gamma_{\phi_m}\left(\sigma + \frac{1}{2}, \frac{1}{2}\widetilde{\alpha_m^2} + \varsigma\right); \quad \widetilde{\phi_m} = \frac{1 + 2\sigma}{\widetilde{\alpha_m^2} + 2\varsigma}$$

The posterior for the precision is also Gamma with the required moment given as

$$Q(\gamma) = \Gamma_\gamma\left(\frac{N}{2} + \rho, \frac{1}{2}\widetilde{\|\mathbf{e}\|^2} + \varrho\right); \quad \widetilde{\gamma} = \frac{N + 2\rho}{\widetilde{\|\mathbf{e}\|^2} + 2\varrho}$$

where

$$\widetilde{\|\mathbf{e}\|^2} = \sum_{n=1}^{N} t_n^2 - 2\sum_{n=1}^{N} t_n \sum_{k=1}^{K} \widetilde{\beta_k} \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}} \mathbf{k}_{kn} + \sum_{i=1}^{K} \sum_{j=1}^{K} \widetilde{\beta_i \beta_j} \Omega_{ij}$$

and $\Omega_{ij} = \sum_{n=1}^{N} \mathbf{k}_{in}^{\mathrm{T}} \widetilde{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{\mathrm{T}} \mathbf{k}_{jn}$. When considering the posterior for the kernel combination weights (right-hand plate of Figure (1) note that the Gaussian and Dirichlet do not form a conjugate-exponential pair and as such there are no closed form representations for $Q(\boldsymbol{\beta})$ or for the required moments of the distribution. However, estimates of the required moments can be obtained using importance sampling. Estimating posterior moments by importance sampling within a variational Bayes setting has been previously employed in (Lawrence et al., 2004). The unnormalised posterior takes the form of

$$Q^*(\boldsymbol{\beta}) = \prod_{k=1}^{K} \beta_k^{\widetilde{\varphi_k}-1} \exp\left\{-\frac{\widetilde{\gamma}}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{b})\right\}$$

where $\boldsymbol{\Omega}$ is the $K \times K$ matrix whose elements are defined above and the K-dimensional vector $\mathbf{b}$ has elements $b_k = \sum_{n=1}^{N} t_n \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}} \mathbf{k}_{kn}$. By drawing S samples from the prior Dirichlet distribution $\boldsymbol{\beta}_s \sim \mathcal{D}_{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\varphi}})$ then the required posterior moments can be estimated as $\widetilde{f(\boldsymbol{\beta})} \approx \sum_{s=1}^{S} f(\boldsymbol{\beta}_s) w(\boldsymbol{\beta}_s)$ where the importance weights are

$$w(\boldsymbol{\beta}_s) = \frac{\exp\left\{-\frac{\widetilde{\gamma}}{2}(\boldsymbol{\beta}_s^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}_s - 2\boldsymbol{\beta}_s^{\mathrm{T}}\mathbf{b})\right\}}{\sum_{s'=1}^{S} \exp\left\{-\frac{\widetilde{\gamma}}{2}(\boldsymbol{\beta}_{s'}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}_{s'} - 2\boldsymbol{\beta}_{s'}^{\mathrm{T}}\mathbf{b})\right\}}$$

and $\widetilde{f(\boldsymbol{\beta})} \equiv \widetilde{\boldsymbol{\beta}}, \quad \widetilde{\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}$ and $\widetilde{\log\{\boldsymbol{\beta}\}}$. The corresponding moments with respect to the posterior distribution over the Dirichlet parameters also require importance sampling. The unnormalised posterior $Q^*(\boldsymbol{\varphi})$ follows as

$$\frac{\Gamma(\sum_k \varphi_k)}{\prod_k \Gamma(\varphi_k)} \prod_k \varphi_k^{\tau-1} \exp\left\{\sum_{k=1}^{K}(\varphi_k - 1)\widetilde{\log \beta_k} - \nu\varphi_k\right\}$$

the required expectation can be obtained by drawing samples $\boldsymbol{\varphi}_s \sim \prod_{k=1}^{K} \Gamma_{\varphi_k}(\tau, \nu)$ and employing these

in the following estimator $\widetilde{f(\varphi)} \approx \sum_{s=1}^{S} f(\varphi_s)w(\varphi_s)$ such that $\widetilde{f(\varphi)} \equiv \widetilde{\varphi}$ and $\widetilde{\log\{\varphi\}}$ where now $w(\varphi_s) = \mu(\varphi_s)/\sum_{s'=1}\mu(\varphi_{s'})$ and

$$\mu(\varphi_s) = \frac{\Gamma(\sum_k \varphi_{sk})}{\prod_k \Gamma(\varphi_{sk})} \exp\left\{\sum_{k=1}^{K}(\varphi_{sk}-1)\widetilde{\log\beta_k}\right\}$$

4.2. Type II ML Hyperparameter Estimation

Point estimates for the set of hyperparameters ψ defining each of the Gamma distributions can be obtained via type II maximum likelihood estimation as defined below.

$$\hat{\sigma}, \hat{\varsigma} = \operatorname*{argmax}_{\sigma,\varsigma} \; E_{Q(\phi)}\left\{\sum_{m=0}^{N}\log P(\phi_m|\sigma,\varsigma)\right\}$$

$$\hat{\tau}, \hat{\nu} = \operatorname*{argmax}_{\tau,\nu} \; E_{Q(\varphi)}\left\{\sum_{k=1}^{K}\log P(\varphi_k|\tau,\nu)\right\}$$

$$\hat{\rho}, \hat{\varrho} = \operatorname*{argmax}_{\rho,\varrho} \; E_{Q(\gamma)}\left\{\log P(\gamma|\rho,\varrho)\right\}$$

As the distributions are all Gamma the required type-II ML solutions for the inverse scale (ς,ν,ϱ) and shape (σ,τ,ρ) parameters can be obtained using a Newton method (Beal, 2003).

4.3. Predictive Distributions

The predictive distribution for a new data point $P(t_{new}|\mathbf{x}_{new},\mathbf{t},\mathbf{X})$ can be approximated under the factorised posterior as

$$\int P(t_{new}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)Q(\boldsymbol{\alpha})Q(\boldsymbol{\beta})Q(\gamma)d\boldsymbol{\alpha} d\boldsymbol{\beta} d\gamma$$

and this can be further approximated by taking the posterior mean values for each of $\boldsymbol{\beta}$ and γ and, due to the conjugate form of the likelihood and posterior for $\boldsymbol{\alpha}$, we obtain the following approximation

$$\int P(t_{new}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\boldsymbol{\alpha},\widetilde{\boldsymbol{\beta}},\widetilde{\gamma})Q(\boldsymbol{\alpha})d\boldsymbol{\alpha}$$

which is Gaussian with associated moments

$$\widetilde{t_{new}} = \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{K}(\mathbf{x})^{\mathrm{T}}\mathbf{m}_{\boldsymbol{\alpha}}; \widetilde{\sigma_{new}^2} = \widetilde{\gamma}^{-1} + \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{K}(\mathbf{x})^{\mathrm{T}}\Sigma_{\boldsymbol{\alpha}}\mathbf{K}(\mathbf{x})\widetilde{\boldsymbol{\beta}}$$

where the $(N+1) \times K$ matrix $\mathbf{K}(\mathbf{x})$ defines the kernel values between the test point $\mathbf{x}_{new}$ and the *training* set (including a bias term) for all K candidate kernels.

4.4. Classification

For classification where the likelihood term is for example a Bernoulli distribution quadratic approximations based on either the Laplace method (Tipping,

2001), Expectation Propagation (EP) (Minka, 2001) or defining a further lower-bound on the likelihood can be employed. In this paper we lower-bound the likelihood term employing the quadratic lower-bound proposed in (Jaakkola, 1997) and adopted in for example (MacKay, 2003; Bishop & Tipping, 2000)

$$\log P(\mathbf{t}|\boldsymbol{\alpha},\boldsymbol{\beta}) \geq \sum_{n=1}^{N}\log\sigma(\xi_n) + \frac{1}{2}\sum_{n=1}^{N}t_n\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{k}_n - \sum_{n=1}^{N}\left(\lambda(\xi_n)\left[(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{k}_n)^2 - \xi_n^2\right] + \frac{\xi_n}{2}\right)$$

where each $t_n = \pm 1$, $\sigma(\cdot)$ represents the logistic function, $\lambda(\xi) = \tanh(\xi/2)/4\xi$ (Jaakkola, 1997) and $\mathbf{k}_n$ represents the composite kernel values for the nth data point. Employing this bound will introduce N additional variational parameters into the model as shown dashed in Figure 1[1]. We lower-bound the likelihood term and the required distributions are $Q(\boldsymbol{\alpha}) = \mathcal{N}_{\boldsymbol{\alpha}}(\mathbf{m}_{\boldsymbol{\alpha}},\boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$ where

$$\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \left(\sum_{i=1}^{K}\sum_{j=1}^{K}\widetilde{\beta_i\beta_j}\sum_{n=1}^{N}2\lambda(\xi_n)\mathbf{k}_{in}\mathbf{k}_{jn}^{\mathrm{T}} + \widetilde{\boldsymbol{\Phi}}\right)^{-1}$$

$$\mathbf{m}_{\boldsymbol{\alpha}} = \frac{1}{2}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\sum_{n=1}^{N}t_n\sum_{k=1}^{K}\widetilde{\beta_k}\mathbf{k}_{kn}$$

The other approximate posterior component which is required is the unnormalised posterior for $\boldsymbol{\beta}$, which now takes the form of

$$Q^*(\boldsymbol{\beta}) = \prod_{k=1}^{K}\beta_k^{\widetilde{\varphi_k}-1}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{b})\right\}$$

where $\Omega_{ij} = \sum_{n=1}^{N}2\lambda(\xi_n)\mathbf{k}_{in}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{k}_{jn}$ and $\mathbf{b}$ is defined as before. As in the case of regression this can be employed in an importance sampler to obtain the required posterior moments. All other components of the ensemble, with the exception of $Q(\gamma)$ which drops out, remain the same as in the case of regression. The variational parameters are obtained using $\xi_n^2 = \mathbf{k}_{\widetilde{\boldsymbol{\beta}}}(\mathbf{x}_n)^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{k}_{\widetilde{\boldsymbol{\beta}}}(\mathbf{x}_n)$ where $\mathbf{k}_{\widetilde{\boldsymbol{\beta}}}(\mathbf{x}_n)$ is the composite kernel defined by the current $\widetilde{\boldsymbol{\beta}}$ values for data point $\mathbf{x}_n$.

5. Maximum a Posteriori Estimators

The variational Bayes approach detailed in the previous section provides an approximate posterior distribution over the parameters of the model which can be

---

[1] In comparison, employing EP would introduce an additional $3N$ parameters associated with the $\boldsymbol{\alpha}$ parameters.

employed in making subsequent predictions. The Maximum a Posteriori (MAP) estimators provide point estimates of the most probable *a posteriori* set of parameters. MAP estimators have a number of significant weaknesses such as basis dependence and the potential of overfitting (MacKay, 2003). Despite these shortcomings it is interesting to consider a MAP estimator in addition to the variational Bayes solution for this problem. The MAP estimator will require a nonlinear optimisation in the general case, however, if we restrict the prior on $\boldsymbol{\beta}$ to uniform sampling from the $K-1$ dimensional simplex ($\varphi_k = 1 \;\; \forall \; k$), as would be the case if *a priori* we have no preference over the number of kernels to enter the model, then we obtain a straightforward constrained quadratic optimisation problem.

$$\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\beta}}, \widehat{\gamma} = \operatorname*{argmax}_{\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\beta}, \gamma} \; \log P(\mathbf{t}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\beta}, \gamma | \boldsymbol{\psi}, \boldsymbol{\varphi} = \mathbf{1})$$

where the *hat* notation represents the MAP estimate. Denoting the $K \times N$ matrix whose i, j^{th} element is $\sum_n \widehat{\alpha_n} K_i(\mathbf{x}_n, \mathbf{x}_j)$ by $\mathbf{Z}_{\widehat{\boldsymbol{\alpha}}}$ then iterating over the following will yield the required MAP solution.

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{K}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \mathbf{K}_{\widehat{\boldsymbol{\beta}}} + \widehat{\gamma}^{-1} \widehat{\boldsymbol{\Phi}} \right)^{-1} \mathbf{K}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \mathbf{t}$$

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_{\widehat{\boldsymbol{\alpha}}}^{\mathrm{T}} \mathbf{Z}_{\widehat{\boldsymbol{\alpha}}} \boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_{\widehat{\boldsymbol{\alpha}}}^{\mathrm{T}} \mathbf{t}$$

$$\mathtt{s.t} \; \sum_{k=1}^{K} \beta_k = 1 \; \& \; \beta_k \geq 0 \; \forall \; k$$

$$\widehat{\phi_m} = \frac{1 + 2\sigma}{\widehat{\alpha_m^2} + 2\varsigma}; \;\; \widehat{\gamma} = \frac{N + 2\rho - 2}{\|\mathbf{t} - \mathbf{K}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{\alpha}}\|^2 + 2\varrho}$$

For classification the *MAP* estimators follow where $\boldsymbol{\Delta}$ is a diagonal matrix whose elements are $2\lambda(\xi_n)$.

$$\widehat{\boldsymbol{\alpha}} = \frac{1}{2} \left(\mathbf{K}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{K}_{\widehat{\boldsymbol{\beta}}} + \widehat{\boldsymbol{\Phi}} \right)^{-1} \mathbf{K}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \mathbf{t}$$

The lower bound is optimised using $\xi_n^2 = \mathbf{k}_{\widehat{\boldsymbol{\beta}}}(\mathbf{x}_n)^{\mathrm{T}} \widehat{\boldsymbol{\alpha}} \widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \mathbf{k}_{\widehat{\boldsymbol{\beta}}}(\mathbf{x}_n)$.

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_{\widehat{\boldsymbol{\alpha}}}^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{Z}_{\widehat{\boldsymbol{\alpha}}} \boldsymbol{\beta} - \frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_{\widehat{\boldsymbol{\alpha}}}^{\mathrm{T}} \mathbf{t}$$

$$\mathtt{s.t} \; \sum_{k=1}^{K} \beta_k = 1 \; \& \; \beta_k \geq 0 \; \forall \; k$$

The remaining hyperparameter *MAP* estimators are identical to those obtained above for regression.

6. Experiments

Some illustrative experiments are now provided to demonstrate the potential of the proposed hierarchic Bayesian models.

6.1. Kernel Target Alignment

In (Cristianini et al., 2002) it is shown that a reweighting of the eigenvector decomposition of a kernel matrix $\mathbf{K}_{\boldsymbol{\lambda}} = \sum_n \lambda_n \mathbf{u}_n \mathbf{u}_n^{\mathrm{T}}$ obtains a new kernel $\mathbf{K}_{\boldsymbol{\beta}} = \sum_n \beta_n \mathbf{u}_n \mathbf{u}_n^{\mathrm{T}}$ which has a greater alignment with a set of target values. Consider the toy data set shown in Figure (2.a) and the associated Gaussian kernel matrix (normalised such that $\sum_n \lambda_n = 1$) in Figure (2.b), note that the points are ordered to aid visualisation. The cluster structure in this data is well characterised by the original kernel matrix (Figure 2.b). However, the kernel is obviously not optimally aligned to the class labels and as such does not fully capture the corresponding class structure. By obtaining a MAP
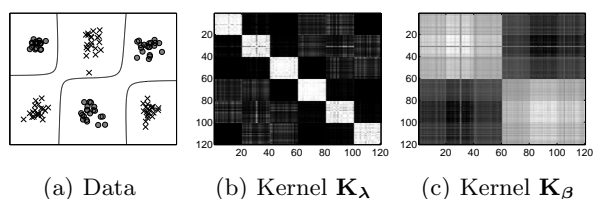

(a) Data   (b) Kernel $\mathbf{K}_{\boldsymbol{\lambda}}$ (c) Kernel $\mathbf{K}_{\boldsymbol{\beta}}$

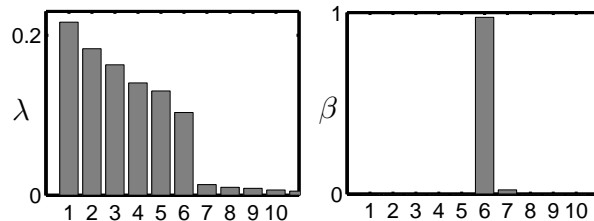*Figure 2.* Scatter plot of two class data sample and associated kernel matrices.

Figure 3. The original kernel matrix eigenvalues λ_n (left) and corresponding MAP kernel weights (right).

estimated classifier it is clearly seen, from the right chart of Figure (3), that the MAP estimates of the kernel matrix weightings $\widehat{\boldsymbol{\beta}}$ is predominantly concentrated on one particular value. The corresponding composite kernel (Figure (2.c)) is now much more aligned with the class labels.

6.2. Multimedia Web Page Classification

The problem of web page classification based on text and image [2] content is considered in (Kolenda et al., 2002). Here we consider the problem of classifying webpages that have been labelled as being related to `Sport` and `Paintball`. From the 800 `Sport` and `Paintball` webpages available 50 random 160/640 train/test splits were used to obtain estimates of pre-

---

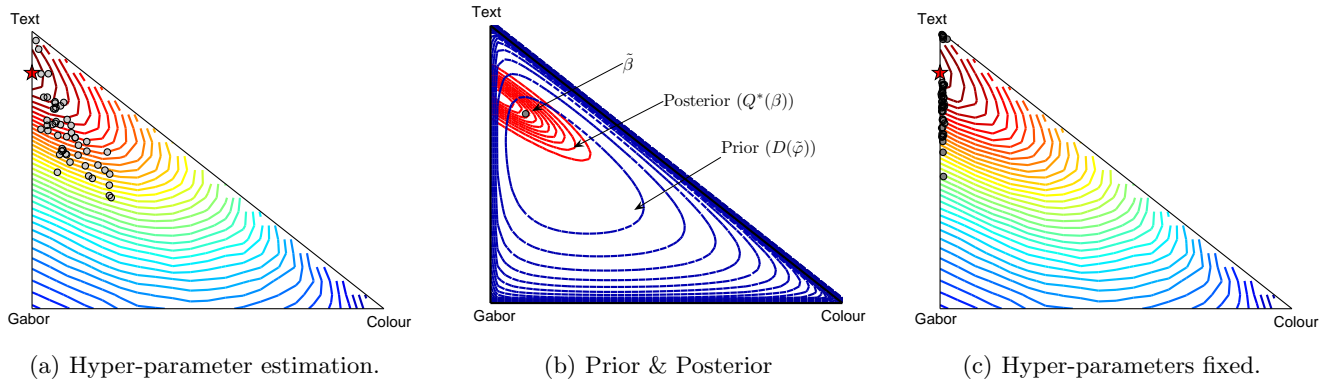[2] Gabor wavelet texture and colour histogram features were extracted

(a) Hyper-parameter estimation.

(b) Prior & Posterior

(c) Hyper-parameters fixed.

Figure 4. Error isocontours and associated posterior means $\widetilde{\boldsymbol{\beta}}$ when hyperparameters τ and ν are estimated (a), the corresponding prior and posterior (b), and when the hyper-parameters are fixed (c).

dictive power. A multinomial diffusion kernel (Lafferty & Lebanon, 2005) was used for the text with two cosine kernels employed for the image texture and colour features. Figure 4 a & c show the isocontours of numbers of correct predictions on the kernel combination simplex obtained by enumerating each feasible weight combination $\boldsymbol{\beta} = [\beta_{Text}, \beta_{Gabor}, \beta_{Color}]$. From the contours it is clear that a maximum number of correct predictions can be achieved when an appropriate combination of the three kernels is employed as denoted by the star. The challenge of course is to find this combination without resorting to extensive enumeration and testing of the possible combinations.

Figure 4 a & c also show the estimated posterior mean values for the combination weights $\widetilde{\boldsymbol{\beta}}$, for each of the 50 random splits of the data, as points on the simplex superimposed on the classification performance contours. On this example we see that hyper-parameter estimation from the data tends to, on average, align the posteriors along the region of lowest error. Figure 4 b gives an example of the prior and associated posterior distribution and its mean value obtained for one of the data splits obtained when estimating the hyper-parameters. It is interesting to note that the prior obtained is giving more weight to regions of the simplex associated with Text although the skew is not so extreme. However, if we fix the hyper-parameters such that $\tau = 0.1$ and $\nu = 1.0$ this has the *a priori* effect of giving higher weight to single kernels. In other words we believe that only a subset of kernels will contribute to achieving a low classification error as the prior Dirichlet will be peaked at the nodes corresponding to either Text, Gabor or Color induced kernels. Figure 4 c shows this effect where the posterior mean values are now strongly concentrated along the edge corresponding to the near optimal mixtures of

Text and Gabor kernels with Color having essentially zero weighting. Figure 5 gives the corresponding distributions for one data split along the Text & Gabor edge. We observe that the prior places most mass on a unit weighting of the Text kernel whilst the posterior has two modes with the estimated mean $\widetilde{\boldsymbol{\beta}}$ being a combination of both kernels. This is a nice example of the dangers of MAP estimators as the maximum of the posterior is achieved when only the Text based kernel is employed.

The performance (in terms of the number of errors made on the test set) for the individual kernels Text, Gabor, Colour are 2.7000 ± 0.0298, 50.0200 ± 0.1194, 60.5400 ± 0.1075, whilst the combination achieves 2.1400 ± 0.0271. This improvement, whilst not dramatic (due to the particular dataset employed) is statistically significant with a p-value of 4.5222×10^{-4} under a paired t-test. The VB method we have proposed has been able to learn a combination of kernels which provide improved performance over that achieved by any of the individual kernels induced from the different forms of data available. Other applications which have genuinely heterogeneous forms of data such as those in Bioinformatics will benefit from the adoption of this Bayesian approach to learning kernels.

6.3. Comparison of Classification Performance

The final illustrative example compares the performance of the VB classification algorithm (using type-II ML estimation of the hyper-parameters) with the Heterogenous Kernel Fisher Discriminant (HKFD) algorithm recently introduced in (Fung et al., 2004) on four datasets employed in (Lanckriet et al., 2004). Following (Lanckriet et al., 2004) three candidate kernels are used which comprise of a Gaussian kernel with
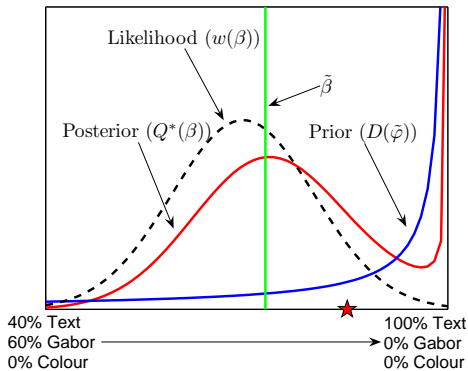
Figure 5. Distributions for β for an example data split viewed along the simplex edge corresponding to combinations of Text and Gabor based kernels.

width (4 for Heart, 5 for 2-norm, and 1 for Sonar & Ionosphere) a linear kernel and a second order polynomial kernel, K_1, K_2 & K_3 respectively. The motivation behind this experiment is to compare the performance of the VB algorithm with hyper-parameter estimation and that of the optimum achievable performance of the HKFD algorithm (i.e. with the single hyper-parameter set by extensive cross validation of the training samples). Thirty (30) random 70% train & 30% test splits of the data were used to obtain the results reported in Table. (1) & Table. (2). Here

Table 1. Classification performance ($\times 100\%$) for Heart, Sonar, Ionosphere & 2-Norm.

HEART	VB	HKFD
K_1	0.8281± 0.0062	0.7992± 0.0237
K_2	0.8259± 0.0078	0.8313± 0.0063
K_3	0.8226± 0.0068	0.7786± 0.0225
$\sum_k^3 K_k$	0.8346 ± 0.0067	0.8292 ± 0.0068
SONAR	VB	HKFD
K_1	0.8038± 0.0089	0.8683± 0.0073
K_2	0.5489± 0.0127	0.7172± 0.0072
K_3	0.5516± 0.0139	0.8038± 0.0093
$\sum_k^3 K_k$	0.8059 ± 0.0080	0.8667 ± 0.0083
IONOSPHERE	VB	HKFD
K_1	0.9038± 0.0044	0.9311± 0.0031
K_2	0.8298± 0.0073	0.9111± 0.0072
K_3	0.9375± 0.0042	0.9305± 0.0041
$\sum_k^3 K_k$	0.9375 ± 0.0043	0.9410 ± 0.0038
2-NORM	VB	HKFD
K_1	0.9589± 0.0074	0.9756± 0.0027
K_2	0.9756± 0.0029	0.9778± 0.0032
K_3	0.9761± 0.0026	0.9783± 0.0028
COMBINED	0.9761 ± 0.0030	0.9783 ± 0.0028

we observe that the results for both methods in terms of overall error for the combined kernel classifiers is comparable with that of the best performing classi-

fier employing a single kernel and this is consistent with results reported by (Lanckriet et al., 2004). We also note similar relative classification performance between both the VB and HKFD methods across all the datasets with the exception of Sonar where HKFD is superior. However we should bear in mind that HKFD requires an additional level of cross-validation to obtain the associated hyper-parameter for each data split which of course is not required in VB. It should also be noted that the composite kernels obtained by both methods are indeed a combination of the candidate kernels (Table. (2)) except in the case of 2-Norm where VB places total weight almost exclusively on the linear kernel. This is exactly what should be expected given that the two classes consists of two overlapping isotropic Gaussians. As the kernels in this example are derived from the same data representation it is unlikely that a combination of the kernels will provide any measurable improvement in performance over the base kernels. However, in the previous example the kernels are derived from independent heterogeneous features (text & images) which provide differing informative representations when combined can, and indeed do, improve predictive ability.

Table 2. Estimated β values for Heart, Sonar, Ionosphere & 2-Norm.

HEART	$\beta_1 \times 10^2$	$\beta_2 \times 10^2$	$\beta_3 \times 10^2$
VB	85.9 ± 6.2	13.0 ± 6.2	0.3 ± 0.1
HKFD	10.6 ± 1.1	22.5 ± 0.5	60.4 ± 5.9
SONAR	$\beta_1 \times 10^2$	$\beta_2 \times 10^2$	$\beta_3 \times 10^2$
VB	95.5 ± 2.4	2.3 ± 1.1	2.2 ± 1.3
HKFD	314 ± 58.5	127 ± 32.8	42.9 ± 7.1
IONOS	$\beta_1 \times 10^2$	$\beta_2 \times 10^2$	$\beta_3 \times 10^2$
VB	32.4 ± 1.8	16.5 ± 0.7	51.2 ± 2.3
HKFD	209 ± 16.8	17.1 ± 1.4	126 ± 10.4
2-NORM	$\beta_1 \times 10^2$	$\beta_2 \times 10^2$	$\beta_3 \times 10^2$
VB	0.8 ± 0.1	98.9 ± 0.1	0.3 ± 0.1
HKFD	15.8 ± 5.1	24.5 ± 1.0	64.8 ± 10.0

7. Conclusions and Discussion

A hierarchic Bayesian modelling framework for kernel learning has been presented and illustrative experiments demonstrate the validity of the approach. In all of the experiments conducted we observe that, at least, the predictive performance of the combined model is comparable to the best individual classifier and if heterogenous data is available overall improvements are achievable, the levels of these improvements are, of course, dataset dependent. This accords with the results of (Lanckriet et al., 2004). Computational scaling is $\mathcal{O}(N^3)$ with storage $\mathcal{O}(KN^2)$ which compares favorably to (Lanckriet et al., 2004; Fung et al., 2004).

Acknowledgments

References

Andrews, D., & Mallows, C. (1974). Scale mixtures of Normal distributions. *Journal of the Royal Statistical Society, Series B, 36*, 99–102.

Bach, F., Lanckriet, G., & Jordan, M. I. (2004). Multiple kernel learning, conic duality and the SMO algorithm. *Proceedings of the Twenty-First International Conference on Machine Learning*.

Bach, F. R., Thibaux, R., & Jordan, M. I. (2005). Computing regularization paths for learning multiple kernels. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems 17*. Cambridge, MA: MIT Press.

Beal, M. (2003). *Variational algorithms for approximate bayesian inference*. Doctoral dissertation, University College London.

Bishop, C., & Tipping, M. (2000). Variational relevance vector machines. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence* (pp. 46–53).

Bousquet, O., & Herrmann, D. J. L. (2003). On the complexity of learning the kernel matrix. In S. T. S. Becker and K. Obermayer (Eds.), *Advances in neural information processing systems 15*, 399–406. Cambridge, MA: MIT Press.

Crammer, K., Keshet, J., & Singer, Y. (2003). Kernel design using boosting. In S. T. S. Becker and K. Obermayer (Eds.), *Advances in neural information processing systems 15*, 537–544. Cambridge, MA: MIT Press.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. (2002). On kernel-target alignment. *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Fung, G., Dundar, M., Bi, J., & Rao, B. (2004). A fast iterative algorithm for Fisher discriminant using heterogeneous kernels. *Proceedings of the twenty-first International Conference on Machine Learning* (pp. 313–320).

Gunn, S., & Kandola, J. (2002). Structural modelling with sparse kernels. *Machine Learning, 48*, 137–163.

Jaakkola, T. (1997). *Variational methods for inference and estimation in graphical models*. Doctoral dissertation, MIT.

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*, 183–233.

Kolenda, T., Hansen, L., Larsen, J., & Winther, O. (2002). Independent component analysis for understanding multimedia content. *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII* (pp. 757–766).

Lafferty, J., & Lebanon, G. (2005). Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research, 6*, 129–163.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L. E., & Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research, 5*, 27–72.

Lawrence, N. D., Milo, M., Niranjan, M., Rashbass, P., & Soullier, S. (2004). Reducing the variability in cDNA microarray image processing by Bayesian inference. *Bioinformatics, 20*, 518–526.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference*. Doctoral dissertation, MIT.

Ong, C. S., Smola, A. J., & Williamson, R. C. (2003). Hyperkernels. In S. T. S. Becker and K. Obermayer (Eds.), *Advances in neural information processing systems 15*, 478–485. Cambridge, MA: MIT Press.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1*, 211–244.

Tsang, I. W., & Kwok, J. T. (2004). Efficient hyperkernel learning using second-order cone programming. *Proceedings of the 15th European Conference on Machine Learning* (pp. 453–464).

Zhang, Z., Yeung, D.-Y., & Kwok, J. T. (2004). Bayesian inference for transductive learning of kernel matrix using the Tanner-Wong data augmentation algorithm. *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 935–942).