Empirical Study of a Bayesian Kernel Model

Kai Mao, Feng Liang, Sayan Mukherjee, and Mike West
Duke University and University of Illinois at Urbana-Champaign
Durham, NC and Urbana-Champaign, IL
{km68,feng,sayan,mike}@stat.duke.edu

Abstract

Kernel models have been used extensively in machine learning due to their flexibility as non-parametric models. Recently a fully Bayesian framework and theory for kernel regression and classification was developed. Models developed in this framework allow for the incorporation of variable selection and other kernel hyper-parameters. This paper is an empirical study of the performance and modeling considerations of this family of models. We first illustrate the efficacy of the model as a point estimator for a variety of simulated and real data. The true benefit of the Bayesian model – estimates and formal modeling of uncertainty – is examined by simulating the posterior distributions of predictions as well the posterior distributions on the relevance of input variables. We will discuss the sensitivity of the posterior to both priors on model parameters and hyperparameters.

1 Introduction

Kernel models for regression and classification have been used extensively in machine learning [9, 16, 11, 12]. The appeal of these models are their predictive accuracy, flexibility, and simple extension to high-dimensional data analysis. In the case of certain kernels they are universal approximators [3] and can thus approximate any square integrable function – the basis for non-parametric modeling.

The basic formulation of kernel methods is a penalized loss functional [3]

$$\hat{f} = \arg\min_{f \in \mathcal{H}} [L(f, \text{data}) + \lambda \|f\|_K^2],$$

where $L$ is a loss function, $\mathcal{H}$ is often an infinite-dimensional reproducing kernel Hilbert space (RKHS), $\|f\|_K^2$ is the norm of a function in this space, and $\lambda$ is a tuning parameter chosen to balance the trade-off between fitting errors and the smoothness of the function. The data are input-output pairs $\{(x_i, y_i)\}_{i=1}^n$ with $x \in \mathbb{R}^p$ and $y \in \{0, 1\}$, for classification. Due to the representer theorem [4] the solution of the penalized loss functional will be a kernel

$$\hat{f}(x) = \sum_{i=1}^n w_i K(x, x_i),$$

where $\{x_i\}_{i=1}^n$ are the $n$ observed input or explanatory variables.

A Bayesian formulation for kernel methods would provide a natural framework to further the richness and interpretability of kernel models – a program driving much research in data mining and machine learning. Crucially it would provide a formal model of uncertainty and allow for the inference of a full posterior rather than a point estimator.

Bayesian versions of kernel models, in particular Bayesian SVMs, have been proposed before [15, 1, 14]. Here the Bayesian analysis is applied on the finite representation from equation (2). This direct adoption of (2) is not a proper model from a Bayesian perspective since the model changes with the sample size without a coherent argument.
A fully Bayesian framework and theory for kernel regression and classification addressing the above issues and specifying priors over the entire RKHS was developed in [5]. This model also allowed for variable or feature selection by placing hyperparameters on the kernel. In this paper we provide a careful empirical study of this method and discuss modeling issues such as sensitivity to the priors and uncertainty in the posterior estimates. The summary of this analysis is that our model performs as well or better than well studied and popular methods such as SVMs but provide a formal measure of uncertainty.

2 The Bayesian kernel model

In this section we specify the model developed in [5]. The most common approach to putting priors over a RKHS is through the duality between RKHS and Gaussian processes [7, 17]. We do not pursue this approach due to theoretical reasons. The RKHS induced by this approach is larger than the one corresponding to the kernel \( k \). See [5, 8] for detailed discussions.

In [5] the space of functions used in modeling is defined as a convolution of the kernel with a signed (Borel) measure

\[
\mathcal{G} = \left\{ f \mid f(x) = \int k(x, u) \, d\gamma(u), \quad \gamma \in \Gamma \right\},
\]

with \( \Gamma(\cdot) \) as a subset of the space of signed Borel measures. Placing a prior on \( \Gamma \) implies a prior on \( \mathcal{G} \). An equivalence between \( \mathcal{G} \) and \( \mathcal{H} \) exists for an appropriate choice of the the prior distribution on \( \Gamma \) [8]. The prior specified in [5] is such a choice.

The model in (3) can be rewritten as

\[
f(x) = \int k(x, u) \, d\gamma(u) = \int k(x, u) \, w(u) \, dF(u),
\]

where the random signed measure \( \gamma(u) \) is modeled by a random probability distribution \( F(u) \) and random coefficients \( w(u) \) and \( F(u) \) and \( \gamma(u) \) share the same support. In [5] it is assumed that \( F = F_X \), the marginal distribution of \( X \), and a Dirichlet Process (DP) prior, \( DP(\alpha, F_0) \), is used to model uncertainty about the distribution function \( F \). A result of the conjugacy of the DP prior and marginalization over the posterior distribution is that given data \( X_n = (x_1, \ldots, x_n) \)

\[
\mathbb{E}[f \mid X_n] = (\alpha + n)^{-1} \left[ \alpha \int k(x, u) \, w(u) \, dF_0(u) + \sum_{i=1}^{n} w(x_i) \, k(x, x_i) \right],
\]

taking the limit \( \alpha \to 0 \), an uninformative prior, the above takes the same form as (2) derived via the representer theorem with \( w_i = \frac{w(x_i)}{n} \).

This results the following classification model

\[
z_i = f(x_i) + \varepsilon = w_0 + \sum_{j=1}^{n} w_j \, k(x_i, x_j) + \varepsilon \quad \text{for } i = 1, \ldots, n,
\]

where \( \varepsilon \sim \text{No}(0, 1) \) and \( z_i \) is a latent variable that is related to \( y_i \) by a probit link function, \( P(y_i = 1) = \Phi(z_i) \), with \( \Phi(\cdot) \) being the standard normal distribution function.

To allow for variable or features selection a scale parameter is added for each coordinate in the kernel [2]

\[
k_{\rho}(x, u) = k(\sqrt{\rho} \otimes x, \sqrt{\rho} \otimes u)
\]

where \( a \otimes b \) is the element-wise product of two vectors and \( \rho = (\rho_1, \ldots, \rho_p) \) is a \( p \)-dimensional vector with \( \rho_k \in [0, \infty] \). For the linear and Gaussian kernels the paramterization is

\[
k_{\rho}(x, u) = \sum_{k=1}^{p} \rho_k \, x_k \, u_k,
\]

\[
k_{\rho}(x, u) = \exp \left( -\sum_{k=1}^{p} \rho_k (x_k - u_k)^2 \right).
\]
The likelihood for the model is
\[
\prod_{i=1}^{n} \Phi(z_i)^{y_i} (1 - \Phi(z_i))^{1-y_i}.
\]
The hierarchical prior involves a decomposition of the kernel, \(K = F\Delta F'\). A result of the decomposition is a reparameterization of the regression into the orthogonal factors, \(F\) with \(Kw = F\beta\). We now state our prior specifications
\[
\begin{align*}
    w_0 &\sim \text{Unif}(\mathbb{R}), & \beta_j &\sim \text{No}(0, \tau_j), & j &= 1, \cdots, m \\
    \rho_j &\sim (1 - \gamma)\delta_0 + \gamma \text{Gamma}(a_p, a_p s), & j &= 1, \cdots, p \\
    \tau_j &\sim \text{InvGamma} \left( \frac{a_\tau}{2}, \frac{b_\tau}{2} \right), & s &\sim \text{Exp}(a_s), & \gamma &\sim \text{Beta}(a_\gamma, b_\gamma)
\end{align*}
\]
a_\tau, a_s, a_\gamma, b_\gamma, a_p, b_p are all hyperparameters that are prespecified. Due to numerical stability not all \(n\) factors are included in the decomposition of the kernel matrix resulting in the parameter \(m < n\).

The posterior is simulated based on a Markov chain Monte Carlo (MCMC) analysis. A novel MCMC methodology was developed in [5] that added Metropolis-Hastings steps appropriately to deal with the fact that the complete conditionals were not of standard forms. The Metropolis-Hastings steps were required to update the scale parameters for each coordinate, \(\bm{\rho}\).

The empirical studies will examine predictive accuracy via draws from the posterior. Given a new input \(x^*\) we provide posterior probabilities of its label \(y^* = 1 = \Phi(z^{*}(d)) = \Phi(w_0^{(d)} + \sum_{j=1}^{m} k_{\rho(j)}(x^*, z_j) w_j^{(d)})\), where \(\{z^{*}(d), w_0^{(d)}, \rho^{(d)}, w_j^{(d)}\}\) are samples \(d = 1, \cdots, D\) from the Markov chain. Note that for different draws the kernel matrix changes since \(F\) changes. The other focus of this paper is to study variable selection which will be done by drawing posterior samples of \(\bm{\rho}\). The Bayesian kernel model will be termed BAKER.

### 3 Experiments

The following experiments on real and simulated data provide an indication of the accuracy of BAKER as a point estimator, a reflection of the uncertainty in the point estimates, and the efficacy and uncertainty in variable selection.

#### 3.1 High-dimensional gene expression data – uncertainty in predictions

We first consider a high-dimensional gene expression data set. The purpose of this analysis is to compare the BAKER to an extensively used classifier, the SVM. This example also highlights the utility of the model uncertainty inherent in Bayesian methods. In this example we do not include the variable selection scale parameters to the model.

The data set [10, 6] consists of 280 samples from patients of which 190 are tumor samples and 90 are normal. For each sample expression data from 16063 genes and ESTs was collected. The data was randomly split into training and test sets with 180 and 100 samples respectively. A linear kernel was used for both BAKER and SVM with accuracy over twenty random test-train splits of 97.2 ± 1.0% (training) and 88.3 ± 4.1% (test) for BAKER and 98.9 ± .5% (training) and 91.3 ± 2.8% (test) for SVM.

The true utility of BAKER is that it provides a predictive distribution and not just a point estimate. We illustrate this on one random test-train split where the test error was 13%. In Figure (1) we plot the mean of the posterior distribution for each test sample in red and provide a point-wise 90% credible interval for each test sample. A result of this analysis is that in addition to misclassified instances there are 7 correctly classified samples with great uncertainty (credible intervals covering 0.5) that may require further investigation.

#### 3.2 Simulated nonlinear classification – variable selection

A data set was generated to illustrate the variable selection aspect of BAKER. The data is perfectly separable by a nonlinear classifier and has two relevant dimensions and ten noise dimensions.
Samples from class 0 were drawn from
\[
(x^1, x^2) = (r \sin(\theta), r \cos(\theta)), \quad \text{where } r \sim \text{Unif}[0, 1] \text{ and } \theta \sim \text{Unif}[0, 2\pi],
\]
\[
x^j \sim \text{Unif}[-2, 2], \quad \text{for } j = 3, \ldots, 12.
\]

Samples from class 1 were drawn from
\[
(x^1, x^2) = (r \sin(\theta), r \cos(\theta)), \quad \text{where } r \sim \text{Unif}[1, 2] \text{ and } \theta \sim \text{Unif}[0, 2\pi],
\]
\[
x^j \sim \text{Unif}[-2, 2], \quad \text{for } j = 3, \ldots, 12.
\]

The first two dimensions are signal with class 1 contained in the unit circle and class 0 contained in an annulus outside the unit circle. A draw of the first two dimensions of the data is shown in Figure (2a).

We drew 100 training and 100 test samples, 50 for each class. A Gaussian kernel was used and the empirical training and test errors were 0% and 3%, respectively. Posterior means of the precision parameters \(\rho_1\) and \(\rho_2\) and the prediction on the region \([-2, 2] \times [-2, 2]\) for the first two dimensions are shown in Figure (2b,c). The posterior means for \(\rho_1\) and \(\rho_2\) are significantly larger than those of dimension 3 to 12 as expected. The prediction on the whole region is accurate. The boundary is approximately the unit circle, the optimal separating boundary.

### 3.3 UCI Data sets

We further examine variable selection and the predictive posterior distribution on several data sets from the UCI Machine Learning Repository\(^1\).

#### 3.3.1 Wisconsin Breast Cancer Dataset

The Wisconsin Breast Cancer data was examined in [18] using a variety of classifiers. The class labels were benign and malignant and after removal of samples with missing attributes 353 samples remained. The input data had nine attributes. The best result reported in [18] with a training set of 200 samples with 100 from each class and the remaining samples as test was 93.7% by 1 nearest neighbor. We repeated this comparison using BAKER, SVM, and a Generalized Linear Model (GLM) with a probit link as the classification algorithms over ten random 200-153 training-test splits. The classifiers perform comparably: 96.0 ± 1.2% (train) and 95.2 ± .9% (test) for Baker; 96.0 ± 0.9% (train) and 95.6 ± 1.7% (test) for SVM; 96.0 ± 1.2% (train) and 94.7 ± 1.3% (test) for GLM.

\(^1\)ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
The posterior predictive distribution for each sample in the test set is displayed in Figure 3. BAKER incorrectly classifies 8 samples, 7 of which are benign. However, the 95% credible intervals for each sample indicate that the uncertainty of some of the patients is much larger than that for some others. The credible intervals for a few incorrectly classified samples do not cover 0.5. It would be interesting to look more carefully at these samples, especially sample 8 which has very low posterior probability.

All 9 attributes seem relevant as reflected by the posterior means in Figure 4. The second variable “Uniformity of Cell Size” and the ninth variable “Mitoses” seem to be weaker than the others.

### 3.3.2 Johns Hopkins University Ionosphere database

The ionosphere data was examined in [13] using a variety of neural networks. The class labels were good and bad. The input data had 34 attributes. They used the first 200 samples which included 100 bad and good samples as the training set and tested on the remaining 151 samples. They achieved test accuracies of 90.7% by a linear perceptron, 92% by a nonlinear perceptron and 96% by their
Figure 4: Boxplots of the posterior distribution of the the nine values of $\rho$.

The implementation of backprop. We repeated this comparison using BAKER and SVM with a Gaussian kernel. The classifiers perform comparably: 97.5 ± .9% (train) and 95.0 ± 1.9% (test) for Baker; 98.9 ± .4% (train) and 96.0 ± 1.1% (test) for SVM.

The posterior predictive distribution on the test set as well as the posteriors for the relevance of each variable are plotted in Figure 5. Classifier is accurate on the test set but the uncertainty for some of the test samples is considerable. Variables 5, 9, 14, 23 have larger relative values and hence should be more significant.

Figure 5: (a) The posterior predictive probability of a sample being good on the test samples. The first 125 samples are good and the remaining 75 are bad. Red stars are posterior means and blue lines are 95% credible intervals. (b) A boxplot showing the relevant significance for the explanatory variables.
4 Modeling considerations

The two most important considerations in Bayesian modeling are the computational efficiency of sampling from the posterior and the sensitivity of the posterior to prior specifications.

We first consider efficiency. The issue is the amount of time required to obtain coverage of the posterior distribution. The majority of time in the MCMC algorithm is in recomputing the kernel matrix for a new draw of $\rho$ and the computation of the kernel decomposition. For example, a typical run on the Wisconsin Breast Cancer dataset with 1,000 iterations takes 153.00 seconds in which 91.2% or 139.54 seconds are spent on recomputing and decomposing the kernel (CPU: Intel Xeon 2×2.8GHz; Memory: 4096M; Matlab). This is the computational cost for feature selection.

The sensitivity of the posterior to prior specifications is another important issue. We have observed for the model without variable selection the posterior is relatively insensitive to the hyperparameters. For the model with variable selection hyperparameters have an effect on the posterior distribution of relevance for each variable, $\rho$. However, the prediction is robust with respect to the hyperparameters.

In the model with variable selection the two considerations are the probability that a $\rho$ is zero or not and the magnitude of a nonzero $\rho$. The parameter $\gamma$ models the prior probability of a variable being relevant reflected by a Beta distribution. We use $a_{\gamma} = b_{\gamma} = 5$ to reflect a prior assumption of half the variables being relevant and the magnitude reflects some degree of concentration. Other values are also possible, but generally do not significantly influence the prediction. The posterior on the $\rho$ values is more sensitive to prior specifications. This is due to the Gamma distribution draws from which can be very large. One possibility to address this is to use a truncated Gamma distribution. Another approach is to examine the hyperparameters that control the tail of the Gamma distribution. The prior magnitude of $\rho$ depends on $s$ which in turn depends on $a_{s}$. We examine the effect of $a_{s}$ on the posterior by running the model setting $a_{s} = 0.5, 1, 5$. The comparisons are run on the Wisconsin Breast Cancer data set. The predictions are robust for all three settings, for example, we examine the first explanatory variable in the first instance of the test set. Figure (6) provides a boxplot of $\rho_1$ for the three settings of $a_{s}$. It is apparent that the posterior means of $\rho_1$ is very stable however the variance of the posterior distribution of $\rho_1$ is not. As $a_{s}$ increases very large values of $\rho_1$ appear in the posterior distribution. These outliers drive the uncertainty in $\rho$. It is very important to consider this hyperparameter when we model the uncertainty of variable relevance.

5 Discussion

We provide a careful empirical examination of a fully Bayesian framework and theory for kernel regression and classification, BAKER, that also allows for variable selection. A conclusion of this study is that BAKER performs comparably well as classifiers such as SVMs while providing a formal measure of model uncertainty. The importance of an estimate of this uncertainty is illustrated on a variety of data sets. The accuracy and models of uncertainty for variable selection is also examined. Lastly we study the priors in the model to better understand setting hyper-parameters.

References


Figure 6: Boxplots for the posterior draws for $\rho_1$ (upper panel) and the posterior probability of the first benign sample under different hyperparameter values of $a_s$. The predictions shown are similar, but the distribution of $\rho_1$ differs greatly due to outliers under larger $a_s$ values.


