ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Generalised Cross Validation for Noise-Free Data

Tony J. Dodd and Tun M. Ladoni
Department of Automatic Control and Systems Engineering
The University of Sheffield, Sheffield S1 3JD, UK
e-mail: {t.j.dodd, tun.ladoni}@shef.ac.uk

Abstract. Whilst machine learning is principally concerned with function approximation from noisy data there are situations where the data maybe noise-free. This arises, for example, in metamodelling where we seek models of computationally expensive high fidelity simulation models. In this paper we derive a noise-free version of generalised cross validation (GCV) which can be used for model selection and hyperparamter estimation in metamodelling. This noise-free GCV measure is applied to the determination of the optimal kernel width in a reproducing kernel Hilbert space interpolation problem.

1 Introduction

Within computational engineering design and simulation, codes such as computational fluid dynamics and finite element methods are routinely used. These are applied, for example, to the solution of aerodynamic flows, electromagnetics and structural analysis [4]. However, the computational costs associated with using such high fidelity simulation models can severely restrict the space of engineering designs which can be successfully searched by, for example, a multi-objective optimisation algorithm [3].

Metamodelling is receiving increasing attention in a number of application areas where the original models are computationally very expensive. Meta- or surrogate models are simply computationally efficient models of models. Due to the high computational cost (often hours or days) of generating data points from the original model, model selection and hyperparameter estimation becomes a real problem in metamodelling.

Algorithms for model selection and hyperparameter estimation are usually based on optimising the generalisation performance of the model. Various measures of generalisation performance exist which are based on the predictive mean-squared error of the model, the most obvious being to measure the generalisation performance on some independent validation data set. However, in metamodelling the computational expense of generating a validation set is often too high.

Various measures of generalisation performance which do not require an

independent validation set have been proposed. These include various forms of cross validation, including ordinary (OCV) and generalised cross validation (GCV) [5], and measures based on the empirical risk plus some additional term which measures model complexity, for example Mallow's C_p statistic or the Akaike Information Criterion (AIC). A common feature of all these measures though is that they are statistically based and assume that the data is noisy. They are therefore not immediately applicable to metamodelling, which is inherently a noise-free problem.

In this paper we derive a noise-free version of the GCV measure for function approximation in reproducing kernel Hilbert spaces (RKHS). Our final result is identical to one given in [5] on the convergence of the GCV measure for spline approximation but not derived there. We also compare our noise-free GCV measure to the asymptotic performance of the traditional OCV and GCV measures.

2 Measuring Generalisation Performance

Given a RKHS, $\mathcal{F}$, the set of reproducing kernels, $\{k_i\}_{i=1}^N \subset \mathcal{F}$, and the set of observations $y_i = L_i f$ the RKHS interpolation problem is to find a function $f \in \mathcal{F}$ such that $y_i = L_i f, i = 1, \ldots, N$. The (Tikhonov) regularised solution is given by [1]

$$f_\lambda = L^*(\lambda I + LL^*)^{-1} y \tag{1}$$

where $L = \sum_{i=1}^N (L_i f) e_i$, e_i is the ith standard basis vector, L^* is the adjoint operator of L and λ is the regularisation parameter.

Define the predictive mean-square error, $T(\lambda)$, as [5]

$$T(\lambda) = \frac{1}{N} \sum_{i=1}^{N} (L_i f_\lambda - L_i f)^2. \tag{2}$$

Obviously if the function estimate, f_λ, exactly interpolates the (noise-free) observations $L_i f$, $T(\lambda)$ will be zero. Define the influence matrix, $A(\lambda)$, by

$$\begin{bmatrix} L_1 f_\lambda \\ \vdots \\ L_N f_\lambda \end{bmatrix} = A(\lambda) y \tag{3}$$

where $A(\lambda) = K(K + \lambda I)^{-1}$ and $K = LL^*$. Then

$$T(\lambda) = \frac{1}{N} \| (I - A(\lambda)) y \|^2. \tag{4}$$

The expectation of $T(\lambda)$ may be expected to provide more information in the noise-free case. Assuming, temporarily that the data are noisy and

$$y_i = L_i f + \epsilon_i, \quad i = 1, \ldots, N \tag{5}$$

and letting $g = [L_1 f, \ldots, L_N f]^T$ we have

$$ET(\lambda) = \frac{1}{N} E\|A(\lambda)(g+\epsilon) - g\|^2 \quad (6)$$

$$= \frac{1}{N}\|(I - A(\lambda)g)\|^2 + \frac{\sigma^2}{N}\mathrm{tr}A^2(\lambda). \quad (7)$$

Since $\sigma^2 = 0$ and $A(\lambda) = I$ for strict interpolation of noise-free data we see that the expected value of $T(\lambda)$ is also equal to zero. Many measures of generalisation peformance, such as Mallow's C_p and AIC are based on similar forms. They are only applicable to noisy data and are not relevant in our case. We therefore seek alternatives measures of interpolation generalisation performance. The cross validation procedure is well-known as a measure of generalisation performance [5].

The well-known GCV measure is given by

$$V(\lambda) = \frac{\frac{1}{N}\|(I - A(\lambda))y\|^2}{[\frac{1}{N}\mathrm{tr}(I - A(\lambda))]^2} \quad (8)$$

where tr is the trace operator. GCV is a predictive mean-square error criteria which was introduced to achieve certain desirable invariance properties which do not hold for ordinary cross validation [5]. It has been widely applied and has been found to provide a reliable estimate of generalisation performance. However, in the case of noise-free data, strict interpolation gives $V(\lambda) = 0/0$ and therefore it is not directly applicable to such cases. In the next section we use l'Hôpital's rule to derive a noise-free version of GCV.

3 GCV for Interpolation

We first re-write (8) in terms of the eigenvalues, λ_i, of K as [2]

$$V(\lambda) = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\lambda}{\lambda_i+\lambda}\right)^2 z_i^2}{\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\lambda}{\lambda_i+\lambda}\right)^2} = \frac{a(\lambda)}{b(\lambda)} \quad (9)$$

where we introduce the last equality for convenience and the $z_i = \Gamma y$ where $A = \Gamma\Phi\Gamma$.

If we set $\lambda = 0$ we have $V(0) = 0/0$. We therefore seek a solution as $\lambda \to 0$ by iteratively applying l'Hôpital's rule:

$$\lim_{\lambda\to 0}\frac{a(\lambda)}{b(\lambda)} = \lim_{\lambda\to 0}\frac{\partial a(\lambda)/\partial\lambda}{\partial b(\lambda)/\partial\lambda} = \cdots = \lim_{\lambda\to 0}\frac{\partial^n a(\lambda)/\partial\lambda^n}{\partial^n b(\lambda)/\partial\lambda^n}. \quad (10)$$

Now

$$\frac{\partial a(\lambda)}{\partial\lambda} = \frac{2}{N}\sum_{i=1}^{N}\frac{\lambda}{\lambda_i+\lambda}\frac{\lambda}{(\lambda_i+\lambda)^2}z_i^2 = \frac{2}{N}\sum_{i=1}^{N}\frac{\lambda\lambda_i}{(\lambda_i+\lambda)^3}z_i^2 \quad (11)$$

and

$$\frac{\partial b(\lambda)}{\partial \lambda} = \left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda}{\lambda_i + \lambda} \right) \left(\frac{1}{N} \sum_{j=1}^{N} \frac{\lambda_j}{(\lambda_j + \lambda)^2} \right). \tag{12}$$

Therefore, applying l'Hôpital's rule,

$$\lim_{\lambda \to 0} \frac{a(\lambda)}{b(\lambda)} = \lim_{\lambda \to 0} \frac{\partial a(\lambda)/\partial \lambda}{\partial b(\lambda)/\partial \lambda} = \frac{0}{0}. \tag{13}$$

Taking the next iteration

$$\frac{\partial^2 a(\lambda)}{\partial \lambda^2} = \frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i(\lambda_i + \lambda)^3 - 3\lambda\lambda_i(\lambda_i + \lambda)^2}{(\lambda_i + \lambda)^6} z_i^2 \tag{14}$$

$$= \frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i^2 - 2\lambda\lambda_i}{(\lambda_i + \lambda)^4} z_i^2 \tag{15}$$

and

$$\begin{aligned} \frac{\partial^2 b(\lambda)}{\partial \lambda^2} &= \left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i}{(\lambda_i + \lambda)^2} \right) \left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i}{(\lambda_i + \lambda)^2} \right) \\ &\quad + \left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda}{\lambda_i + \lambda} \right) \left(\frac{1}{N} \sum_{i=1}^{N} \frac{2\lambda_i(\lambda_i + \lambda)}{(\lambda_i + \lambda)^4} \right) \\ &= \left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i}{(\lambda_i + \lambda)^2} \right)^2 + \left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda}{\lambda_i + \lambda} \right) \left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i}{(\lambda_i + \lambda)^3} \right). \end{aligned}$$

We now have

$$\lim_{\lambda \to 0} \frac{\partial^2 a(\lambda)/\partial \lambda^2}{\partial^2 b(\lambda)/\partial \lambda^2} = \frac{\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i^2}{\lambda_i^4} z_i^2}{\left(\frac{2}{N} \sum_{i=1}^{N} \frac{\lambda_i}{\lambda_i^2} \right)^2} = \frac{\frac{1}{N} \sum_{i=1}^{N} \frac{z_i^2}{\lambda_i^2}}{\left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\lambda_i} \right)^2}. \tag{16}$$

Finally, this is equivalent to

$$V(0) = \lim_{\lambda \to 0} \frac{\partial^2 a(\lambda)/\partial \lambda^2}{\partial^2 b(\lambda)/\partial \lambda^2} = \frac{\frac{1}{N} \|K^{-1} y\|^2}{\left(\frac{1}{N} \mathrm{tr} K^{-1} \right)^2}. \tag{17}$$

4 Example

The noise-free GCV measure was applied to the determination of the optimum value of β in the noise-free interpolation problem of estimating the function

$$y(x) = \mathrm{sinc}(6x) \tag{18}$$

on the interval $[-0.5, 0.5]$ using 10 data points and the reproducing kernel $k_i(x) = \exp(-\beta\|x - x_i\|^2)$. Figure 1 shows the variation with λ of the normal GCV, OCV and calculated leave-one-out cross validation (LOOCV) measures for 10 uniformly spaced data points. For $\lambda < 10^{-5}$ the values are seen to converge before those for GCV and OCV suddenly increase around $\lambda = 10^{-14}$. This is due to the limited machine precision preventing accurate calculation of GCV and OCV. Theoretically these estimates are computable for any $\lambda > 0$ and the equations will only fail in the specific case of $\lambda = 0$. The value of GCV in the plateau around $10^{-13} \leq \lambda \leq 10^{-5}$ is identical to the value of $V(0) = 0.0016$ given by (17). The form (17) is therefore consistent with the trend of the usual GCV as $\lambda \to 0$ before machine precision becomes a problem.

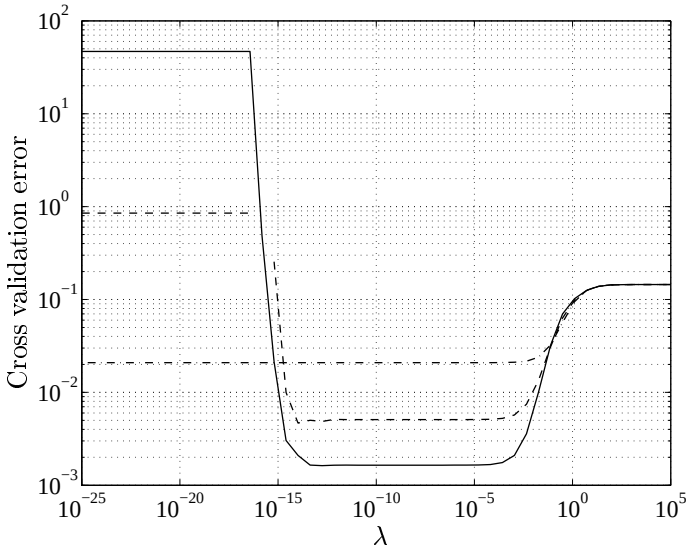

Figure 1: Comparison of cross validation measures for simple interpolation problem. Shown are GCV ('—'), OCV ('– –') and LOOCV ('– · –') estimates.

The variation of $V(0)$ with β is shown in Figure 2 where the solid line represents the average GCV error of a Monte Carlo study of 50 realisations of the data points with the inputs drawn from a uniform distribution. The dashed line is from a single run with the inputs uniformly spaced in the interval $[-0.5, 0.5]$. In both cases a clearly defined optimum value of β can be located.

5 Conclusions

A noise-free measure of generalisation performance for RKHS based machine learning algorithms has been derived. This was motivated by the problem of metamodelling which requires computationally efficient methods for assessing generalisation performance. The potential of the measure was demonstrated

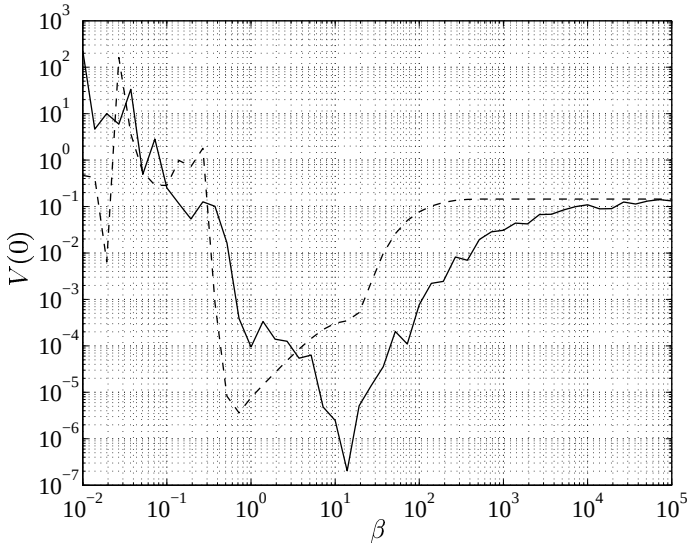

Figure 2: Variation of $V(0)$ with β for Monte Carlo study ('—') and linearly space data ('−−').

on an example function interpolation problem where the optimum value of the hyperparameter β in the well-known Gaussian kernel was determined. Future work will focus on how to calculate the measure for large data sets where ill-conditioning of the kernel Gram matrix becomes a problem and investigating the consistency of $V(0)$ for estimating model hyperparameters.

References

[1] T.J. Dodd and R.F. Harrision. The gradient iteration for approximation in reproducing kernel Hilbert spaces. *IMA Journal of Mathematical Control and Information*, 21:359–376, 2004.

[2] T.J. Dodd and T.M. Ladoni. A new measure of generalisation performance for surrogate models. Technical Report 879, Department of Automatic Control and Systems Engineering, University of Sheffield, UK, 2004.

[3] Y. Jin. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing*, 2003.

[4] Y.S. Ong, P.B. Nair, and A.J. Keane. Evolutionary optimization of computationally expensive problems via surrogate modeling. *AIAA Journal*, 41(4):687–696, 2003.

[5] G. Wahba. *Spline Models for Observational Data*, volume 50 of *Series in Applied Mathematics*. SIAM, Philadelphia, 1990.