ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 133-138

Sparse Bayesian Kernel Logistic Regression

Gavin C. Cawley and Nicola L. C. Talbot

School of Computing Sciences
University of East Anglia
Norwich, U.K. NR4 7TJ
gcc@sys.uea.ac.uk

Abstract. In this paper we present a simple hierarchical Bayesian treatment of the sparse kernel logistic regression (KLR) model based MacKay's evidence approximation. The model is re-parameterised such that an isotropic Gaussian prior over parameters in the kernel induced feature space is replaced by an isotropic Gaussian prior over the transformed parameters, facilitating a Bayesian analysis using standard methods. The Bayesian approach allows the selection of "good" values for the usual regularisation and kernel parameters through maximisation of the marginal likelihood. Results obtained on a variety of benchmark datasets are provided indicating that the Bayesian kernel logistic regression model is competitive, whilst having one less parameter to determine during model selection.

1 Introduction

Given labelled training data, $\mathcal{D} = \{(\boldsymbol{x}_i,\ t_i)\}_{i=1}^{\ell},\ \boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d,\ t_i \in 0,\ 1$, kernel logistic regression [1] aims to construct a statistical decision rule of the form

$$\text{logit}\{y(\boldsymbol{x};\boldsymbol{\alpha})\} = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}),$$

where $\mathcal{K}$ is a kernel function, commonly the Gaussian radial basis function, $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\eta\|\boldsymbol{x} - \boldsymbol{x}'\|^2\}$ (note we have omitted the usual bias parameter for ease of exposition). The output of the model can be interpreted as an estimate of *a-posteriori* probability, i.e. $y(\boldsymbol{x}) \approx p(t = 1|\boldsymbol{x})$. The optimal model parameters $\boldsymbol{\alpha}$ are determined by minimising a regularised [2] likelihood training criterion,

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} C\left(t_i, y(\boldsymbol{x}_i;\boldsymbol{\alpha})\right) + \frac{\mu}{2}\boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha}, \tag{1}$$

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 133-138

where $C(t, y) = t \log y + (1 - t) \log(1 - y)$ is the familiar cross-entropy loss function, μ is a regularisation parameter controlling the bias-variance trade off [3] and the Gram matrix $\boldsymbol{K} = [\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{\ell}$. This optimisation problem can be solved via the iteratively re-weighted least-squares (IRWLS) procedure, e.g. [4]. The value of the regularisation parameter, μ, is critical to obtaining optimal generalisation; in this paper we demonstrate a simple Bayesian method to determine a good value for this parameter via the evidence framework due to MacKay [5–7].

2 Method

We begin by re-parameterising our model such that the prior anisotropic Gaussian prior over the coefficients of the kernel expansion is replaced by an isotropic Gaussian prior over the transformed parameters, i.e. $\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} = \boldsymbol{\beta}^T \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector or transformed parameters. Let $\boldsymbol{R}$ represent the upper triangular Cholesky factor [8] of a symmetric positive-definite matrix $\boldsymbol{K}$, such that $\boldsymbol{K} = \boldsymbol{R}^T \boldsymbol{R}$. By inspection, the desired parameterisation is given then by

$$\boldsymbol{\beta} = \boldsymbol{R}\boldsymbol{\alpha} \quad \Longrightarrow \quad \boldsymbol{\alpha} = \boldsymbol{R}^{-1}\boldsymbol{\beta}.$$

The Gram matrix, $\boldsymbol{K}$ for a radial basis function kernel is at least in principle of full rank, assuming that $\boldsymbol{x}_i \neq \boldsymbol{x}_j$, $\forall\ i, j \in \{1, 2, \ldots, \ell\}$ [9]; however it is possible for $\boldsymbol{K}$ to be *numerically* rank-deficient in which case the Cholesky factor $\boldsymbol{R}$ becomes ill-conditioned. We therefore use the incomplete Cholesky factorisation with symmetric pivoting, due to Fine and Scheinberg [10], to construct the Cholesky factor $\hat{\boldsymbol{R}}$, of a numerically full-rank symmetric submatrix of $\boldsymbol{K}$. Without loss of generality, we assume that only the first W columns of $\boldsymbol{K}$ can be used to form $\hat{\boldsymbol{R}}$; the remaining columns are then linearly dependent, or close to being linearly dependent, on columns $1, 2, \ldots, W$, and can be deleted prior to training (c.f. [11]). The optimisation criterion then becomes

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{\ell} C\left(t_i, y(\boldsymbol{x}_i; \boldsymbol{\beta})\right) + \frac{\mu}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} = E_{\mathcal{D}} + \frac{\mu}{2}E_{\mathcal{W}}, \tag{2}$$

and the output of the re-parameterised model is given by

$$\text{logit}\{y(\boldsymbol{x}; \boldsymbol{\beta})\} = \boldsymbol{k}(\boldsymbol{x})\hat{\boldsymbol{R}}^{-1}\boldsymbol{\beta}, \qquad \text{where} \qquad \boldsymbol{k}(\boldsymbol{x}) = [\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x})]_{i=1}^{W}.$$

Again the optimal model parameters, $\boldsymbol{\beta}$, can be determined via the IRWLS procedure. Minimising the the criterion given in equation 2 is equivalent to maximising the posterior distribution

$$p(\boldsymbol{\beta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mu)}{p(\mathcal{D})} \tag{3}$$

where the likelihood is given by the Bernoulli distribution

$$p(\mathcal{D}|\boldsymbol{\beta}) = \prod_{i=1}^{\ell} y(\boldsymbol{x}_i; \boldsymbol{\beta})^{t_i}[1 - y(\boldsymbol{x}_i; \boldsymbol{\beta})]^{(1-t_i)},$$

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 133-138

and the prior over model parameters by a multivariate Gaussian distribution,

$$p(\boldsymbol{\beta}) = \left[\frac{\mu}{2\pi}\right]^{W/2} \exp\left\{-\frac{\mu}{2}\|\boldsymbol{\beta}\|^2\right\}.$$

The Taylor expansion of $L(\boldsymbol{\beta}, \mu)$ around the most probable value, $\boldsymbol{\beta}^{\mathrm{MP}}$, gives rise to familiar Gaussian approximation to the posterior distribution, known as the "Laplace approximation",

$$p(\boldsymbol{w}|\mathcal{D}) \approx \frac{1}{Z^*} \exp\left\{-L\left(\boldsymbol{\beta}^{\mathrm{MP}}\right) - \frac{1}{2}\Delta\boldsymbol{\beta}^T \boldsymbol{A} \Delta\boldsymbol{\beta}\right\}, \tag{4}$$

where z^* is an appropriate normalising constant, $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{MP}}$ and $\boldsymbol{A} = \nabla\nabla L(\boldsymbol{\beta}) = \nabla\nabla E_{\mathcal{D}} + \mu\boldsymbol{I}$ is the Hessian of $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. For further details, see e.g. Bishop [12].

2.1 The Evidence Approximation for μ

The evidence approximation of [5–7] assumes that the posterior distribution for the regularisation parameter, $p(\mu|\mathcal{D})$, is sharply peaked about its most probable value, μ^{MP}, suggesting the following approximation to the posterior distribution for $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}|\mathcal{D}) = \int p(\boldsymbol{\beta}|\mu, \mathcal{D}) p(\mu|\mathcal{D}) d\mu \approx p(\boldsymbol{\beta}|\mu^{\mathrm{MP}}, \mathcal{D}).$$

Thus, rather than integrate out the regularisation parameter entirely (e.g. Buntine and Weigend [13]), we simply proceed with the analysis using the regularisation parameter fixed at its most likely value. For a discussion of the validity of this approach, see MacKay [14]. We seek therefore to maximise the posterior distribution,

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}.$$

If the prior, $p(\mu)$ is relatively insensitive to the value μ, then maximising the posterior is approximately equivalent to maximising the likelihood term, $p(\mathcal{D}|\mu)$, known as the *evidence* for μ. Adopting the Gaussian approximation to the the posterior for the model parameters, the log-evidence is given by

$$\log p(\mathcal{D}|\mu) = -E_{\mathcal{D}}^{\mathrm{MP}} - \mu E_{\mathcal{W}}^{\mathrm{MP}} - \frac{1}{2}\log|\boldsymbol{A}| + \frac{W}{2}\log\mu. \tag{5}$$

Noting that $\boldsymbol{A} = \boldsymbol{H} + \mu\boldsymbol{I}$, where $\boldsymbol{H}$ is the Hessian of $E_{\mathcal{D}}$ with respect to $\boldsymbol{\beta}$, if the eigenvalues of $\boldsymbol{H}$ are $\lambda_1, \lambda_2, \ldots, \lambda_W$, then the eigenvalues of $\boldsymbol{A}$ are $(\lambda_1 + \mu), (\lambda_2 + \mu), \ldots, (\lambda_W + \mu)$. The derivative of $\log|\boldsymbol{A}|$ with respect to μ (assuming that the eigenvalues of $\boldsymbol{H}$ are independent of μ) is then given by

$$\frac{d}{d\mu}\log|\boldsymbol{A}| = \frac{d}{d\mu}\log\left\{\prod_{i=1}^{W}(\lambda_i + \mu)\right\} = \sum_{i=1}^{W}\frac{1}{\lambda_i + \mu}.$$

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 133-138

Setting the derivative of the log-evidence with respect to μ to zero, we have that

$$2\mu E_{\mathcal{W}}^{\mathrm{MP}} = W - \sum_{i=1}^{W} \frac{\mu}{\lambda_i + \mu} = \sum_{i=1}^{W} \frac{\lambda_i}{\lambda_i + \mu} = \gamma,$$

where γ is the number of well determined parameters in the model. This leads to a simple update formula for the regularisation parameter:

$$\mu^{\mathrm{new}} = \frac{\gamma}{2E_{\mathcal{W}}^{\mathrm{MP}}}. \tag{6}$$

The training procedure then alternates between updates of the primary model parameters using the IRWLS procedure and updates of the regularisation parameter according to equation 6.

3 Results

Figure 1 shows the (unmoderated) output of a Bayesian kernel logistic regression model, based on an isotropic radial basis function kernel, for the synthetic dataset described by Ripley [15]. The regularisation parameter, μ, was optimised via the update formula given by equations; the kernel parameter, η, was selected by maximising the marginal likelihood via a simple line search procedure. Clearly Bayesian kernel logistic regression is able to form a good model of the data, with little sign of over-fitting.

Table 1 presents the test set cross-entropy and error rate over six datasets for Bayesian and conventional kernel logistic regression models. The regularisation and kernel parameters for the conventional kernel logistic regression model were determined by minimisation of a ten-fold cross-validation [16] estimate of the cross-entropy criterion via the Nelder-Mead simplex optimisation algorithm [17]. Note that the differences in performance between the Bayesian and conventional kernel logistic regression model are not generally significant. However the model selection process for the Bayesian approach is somewhat less computationally expensive as the regularisation parameter is optimised by efficient update formula.

4 Conclusions

In this paper we have proposed a simple hierarchical Bayesian treatment of the kernel logistic regression model. The Bayesian approach is found to be competitive with conventional kernel logistic regression, but greatly reduces the computational expense of the model selection process. The key feature of this approach is that the model is re-parameterised such that an isotropic Gaussian prior over model parameters is obtained, facilitating simple implementation of MacKay's evidence approximation via standard methods. Note that this approach is quite general and could easily be applied to any kernel model minimising a regularised likelihood criterion.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 133-138

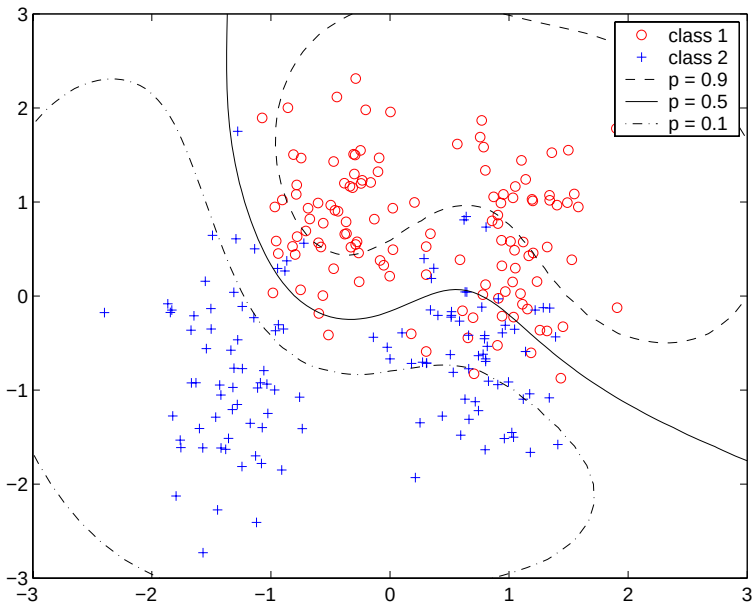

Figure 1: Output of a Bayesian kernel logistic regression (BKLR) model for Ripley's synthetic benchmark problem [15], the scale parameter of the RBF kernel chosen so as to maximise the marginal likelihood.

Table 1: Cross-entropy and error rate calculated over the test set for kernel logistic regression models with kernel and regularisation parameters determined via the evidence approximation and ten-fold cross-validation for six benchmark datasets.

Dataset	**Evidence**		**Cross-Validation**	
	xent	**error**	**xent**	**error**
Breast cancer	40.630	0.2597	40.674	0.2597
Diabetis	143.127	0.2400	143.158	0.2467
Pima	146.616	0.2018	146.215	0.2018
Synthetic	230.636	0.0950	230.281	0.0960
Thyroid	6.106	0.0267	3.274	0.0267
Titanic	1066.158	0.2292	1044.719	0.2292

Acknowledgements

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council (grant number 83/D17534).

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 133-138

References

[1] S. S. Keerthi, K. Duan, S. K. Shevade, and A. N. Poo. A fast dual algorithm for kernel logistic regression. In *Proceedings of the International Conference on Machine Learning*, 2002.

[2] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. John Wiley, New York, 1977.

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

[4] I. T. Nabney. Efficient training of RBF networks for classification. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, volume 1, pages 210–215, September 7–10 1999.

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

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

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

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

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

[10] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, December 2001.

[11] G. Baudat and F. Anouar. Kernel-based methods and function approximation. In *Proc. IJCNN*, pages 1244–1249, Washington, DC, July 2001.

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

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

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

[15] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.

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

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