ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Model Selection for Kernel Probit Regression

Gavin C. Cawley

University of East Anglia - School of Computing Sciences
Norwich, Norfolk NR4 7TJ - United Kingdom

Abstract. The convex optimisation problem involved in fitting a kernel probit regression (KPR) model can be solved efficiently via an iteratively re-weighted least-squares (IRWLS) approach. The use of successive quadratic approximations of the true objective function suggests an efficient approximate form of leave-one-out cross-validation for KPR, based on an existing exact algorithm for the weighted least-squares support vector machine. This forms the basis for an efficient gradient descent model selection procedure used to tune the values of the regularisation and kernel parameters. Experimental results are given demonstrating the utility of this approach.

1 Introduction

Assume we are given labelled training data $\mathcal{D} = \{(\boldsymbol{x_i}, t_i)\}_{i=1}^{\ell}$, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ represents a vector of attributes representing the i^{th} example and $t_i \in \{0,\ 1\}$ represents the desired class label. Kernel probit regression (c.f. [1]) aims to fit a probabilistic model of the form,

$$p(y_i = 1|\boldsymbol{x}_i) = \Phi\left\{\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b\right\} \qquad \text{where} \qquad \Phi\{z\} = \frac{1}{2}\left[1 + \operatorname{erf}\{z\}\right]$$

and $\operatorname{erf}(z) = 2\pi^{\frac{1}{2}} \int_0^z e^{-t^2} dt$ is the error function. The model is constructed in a feature space, $\mathcal{F}$, defined by a fixed non-linear transformation $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$. However, rather than specify the transformation explicitly, it is instead induced by a kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, giving the inner product between vectors in the feature space, i.e. $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$ (for a detailed introduction to kernel learning methods, see e.g. [2]). Any positive definite kernel function may be used, in this study we will adopt the spherical squared exponential kernel

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

where η is a kernel parameter controlling the sensitivity of the kernel. The optimal model parameters, $(\boldsymbol{w},\ b)$, are given by the solution of a convex penalised maximum likelihood cost function,

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \frac{\gamma}{2}\sum_{i=1}^{\ell}\left[t_i \log p_i + (1 - t_i)\log(1 - p_i)\right], \tag{2}$$

where γ is a regularisation parameter [3] controlling the bias-variance trade-off [4]. The representer theorem [5] states that the solution of an optimisation problem of this nature is given by an expansion over the training patterns of the form,

$$\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}) \qquad \Longrightarrow \qquad \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}).$$

The dual parameters of the kernel probit regression machine, $(\boldsymbol{\alpha},\ b)$ can be determined efficiently via an iteratively re-weighted least-squares (IRWLS) procedure [6]. We begin by forming a univariate quadratic approximation for each term comprising the negative log-likelihood,

$$l_i = -t_i \log p_i - (1 - t_i)\log(1 - p_i) \quad \text{where} \quad p_i = \Phi(z_i) \quad \text{and} \quad z_i = \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) + b.$$

The partial derivatives of l_i with respect to z_i are given by

$$\frac{\partial l_i}{\partial p_i} = \frac{t - p}{p(p-1)} \qquad \text{and} \qquad \frac{\partial p_i}{\partial z_i} = \frac{\exp\left\{-z_i^2\right\}}{\sqrt{\pi}}$$

$$\frac{\partial l_i}{\partial z_i} = \frac{\exp\left\{-z_i^2\right\}}{\sqrt{\pi}} \frac{t_i - p_i}{p_i(p_i - 1)}$$

and

$$\frac{\partial^2 l_i}{\partial p_i^2} = \frac{t_i}{p_i^2} + \frac{1-t}{(1-p_i)^2} \qquad \text{and} \qquad \frac{\partial p_i}{\partial z_i} = -\frac{\sqrt{2} z_i \exp\left\{-z_i^2\right\}}{\sqrt{\pi}}$$

$$\frac{\partial^2 l_i}{\partial z_i^2} = -z_i \exp\left\{-z_i^2\right\} \left[\frac{t_i}{p_i^2} + \frac{1-t}{(1-p_i)^2}\right] \frac{\sqrt{2}}{\sqrt{\pi}}$$

A local quadratic approximation of the regularised loss (2) is then given by

$$\tilde{L} = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{2}\sum_{i=1}^{\ell} \beta_i \left[\eta_i - z_i\right]^2 \tag{3}$$

where

$$\beta_i = \frac{\partial^2 l_i}{\partial z_i^2} \qquad \text{and} \qquad \eta_i = z_i - \frac{\partial l_i}{\partial z_i}\left[\frac{\partial^2 l_i}{\partial z_i^2}\right]^{-1}. \tag{4}$$

The quadratic approximation (3) represents a weighted least-squares problem, where the optimal model parameters are given by the solution of a system of linear equations,

$$\begin{bmatrix} \boldsymbol{K} + \gamma \boldsymbol{B} & \boldsymbol{1} \\ \boldsymbol{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{\eta} \\ 0 \end{bmatrix} \tag{5}$$

where $\boldsymbol{K} = [k_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{\ell}$ and $\boldsymbol{B} = \mathrm{diag}(\{\beta_1^{-1}, \beta_2^{-1}, \ldots, \beta_\ell^{-1}\})$ [7]. The training procedure then alternates between updates of the model parameters $(\boldsymbol{\alpha},\ b)$ via (5) and updates of the local quadratic approximation using (4). However, while an efficient training algorithm for the kernel probit regression model is easily implemented, a model selection scheme used to select good values for the kernel and regularisation parameters is less straightforward, and is the subject of the remainder of this paper.

2 Model Selection for Kernel Probit Regression

Cross-validation [8] commonly forms the basis for model selection schemes employed in practical applications of kernel learning methods. Under a k-fold cross-validation strategy, the data are partitioned into k disjoint subsets of approximately equal size. Models are then fitted on each of the k combinations of $k-1$ subsets, and the performance of each model evaluated on the unused subset in each fold. The average performance over the k trials generally provides a reliable estimate of performance on unseen data. The most extreme form of cross-validation, known as leave-one-out cross-validation [9, 10] partitions the data into ℓ subsets, each consisting of a single example. Fortunately, leave-one-out cross-validation can be implemented very efficiently in closed form for weighted least-squares based models (e.g. [11, 12]). These procedures provide the basis for an efficient approximate leave-one-out method for kernel probit regression, using the quadratic approximation of the true regularised loss used in the final iteration of the iteratively re-weighted least-squares procedure (c.f. [13, 14]). The matrix on the left-hand side of the system of linear equations (5) can be partitioned as follows,

$$\left[\begin{array}{cc} \boldsymbol{K}+\gamma\boldsymbol{B} & \boldsymbol{1} \\ \boldsymbol{1}^T & 0 \end{array}\right] = \left[\begin{array}{cc} c_{11} & \boldsymbol{c}_1^T \\ \boldsymbol{c}_1 & \boldsymbol{C}_1 \end{array}\right] = \boldsymbol{C}.$$

Let $[\boldsymbol{\alpha}^{(-i)}; b^{(-i)}]$ represent the parameters of the kernel probit regression model during the i^{th} iteration of the leave-one-out cross-validation procedure, then in the first iteration, in which the first training pattern is excluded,

$$\left[\begin{array}{c} \boldsymbol{\alpha}^{(-1)} \\ b^{(-1)} \end{array}\right] = \boldsymbol{C}_1^{-1}\left[\eta_2,\ldots,\eta_\ell,0\right]^T.$$

The leave-one-out prediction for the first training pattern is then given by,

$$\hat{z}_1^{(-1)} = \boldsymbol{c}_1^T \left[\begin{array}{c} \boldsymbol{\alpha}^{(-1)} \\ b^{(-1)} \end{array}\right] = \boldsymbol{c}_1^T\boldsymbol{C}_1^{-1}\left[\eta_2,\ldots,\eta_\ell,0\right]^T$$

Considering the last ℓ equations in the system of linear equations (5), it is clear that $[\boldsymbol{c}_1\ \boldsymbol{C}_1]\left[\alpha_2,\ldots,\alpha_\ell,b\right]^T = \left[\eta_2,\ldots,\eta_\ell,0\right]^T$, and so

$$\hat{z}_1^{(-1)} = \boldsymbol{c}_1^T\boldsymbol{C}_1^{-1}\left[\boldsymbol{c}_1\ \boldsymbol{C}_1\right]\left[\boldsymbol{\alpha}^T,b\right]^T = \boldsymbol{c}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{c}_1\alpha_1 + \boldsymbol{c}_1\left[\alpha_2,\ldots,\alpha_\ell,b\right]^T.$$

Noting, from the first equation in the system of linear equations (5), that $\eta_1 = c_{11}\alpha_1 + \boldsymbol{c}_1^T\left[\alpha_2,\ldots,\alpha_\ell,b\right]^T$, thus

$$\hat{z}_1^{(-1)} = \eta_1 - \alpha_1\left(c_{11} - \boldsymbol{c}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{c}_1\right)$$

Finally, via the block matrix inversion lemma,

$$\left[\begin{array}{cc} c_{11} & \boldsymbol{c}_1^T \\ \boldsymbol{c}_1 & \boldsymbol{C}_1 \end{array}\right]^{-1} = \left[\begin{array}{cc} \kappa^{-1} & -\kappa^{-1}\boldsymbol{c}_1\boldsymbol{C}_1^{-1} \\ \boldsymbol{C}_1^{-1}+\kappa^{-1}\boldsymbol{C}_1^{-1}\boldsymbol{c}_1^T\boldsymbol{c}_1\boldsymbol{C}_1^{-1} & -\kappa^{-1}\boldsymbol{C}_1^{-1}\boldsymbol{c}_1^T \end{array}\right],$$

where $\kappa = c_{11} - \boldsymbol{c}_1^T \boldsymbol{C}_1^{-1} \boldsymbol{c}$, and noting that the system of linear equations (5) is insensitive to permutations of the ordering of the equations and of the unknowns, we have that,

$$\hat{z}_i^{(-i)} = \eta_i - \frac{\alpha_i}{\boldsymbol{C}_{ii}^{-1}}. \tag{6}$$

This means that, assuming the system of linear equations (5) is solved via explicit inversion of $\boldsymbol{C}$, an approximate leave-one-out cross-validation estimate of the test cross-entropy can be evaluated using information already available as a by-product of training the least-squares support vector machine on the entire dataset. This approximation is based on the assumption that the quadratic approximation of the regularised loss function is unchanged during the leave-one-out cross-validation procedure (c.f. [13]). The partial derivatives of $z_i^{(-i)}$ with respect to the kernel and regularisation parameters are easily obtained. This provides the basis for an efficient model selection scheme, based on minimisation of the approximate leave-one-out cross-validation estimate of the test cross-entropy, via scaled conjugate gradient descent [15].

3 Results

Table 1 illustrates the generalisation performance of kernel probit regression algorithms with leave-one-out and conventional k-fold cross-validation based model selection schemes, over the suite of thirteen benchmark datasets used in the study by Mika *et al.* [16]. Results obtained using a range of state-of-the-art classifiers are also displayed for comparison. The expectation-propagation based Gaussian Process classifier (e.g. [17]) is of particular interest, as it represents a classifier with the same basic structure as the KPR model, with similar design goals. Each benchmark consists of 100 random partitionings (20 in the case of `image` and `splice`) of the data to form the training and test sets for each trial. Model selection was performed separately in each trial, in order to avoid any possibility of selection bias. The use of multiple realisations of the data also allows the use of significance tests, via the z-score. Comparing leave-one-out and 5-fold cross-validation based selection methods for the kernel probit regression model, neither model is significantly better than the other on any of the thirteen benchmarks, at the 95% level. The KPR with leave-one-out cross-validation based model selection is significantly superior to the EP-GPC on four benchmarks (`ringnorm`, `splice`, `twonorm` and `waveform`) and statistically inferior on only two (`banana` and `image`). The computational expense of the proposed approximate leave-one-out cross-validation method is however negligible as it can be computed as a by-product of the training algorithm.

4 Conclusions

In this paper, we have presented an efficient model selection procedure for kernel probit regression models, based on a closed-form approximation of the leave-one-out cross-validation estimate of the test cross-entropy. The partial derivatives

Table 1: Error rates of state-of-the-art classifiers over thirteen benchmarks, kernel probit regression using the proposed leave-one-out model selection procedure (LOO-KPR), kernel probit regression with conventional 5-fold cross-validation based model selection (KPR), Gaussian process classifier implemented via the expectation propagation algorithm (EP-GPC). The results for the support vector machine (SVM) kernel Fisher discriminant classifier (KFD), radial basis function (RBF) network, AdaBoost (AB) and regularised AdaBoost (AB$_R$) are taken from the study by Mika *et al.* [16]. The results for the EP-GPC were obtained using the MATLAB software by Rasmussen and Williams [17]. The results 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 datasets), along with the associated standard error. The best results are shown in boldface and the second best in italics.

Dataset	**LOO-KPR**	**KPR**	**EP-GPC**	**SVM**	**KFD**	**RBF**	**AB**	**AB$_R$**
Banana	10.6 ± 0.06	*10.5 ± 0.05*	**10.4 ± 0.05**	11.5 ± 0.07	10.8 ± 0.05	10.8 ± 0.06	12.3 ± 0.07	10.9 ± 0.04
Breast cancer	26.8 ± 0.48	26.6 ± 0.49	26.5 ± 0.49	*26.0 ± 0.47*	**25.8 ± 0.46**	27.6 ± 0.47	30.4 ± 0.47	26.5 ± 0.45
Diabetes	23.4 ± 0.18	23.5 ± 0.17	*23.3 ± 0.18*	23.5 ± 0.17	**23.2 ± 0.16**	24.3 ± 0.19	26.5 ± 0.23	23.8 ± 0.18
Flare solar	34.3 ± 0.17	34.3 ± 0.18	34.2 ± 0.21	**32.4 ± 0.18**	*33.2 ± 0.17*	34.4 ± 0.20	35.7 ± 0.18	34.2 ± 0.22
German	*23.5 ± 0.21*	**23.4 ± 0.22**	**23.4 ± 0.21**	23.6 ± 0.21	23.7 ± 0.22	24.7 ± 0.24	27.5 ± 0.25	24.3 ± 0.21
Heart	16.6 ± 0.31	16.6 ± 0.31	16.7 ± 0.29	**16.0 ± 0.33**	*16.1 ± 0.34*	17.6 ± 0.33	20.3 ± 0.34	16.5 ± 0.35
Image	3.1 ± 0.13	3.2 ± 0.13	*2.8 ± 0.12*	3.0 ± 0.06	3.3 ± 0.06	3.3 ± 0.06	**2.7 ± 0.07**	**2.7 ± 0.06**
Ringnorm	*1.6 ± 0.02*	*1.6 ± 0.02*	4.4 ± 0.06	1.7 ± 0.01	**1.5 ± 0.01**	1.7 ± 0.02	1.9 ± 0.03	*1.6 ± 0.01*
Splice	11.1 ± 0.18	11.1 ± 0.17	11.6 ± 0.18	10.9 ± 0.07	10.5 ± 0.06	*10.0 ± 0.1*	10.1 ± 0.05	**9.5 ± 0.07**
Thyroid	*4.4 ± 0.22*	4.8 ± 0.21	*4.4 ± 0.22*	4.8 ± 0.22	**4.2 ± 0.21**	4.5 ± 0.21	*4.4 ± 0.22*	4.6 ± 0.22
Titanic	*22.5 ± 0.11*	22.7 ± 0.13	22.6 ± 0.13	*22.4 ± 0.10*	23.2 ± 0.20	23.3 ± 0.13	22.6 ± 0.12	22.6 ± 0.12
Twonorm	2.9 ± 0.03	2.9 ± 0.03	3.1 ± 0.03	3.0 ± 0.02	**2.6 ± 0.02**	2.9 ± 0.03	3.0 ± 0.03	*2.7 ± 0.02*
Waveform	*9.9 ± 0.04*	**9.8 ± 0.04**	10.1 ± 0.05	*9.9 ± 0.04*	*9.9 ± 0.04*	10.7 ± 0.11	10.8 ± 0.06	**9.8 ± 0.08**

of this criterion with respect to the hyper-parameters are easily computed, permitting the use of efficient scaled conjugate-gradient optimisation methods. An extensive evaluation over thirteen benchmark datasets reveals this approach to be comparable with conventional k-fold cross-validation based methods in terms of generalisation. Furthermore, the kernel probit regression model is also competitive with the state-of-the-art Gaussian process classifier, based on the expectation propagation algorithm.

References

[1] C. M. Bishop. *Pattern Recognition amd Machine Learning.* Springer, 2006.

[2] J. Shawe-Taylor and N. Cristianini. *Kernel methods for Pattern Analysis.* Cambridge University Press, 2004.

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

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

[5] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[6] 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, Edinburgh, United Kingdom, September 7–10 1999.

[7] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vanderwalle. *Least Squares Support Vector Machines.* World Scientific Publishing, 2002.

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

[9] P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, February 1968.

[10] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition (in Russian). *Techicheskaya Kibernetica*, 3, 1969.

[11] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression.* Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1982.

[12] S. Weisberg. *Applied linear regression.* John Wiley and Sons, New York, second edition, 1985.

[13] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models - A Roughness Penalty Approach*, volume 58 of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC, 1994.

[14] G. C. Cawley and N. L. C. Talbot. Efficient model selection for kernel logistic regression. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR-2004)*, volume 2, pages 439–442, Cambridge, United Kingdom, August 23–26 2004.

[15] P. M. Williams. A Marquardt algorithm for choosing the step-size in backpropagation learning with conjugate gradients. Cognitive Science Research Paper CSRP-229, University of Sussex, Brighton, U.K., February 1991.

[16] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. J. Smola, and K.-R. Müller. Invariant feature extraction and classification in feature spaces. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 526–532. MIT Press, 2000.

[17] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* Adaptive Computation and Machine Learning. MIT Press, 2006.