Efficient Cross-Validation of Kernel Fisher Discriminant Classifiers

Gavin C. Cawley and Nicola L. C. Talbot

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

Abstract.
Mika *et al.* [1] introduce a non-linear formulation of the Fisher discriminant based the well-known "kernel trick", later shown to be equivalent to the Least-Squares Support Vector Machine [2, 3]. In this paper, we show that the cross-validation error can be computed very efficiently for this class of kernel machine, specifically that leave-one-out cross-validation can be performed with a computational complexity of only $\mathcal{O}(\ell^3)$ operations (the same as that of the basic training algorithm), rather than the $\mathcal{O}(\ell^4)$ of a direct implementation. This makes leave-one-out cross-validation a practical proposition for model selection in much larger scale applications of KFD classifiers.

1 Introduction

Assume we are given training data $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell\} = \{\mathcal{X}_1, \mathcal{X}_2\} \subset \mathbb{R}^d$, where $\mathcal{X}_1 = \{\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \ldots, \boldsymbol{x}_{\ell_1}^1\}$ is a set of patterns belonging to class $\mathcal{C}_1$ and similarly $\mathcal{X}_2 = \{\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \ldots, \boldsymbol{x}_{\ell_2}^2\}$ is a set of patterns belonging to class $\mathcal{C}_2$; Fisher's linear discriminant (FLD) attempts to find a linear combination of input variables, $\boldsymbol{w} \cdot \boldsymbol{x}$, that maximises the average separation of the projections of points belonging to $\mathcal{C}_1$ and $\mathcal{C}_2$, whilst minimising the within class variance of the projections of those points. The Fisher discriminant is given by the vector $\boldsymbol{w}$ maximising

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}, \tag{1}$$

where $\boldsymbol{S}_B$ is the between class scatter matrix $\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$, $\boldsymbol{m}_j = \ell_j^{-1} \sum_{i=1}^{\ell_j} \boldsymbol{x}_i^j$ and $\boldsymbol{S}_W$ the within class scatter matrix

$$\boldsymbol{S}_W = \sum_{i \in \{1,2\}} \sum_{j=1}^{\ell_i} (\boldsymbol{x}_j^i - \boldsymbol{m}_i)(\boldsymbol{x}_j^i - \boldsymbol{m}_i)^T.$$

The innovation introduced by Mika *et al.* [1] is to construct Fisher's linear discriminant in a fixed feature space $\mathcal{F}$ ($\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{F}$) induced by a positive definite *Mercer* kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defining the inner product $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$ (see e.g. Cristianini and Shawe-Taylor [4]). Let the kernel matrices for the entire dataset, $\boldsymbol{K}$, and for each class, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ be defined as follows:

$$\boldsymbol{K} = [k_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{\ell} \quad \text{and} \quad \boldsymbol{K}_i = [k_{jk}^i = \mathcal{K}(\boldsymbol{x}_j, \boldsymbol{x}_k^i)]_{j,k=1}^{j=\ell, k=\ell_i}.$$

The theory of reproducing kernels indicates that $\boldsymbol{w}$ can then be written as an expansion of the form

$$\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i). \tag{2}$$

The objective function (1) can also be written such that the data $\boldsymbol{x} \in \mathcal{X}$ appear only within inner products, giving

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{N} \boldsymbol{\alpha}}, \tag{3}$$

where $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^{\ell}$, $\boldsymbol{M} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$, $\boldsymbol{m}_i = \boldsymbol{K}_i \boldsymbol{u}_i$, $\boldsymbol{u}_i$ is a column vector containing ℓ_i elements with a common value of ℓ_i^{-1} and

$$\boldsymbol{N} = \sum_{i \in \{1,2\}} \boldsymbol{K}_i (\boldsymbol{I} - \boldsymbol{U}_i) \boldsymbol{K}_i^T,$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{U}_i$ is a matrix with all elements equal to ℓ_i^{-1}. The coefficients, $\boldsymbol{\alpha}$, of the expansion (2) are then given by the leading eigenvector of $\boldsymbol{N}^{-1}\boldsymbol{M}$. Note that $\boldsymbol{N}$ is likely to be singular, or at best ill-conditioned, and so a regularised solution is obtained by substituting $\boldsymbol{N}_\mu = \boldsymbol{N} + \mu \boldsymbol{I}$, where μ is a regularisation constant. To complete the kernel Fisher discriminant classifier, $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b$, the bias, b, is given by

$$b = -\boldsymbol{\alpha} \frac{\ell_1 \boldsymbol{M}_1 + \ell_2 \boldsymbol{M}_2}{\ell}.$$

Xu *et al.* [3] show that the parameters of the kernel Fisher discriminant classifier are also given by the solution of the following system of linear equations:

$$\begin{bmatrix} \boldsymbol{K}\boldsymbol{K} + \mu \boldsymbol{I} & \boldsymbol{K}\boldsymbol{1} \\ (\boldsymbol{K}\boldsymbol{1})^T & \ell \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{K} \\ \boldsymbol{1} \end{bmatrix} \boldsymbol{y}, \tag{4}$$

where $\boldsymbol{1}$ is a column vector of ℓ ones and $\boldsymbol{y}$ is a column vector with elements $y_i = \ell/\ell_j \; \forall i : \boldsymbol{x}_i \in \mathcal{X}_j$. This illustrates the similarities between the kernel Fisher discriminant and the least-squares support vector machine (LS-SVM) [2]. The kernel Fisher discriminant (KFD) classifier has been shown experimentally to demonstrate near state-of-the-art performance on a range of artificial and real world benchmark datasets [1] and so is worthy of consideration for small to medium scale applications. In this paper we present an efficient algorithm for approximate cross-validation of kernel Fisher discriminant models, providing a practical criterion for model selection.

2 Method

The system of linear equations (4) can be written more concisely in the form

$$\left[\begin{array}{c} \boldsymbol{\alpha} \\ b \end{array}\right] = \left[\boldsymbol{R} + \boldsymbol{Z}^T\boldsymbol{Z}\right]^{-1}\boldsymbol{Z}^T\boldsymbol{y},$$

where $\boldsymbol{Z} = [\boldsymbol{K}\ \boldsymbol{1}]$ and $\boldsymbol{R} = \mathrm{diag}([\mu\boldsymbol{1}\ 0])$ (n.b. this is very similar to the set of *normal equations* to be solved in multi-variate linear regression). At each step of the leave-one-out cross-validation procedure, a kernel Fisher discriminant classifier is constructed excluding a single training pattern from the data. The vector of model parameters, $\{\boldsymbol{\alpha}_{(i)}, b_{(i)}\}$ at the i^{th} iteration is then given by the solution of a modified system of linear equations,

$$\left[\begin{array}{c} \boldsymbol{\alpha}_{(i)} \\ b_{(i)} \end{array}\right] = \left[\boldsymbol{R} + \boldsymbol{Z}_{(i)}^T\boldsymbol{Z}_{(i)}\right]^{-1}\boldsymbol{Z}_{(i)}^T\boldsymbol{y}$$

where $\boldsymbol{Z}_{(i)}$ is the sub-matrix formed by omitting the i^{th} row of $\boldsymbol{Z}$. Normally the most computationally expensive step is the inversion of the matrix $\boldsymbol{C}_{(i)} = \left[\boldsymbol{R} + \boldsymbol{Z}_{(i)}^T\boldsymbol{Z}_{(i)}\right]$, with a complexity of $\mathcal{O}(\ell^3)$ operations. Fortunately $\boldsymbol{C}_{(i)}$ can be written as a rank one modification of a matrix $\boldsymbol{C}$,

$$\boldsymbol{C}_{(i)} = \left[\boldsymbol{R}_{(i)} + \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{z}_i\boldsymbol{z}_i^T\right] = \left[\boldsymbol{C} - \boldsymbol{z}_i\boldsymbol{z}_i^T\right], \tag{5}$$

where $\boldsymbol{z}_i$ is the i^{th} row of $\boldsymbol{Z}$. The following matrix inversion lemma then allows $\boldsymbol{C}_{(i)}^{-1}$ to be found in only $\mathcal{O}(\ell^2)$ operations, given that $\boldsymbol{C}^{-1}$ is already known:

Lemma 1 (Bartlett Matrix Inversion Formula) *Given an invertible matrix $\boldsymbol{A}$ and column vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, then assuming $1 - \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \neq 0$,*

$$\left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}. \tag{6}$$

This is known as the Bartlett matrix inversion formula [5].

Applying the Bartlett formula to the matrix inversion problem given in (5), we have that

$$\boldsymbol{C}_{(i)}^{-1} = [\boldsymbol{C} - \boldsymbol{z}_i\boldsymbol{z}_i^T]^{-1} = \boldsymbol{C} + \frac{\boldsymbol{C}^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{C}^{-1}}{1 - \boldsymbol{z}_i^T\boldsymbol{C}^{-1}\boldsymbol{z}_i}.$$

The computational complexity of the leave-one-out cross-validation process is thus reduced to only $\mathcal{O}(\ell^3)$ operations, which is the same as that of the basic training algorithm for the kernel Fisher discriminant classifier. In the case of S-fold cross-validation, ℓ/S applications of the Bartlett correction formula (6) are performed in each trial to "erase" one of S disjoint sets of ℓ/S training patterns from the kernel Fisher discriminant classifier trained on the entire dataset, again resulting in a computational complexity of $\mathcal{O}(\ell^3)$ operations.

2.1 A Further Refinement

For model selection purposes, we are not principally concerned with the values of the model parameters themselves, but only statistics such as the leave-one-out error rate

$$E = \frac{1}{\ell}\text{card}\{i : y_i(\boldsymbol{w}_{(i)} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b_{(i)}) \leq 0\},$$

or equivalently

$$E = \frac{1}{\ell}\text{card}\{i : \text{sign}(y_i)\{r_{(i)}\}_i \leq -1\},$$

where $\left\{r_{(i)}\right\}_i = \text{sign}(y_i) - \boldsymbol{w}_{(i)} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b_{(i)}$ is the residual error for the i^{th} training pattern during the i^{th} iteration of the leave-one-out cross-validation procedure. It can be shown that

$$\left\{r_{(i)}\right\}_i = \frac{1}{1 - h_{ii}} r_i.$$

where $r_i = \text{sign}(y_i) - \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b$ is the residual error for the i^{th} training pattern for a kernel Fisher discriminant classifier trained on the entire dataset, $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T$ is the *hat* matrix of which h_{ii} is the i^{th} element of the leading diagonal. The leave-one-out error rate can thus be evaluated in closed form without explicit inversion of $\boldsymbol{C}_{(i)}$ $\forall i \in \{1, 2, \ldots, \ell\}$, with a computational complexity of only $\mathcal{O}(\ell^3)$ operations.

3 Results

The proposed approximate leave-one-out cross-validation method is evaluated over a series of randomly generated synthetic datasets. In each case, approximately one quarter of the data belong to class $\mathcal{C}_1$ and three-quarters to class $\mathcal{C}_2$. The patterns comprising class $\mathcal{C}_1$ are drawn from a bivariate Normal distribution with zero mean and unit variance. The patterns forming class $\mathcal{C}_2$ form an annulus; the radii of the data are drawn from a normal distribution with a mean of 4 and unit variance, and the angles uniformly distributed. The datasets vary in size between 10 and 500 patterns. Figure 1 shows a graph of run-time as a function of the number of training patterns for direct and fast approximate leave-one-out cross-validation methods. Clearly the fast approximate method is considerably faster and exhibits significantly better scaling properties than the direct implementation. Let the relative approximation error be defined as

$$e = \frac{\|\tilde{\boldsymbol{r}} - \hat{\boldsymbol{r}}\|^2}{\|\tilde{\boldsymbol{r}}\|^2},$$

where $\tilde{\boldsymbol{r}}$ is a vector of leave-one-out residual errors computed via the direct approach and $\hat{\boldsymbol{r}}$ is the corresponding vector of residual errors resulting from the proposed approximation. Figure 2 shows a graph of the mean relative approximation error, as a function of the number of training patterns. The approximation error is small for datasets of more than ≈ 30 training patterns.

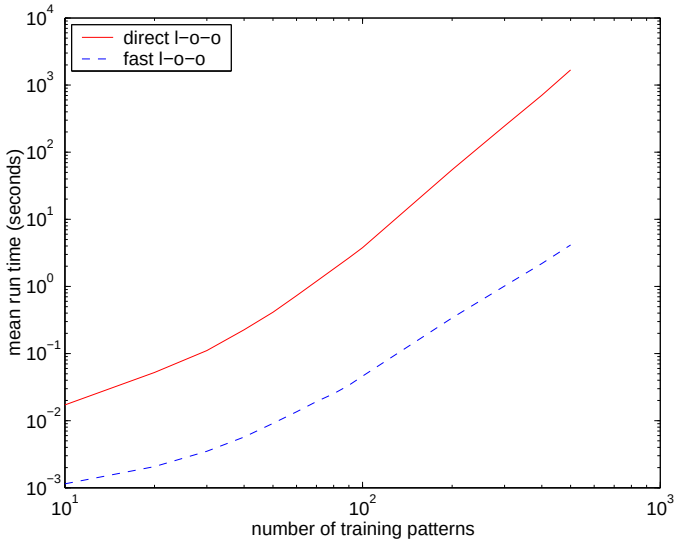

Figure 1: Graph of run-time as a function of the number of training patterns for leave-one-out cross-validation of kernel Fisher discriminant classifiers via direct and fast approximate methods (mean of 20 trials).

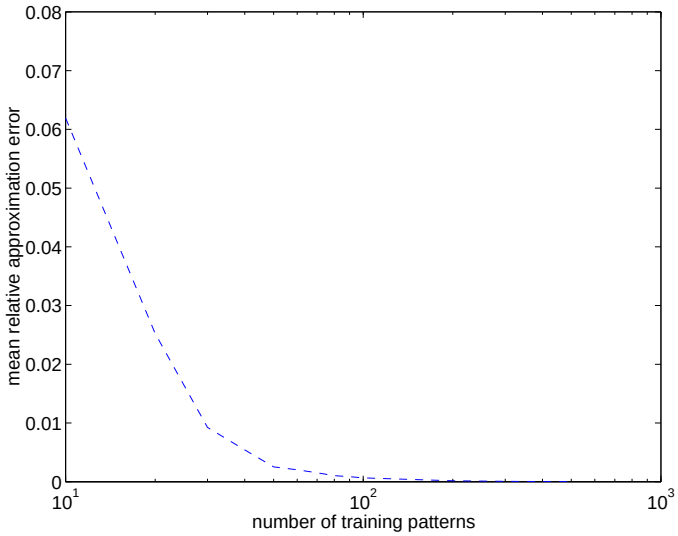

Figure 2: Graph of the mean relative approximation error as a function of the number of training patterns for the proposed fast approximate leave-one-out cross-validation method (mean of 20 trials).

4 Summary

In this paper we have generalised an existing algorithm for leave-one-out cross-validation of multi-variate linear regression models (see e.g. [6]) to provide an approximation to the leave-one-out error rate of kernel Fisher discriminant classifiers. The proposed algorithm allows approximate leave-one-out cross-validation of this class of model with a computational complexity of only $\mathcal{O}(\ell^3)$ operations, instead of the $\mathcal{O}(\ell^4)$ of a direct approach. Furthermore, profiling information reveals that, providing $\boldsymbol{C}^{-1}$ is cached, the time taken to estimate the leave-one-out error rate is considerably *less* than the time taken to train the KFD classifier. As a result leave-one-out cross-validation becomes a practical model selection criterion in far larger scale applications of KFD models.

5 Acknowledgements

The authors would like to thank Rob Foxall for his helpful comments on previous drafts of this manuscript. This work was supported by the Royal Society (grant number RSRG-22270).

References

[1] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, volume IX, pages 41–48. IEEE Press, New York, 1999.

[2] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, June 1990.

[3] J. Xu, X. Zhang, and Y. Li. Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR. In *Proc. IJCNN*, pages 1486–1491, Washington, DC, July 2001.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge University Press, Cambridge, U.K., 2000.

[5] M. S. Bartlett. An inverse matrix adjustment arising in disciminant analysis. *Annals of Mathematical Statistics*, 22(1):107–111, 1951.

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