A variable metric probabilistic k-nearest-neighbours classifier

Richard M. Everson and Jonathan E. Fieldsend

Department of Computer Science, University of Exeter, UK

Revision: 1.14
12:06 30th March 2004

Abstract The k-nearest neighbour (k-nn) model is a simple, popular classifier. Probabilistic k-nn is a more powerful variant in which the model is cast in a Bayesian framework using (reversible jump) Markov chain Monte Carlo methods to average out the uncertainy over the model parameters.
The k-nn classifier depends crucially on the metric used to determine distances between data points. However, scalings between features, and indeed whether some subset of features is redundant, are seldom known *a priori*. Here we introduce a variable metric extension to the probabilistic k-nn classifier, which permits averaging over all rotations and scalings of the data. In addition, the method permits automatic rejection of irrelevant features. Examples are provided on synthetic data, illustrating how the method can deform feature space and select salient features, and also on real-world data.

1 Introduction

One of the most popular methods of statistical classification is the k-nearest neighbour model (k-nn). Although the method has a ready statistical interpretation, and has been shown to have an asymptotic error rate no worse than twice the Bayes error rate [1], it appears in symbolic AI under the guise of case-based reasoning. The method is essentially geometrical, assigning the class of an unknown exemplar to the class of the majority of its k nearest neighbours in some training data. More precisely, in order to assign to a datum $\mathbf{x} \in \mathbb{R}^D$ a class y, given the known training data $\mathcal{D} = \{y_i, \mathbf{x}_i\}_{i=1}^N$, the k-nn method first calculates the distances $d_i = ||\mathbf{x} - \mathbf{x}_i||$. If there are Q classes, each of which is *a priori* equally likely, the probability that $\mathbf{x}$ belongs to the q-th class is then evaluated as $p(y \,|\, \mathbf{x}, k, \mathcal{D}) = k_q/k$, where k_q is the number of the k data points with the smallest d_i belonging to class q.

Classification thus crucially depends upon the metric used to determine the distances d_i. Usual practice is to normalise the data so that each of the D coordinates has, say, unit variance, after which Euclidean distance is used. However, as others have shown and we illustrate here, this may result in suboptimal classification rates. In this paper we use a variable metric of the form

$$d(\mathbf{x}_1, \mathbf{x}_2) = \{(\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{M} (\mathbf{x}_1 - \mathbf{x}_2)\}^{1/2} \tag{1}$$

where $\mathbf{M}$ is a $D \times D$ symmetric matrix. Rather than seek a single, optimal metric we adopt the Bayesian point of view and average over all metrics each weighted by its posterior probability. Essential to this programme is Holmes & Adams' [2] recent recasting of the k-nn model in a Bayesian framework, which we briefly describe in the remainder of this section. In section 2 we describe how the k-nn model may be augmented with a variable metric. Illustrative results are presented in section 3 and the paper concludes with a brief discussion.

1.1 The probabilistic k-nn model

Holmes & Adams [2] have extended the traditional k-nn classifier by adding a parameter β which controls the 'strength of association' between neighbours. The likelihood of the data given parameters $\boldsymbol{\theta} = \{k, \beta\}$ is defined as

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{\exp[\beta \sum_{\mathbf{x}_j \sim \mathbf{x}_i}^{k} u(d(\mathbf{x}_i, \mathbf{x}_j))\delta_{y_i y_j}]}{\sum_{q=1}^{Q} \exp[\beta \sum_{j \sim i}^{k} u(d(\mathbf{x}_i, \mathbf{x}_j))\delta_{q y_j}]}. \tag{2}$$

Here δ_{mn} is the Kronecker delta and $\sum_{\mathbf{x}_j \sim \mathbf{x}_i}^{k}$ means the sum over the k nearest neighbours of $\mathbf{x}_i$ (excluding $\mathbf{x}_i$ itself). If the non-increasing function of distance $u(\cdot) = 1/k$, then the term $\sum_{\mathbf{x}_j \sim \mathbf{x}_i}^{k} u(d(\mathbf{x}_i, \mathbf{x}_j))\delta_{y_i y_j}$ counts the fraction of nearest neighbours of k in the same class y_i as $\mathbf{x}_i$. In the work reported here we choose u to be the tricube kernel [3] which gives decreasing weight to distant neighbours.

Holmes & Adams implement an efficient reversible jump Markov chain Monte Carlo (RJMCMC) [4,5] scheme to draw samples $\boldsymbol{\theta}^{(t)} = \{k^{(t)}, \beta^{(t)}\}$ from the posterior distribution of the parameters $p(\boldsymbol{\theta} \,|\, \mathcal{D})$. Uncertainty in k and β when classifying $\mathbf{x}$ can then be taken into account by averaging over all values of k and β:

$$p(y \,|\, \mathbf{x}, \mathcal{D}) = \int p(y \,|\, \mathbf{x}, \boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta} \,|\, \mathcal{D})\, d\boldsymbol{\theta} \approx \frac{1}{T} \sum_{t=1}^{T} p(y \,|\, \mathbf{x}, \boldsymbol{\theta}^{(t)}, \mathcal{D}) \tag{3}$$

where the predictive likelihood is

$$p(y \,|\, \mathbf{x}, \boldsymbol{\theta}, \mathcal{D}) = \frac{\exp[\beta \sum_{\mathbf{x}_j \sim \mathbf{x}}^{k} u(d(\mathbf{x}, \mathbf{x}_j))\delta_{y y_j}]}{\sum_{q=1}^{Q} \exp[\beta \sum_{\mathbf{x}_j \sim \mathbf{x}}^{k} u(d(\mathbf{x}, \mathbf{x}_j))\delta_{q y_j}]}. \tag{4}$$

2 Variable metric and feature selection

The relative scales on which features are measured are not usually clear *a priori* and the standard practice of normalisation to zero mean and unit variance may be detrimental to the overall classification rate. Many classifiers, such as linear discriminators or neural networks, discriminate on the basis of linear or nonlinear weighted combinations of the feature variables; these weights are adjusted during learning and may thus compensate for improper scaling of the data. The k-nn

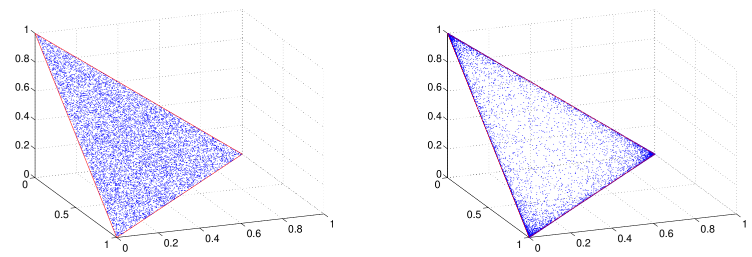

Figure 1. Samples from 3-dimensional Dirichlet distribution. *Left:* Non-informative, $Dir(1, 1, 1)$. *Right:* Sparse distribution, $Dir(0.2, 0.2, 0.2)$.

classifier contains no such implicit scaling, making classifications on the basis of the exemplars within a spherical region of feature space. In order to compensate for this, we therefore adopt a metric of the form (1).

It is useful to write $\mathbf{M} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with non-negative entries λ_d on the diagonal, and $\mathbf{Q}$ is an orthogonal matrix. Then writing $\hat{\mathbf{x}} = \boldsymbol{\Lambda}^{1/2}\mathbf{Q}\mathbf{x}$ shows that using metrics of this form is equivalent to rotating and scaling the data before using the standard Euclidean distance.

The $\mathbf{M} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$ decomposition is convenient for setting priors as well as for computation. We augment the parameters $\boldsymbol{\theta}$ of the probabilistic k-nn classifier with $\mathbf{Q}$ and $\boldsymbol{\lambda} = \mathrm{diag}(\boldsymbol{\Lambda})$, and extend the Holmes & Adams RJMCMC sampler to draw samples from the joint posterior density $p(k, \beta, \boldsymbol{\lambda}, \mathbf{Q} \,|\, \mathcal{D})$. The likelihoods (2) and (4) are unchanged except that $d(\cdot, \cdot)$ depends upon $\boldsymbol{\lambda}$ and $\mathbf{Q}$.

Priors. As in the probabilistic k-nn model [2], we adopt a uniform prior $p(k) = \min(1/k_{max}, N)$, with $k_{max} = 250$. The prior on β expresses a mild preference for small β: $p(\beta) = 2\mathcal{N}(0, 100)I(\beta > 0)$, where $I(\cdot)$ is the set indicator function.

Since we usually have no *a priori* preference for a particular rotation, a uniform prior over $\mathbf{Q}$ is appropriate. The prior for $\boldsymbol{\lambda}$ is taken to be a uniform Dirichlet density:

$$p(\boldsymbol{\lambda}) = Dir(\boldsymbol{\lambda} \,|\, \alpha, \ldots, \alpha). \tag{5}$$

This prior ensures that $\lambda_d \geq 0$ so that $\mathbf{M}$ is non-negative definite; changes of sign in the scalings are irrelevant to the classification. In addition, only relative scales are important, which is enforced by fact that the weights λ_d are confined to a $(D-1)$-dimensional simplex: $\sum_{d=1}^{D} \lambda_d = 1$. The parameter α determines the shape of the prior distribution. As Figure 1 shows, when $\alpha = 1$ the prior is non-informative so that all scalings (points on the simplex) are equally likely. Setting $\alpha > 1$ encodes a belief that the scalings should be the same for each variable. Here it is more useful to set $\alpha < 1$ which reflects a belief that the scale factors are unequal, although as Figure 1 illustrates, no particular feature is favoured. If particular features or (with rotations by $\mathbf{Q}$) linear combinations of features are likely to be irrelevant, setting $\alpha < 1$ is tantamount to a sparse prior over the scalings, leading to the suppression of irrelevant feature combinations. Here we have used a mildly sparse prior, $\alpha = 0.8$.

Sampling. The Holmes & Adams RJMCMC sampler makes reversible proposals (k', β') from the current (k, β). To sample from the full parameter set we make additional proposals $\mathbf{Q}'$ and $\boldsymbol{\lambda}'$; the proposal $(k', \beta', \boldsymbol{\lambda}', \mathbf{Q}')$ being accepted with probability:

$$\min \left\{ 1, \frac{p(\mathcal{D} \,|\, k', \beta', \boldsymbol{\lambda}', \mathbf{Q}')}{p(\mathcal{D} \,|\, k, \beta, \boldsymbol{\lambda}, \mathbf{Q})} \frac{p(\boldsymbol{\lambda}')}{p(\boldsymbol{\lambda})} \frac{p(\beta')}{p(\beta)} \frac{p(\boldsymbol{\lambda} \,|\, \boldsymbol{\lambda}')}{p(\boldsymbol{\lambda}' \,|\, \boldsymbol{\lambda})} \right\}. \tag{6}$$

Proposals to change the scaling are made from a Dirichlet whose expected value is the current $\boldsymbol{\lambda}$; thus $\boldsymbol{\lambda}' \sim Dir(c\boldsymbol{\lambda})$ [6]. Proposals are thus centred on $\boldsymbol{\lambda}$ with their spread controlled by c, which is chosen during burn-in to achieve a satisfactory acceptance rate; in the work reported here $c = 400$. The proposal ratio for $\boldsymbol{\lambda}$ can be shown to be

$$\frac{p(\boldsymbol{\lambda} \,|\, \boldsymbol{\lambda}')}{p(\boldsymbol{\lambda}' \,|\, \boldsymbol{\lambda})} = \prod_{d=1}^{D} \frac{(\lambda'_d)^{c\lambda_d - 1} \Gamma(c\lambda'_d)}{(\lambda_d)^{c\lambda'_d - 1} \Gamma(c\lambda_d)}. \tag{7}$$

To ensure that $\mathbf{Q}$ is orthogonal, we use Cayley coordinates in which $\mathbf{Q}$ is represented as the matrix exponential $\mathbf{Q} = \exp(\mathbf{S})$ of a skew-symmetric matrix $\mathbf{S} = -\mathbf{S}^T$. Consequently $\mathbf{S}$ has $D(D-1)/2$ independent entries corresponding to the $D(D-1)/2$ degrees of freedom in a D-dimensional orthogonal matrix.

Proposals $\mathbf{Q}'$ are generated by perturbing $\mathbf{S}$ as follows: $\mathbf{S}' = \mathbf{S} + \mathbf{R} - \mathbf{R}^T$, where $R_{ij} \sim \mathcal{N}(0, \sigma^2)$. The variance of the perturbations to $\mathbf{S}$ is adjusted during burn-in to achieve a satisfactory acceptance rate. If the perturbations to $\mathbf{S}$ are symmetric, it is straightforward to show that $p(\mathbf{Q}' \,|\, \mathbf{Q}) = p(\mathbf{Q} \,|\, \mathbf{Q}')$, so there is no contribution to the acceptance probability (6) from the $\mathbf{Q}$ proposal ratio.

3 Illustration

We first illustrate the ability of the variable metric to deal with scaled, rotated and irrelevant data by applying it to synthetic data. We then apply it to well-known problems from the UCI machine learning repository [7].

The ability of the variable metric to adjust and compensate for improperly scaled data can be illustrated by the data show in Figure 2. These two-dimensional, two-class data are separated by an elliptical boundary which is rotated with respect to the coordinate axes; there are 500 training and 284 testing examples. Although the variances of the data projected onto either of the features are equal, optimal classification is achieved if the data are rotated and scaled so that the class boundary is roughly circular.

As an initial illustration we consider the ellipse data, but with the axes of the ellipse aligned with the coordinate axes. As shown in Table 1, the probabilistic k-nn has a misclassification rate of 2.82%, whereas for the variable metric method the error is lowered to 1.76%. The mean posterior scaling factors are $\bar{\lambda}_1 = 0.37$ and $\bar{\lambda}_2 = 0.63$, which is close to the ratio $1:2$, the ratio of ellipse axes.

Simple scaling of the data is insufficient to render the class boundary circular when the class boundary ellipse is not aligned with the coordinate axes (see

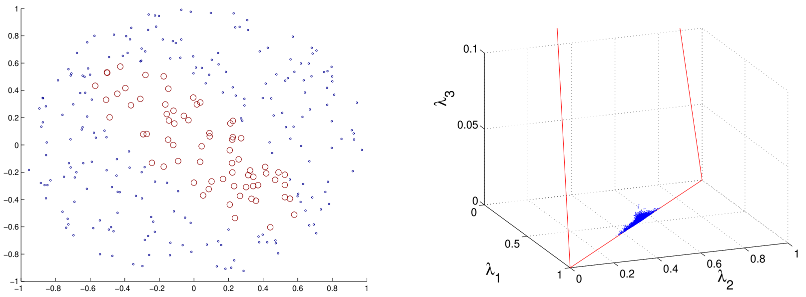

Figure 2. *Left:* Synthetic data with rotated elliptical class boundaries. *Right:* Samples from posterior λ for synthetic rotated data, with an additional irrelevant feature.

Table 1. Mean classification error (%) for probabilistic and variable metric k-nn.

	Ellipse			UCI		
	Scaled	Rotated	Irrel.	Ion.	Pima	Wine
MAP $\beta = 1$ k-nn	2.82	2.82	6.69	7.94	26.04	4.49
probabilistic k-nn	2.82	2.82	4.23	5.30	22.92	3.37
variable metric k-nn	1.76	1.41	1.41	1.99	21.35	1.12

Figure 2); and the misclassification error is 3.16% for a variable metric method but with $\mathbf{Q} = \mathbf{I}$ fixed. However, permitting $\mathbf{Q}$ to vary recovers the lower error rate, and the mean posterior $\mathbf{Q}$ corresponds to rotations of $\pm 45°$.

As a final illustration on synthetic data, we add to the ellipse data a third variable x_3 which is irrelevant to classification. The right hand panel of Figure 2 shows samples from posterior distribution of λ_d. It can be see that the irrelevant direction, corresponding here to λ_3, has been effectively suppressed so that the posterior $\boldsymbol{\lambda}$ is very close to the two-dimensional simplex $\lambda_1 + \lambda_2 = 1$.

We also evaluate the performance of the variable metric k-nn classifier on three well-known data sets [7]. The Ionosphere dataset comprises 33 inputs, with 200 training and 151 test examples. The Pima indian diabetes data has 8 predictive variables. *A priori* choice of a metric here is difficult as these variables are measured in disparate units (e.g., kg/m^2, years, mm Hg). They were all normalised to zero mean and unit variance before classification. There were 576 training and 192 test samples. The Wine recognition data set has 13 continuous features on the basis of which an instance should be assigned to one of three classes. There is no standard training/test split for these data, so they were partitioned at random into training and test sets, each with 89 examples.

Following 10^5 burn-in MCMC steps, 10^4 samples (every 7th step) were collected for classification. Table 1 compares mean classification rates for the standard probabilistic k-nn method, the *maximum a posteriori* standard probabilistic k-nn classifier with β held at 1 and the variable metric model. The additional flexibility in the variable metric method clearly permits better classification, achieving rates at least has high as those reported elsewhere, e.g., [8].

4 Conclusion

We have presented a straightforward scheme for learning a metric for the probabilistic k-nn classifier. Results on synthetic data and real data show that the variable metric method yields improved mean classification rates and is able to reject irrelevant features. The variable metric methods presented here are particularly pertinent for k-nn classifiers, which otherwise have no mechanism for learning the data scaling, but the method can be applied to other classifiers. Although the variable metric method yields improved classification rates, it is computationally more expensive than fixed metric methods because distances must be recomputed for each $(\boldsymbol{\lambda}^{(t)}, \mathbf{Q}^{(t)})$ sample.

Here we have considered only linear scalings and rotations of the data. Current work involves expanding the class of metrics to include Minkowski metrics in the rotated space. Finally, we remark that it would be valuable to extend the considerable work (e.g., [9]) on *local* variable metrics to the Bayesian context.

Acknowledgements

We thank Trevor Bailey, Adolfo Hernandez, Wojtek Krzanowski, Derek Partridge, Vitaly Schetnin and Jufen Zhang for their helpful comments. JEF gratefully acknowledges support from the EPSRC, grant GR/R24357/01.

References

1. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13** (1967) 21–27
2. Holmes, C., Adams, N.: A probabilistic nearest neighbour method for statistical pattern recognition. Journal Royal Statistical Society B **64** (2002) 1–12 See also code at `http://www.stats.ma.ic.ac.uk/~ccholmes/Book_code/book_code.html`.
3. Fan, J., Gijbels, I.: Local polynomial modelling and its applications. Chapman & Hall, London (1996)
4. Green, P.: Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika **82** (1995)
5. Denison, D., Holmes, C., Mallick, B., Smith, A.: Bayesian Methods for Nonlinear Classification and Regression. Wiley (2002)
6. Larget, B., Simon, D.: Markov Chain Monte Carlo Algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution **16** (1999) 750–759
7. Blake, C., Merz, C.: UCI repository of machine learning databases (1998) `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
8. Sykacek, P.: On input selection with reversible jump Markov chain Monte Carlo sampling. In Solla, S., Leen, T., Müller, K.R., eds.: NIPS* 12. (2000) 638–644
9. Henley, W., Hand, D.: A k-nearest-neighbour classfier for assessing consumer credit risk. The Statistician **45** (1996) 77–95