Classification via kernel regression based on univariate product density estimators

Bezza Hafidi[1], Abdelkarim Merbouha[2], and Abdallah Mkhadri[1⋆]

[1] Department of Mathematics,
Cadi Ayyad University, BP 3290 Marrakech, Moroco
(e-mail: b.hafidi@ucam.ac.ma, mkhadri@ucam.ac.ma)
[2] Department of Mathematics
FST-Beni-Mellal, Morocco
(e-mail: merbouhak@yahoo.fr)

Abstract. We propose a nonparametric discrimination method based on a nonparametric Nadaray-Watson kernel regression type-estimator of the posterior probability that an incoming observed vector is a given class. To overcome the curse of dimensionality of the multivariate kernel density estimate, we introduce a variance stabilizing approach which constructs independent predictor variables. Then, the multivariate kernel estimator is replaced by the univariate kernel product estimators. The new procedure is illustrated in simulated data sets and real example, confirming the usefulness of our approach.
Keywords: Classification, Density estimation, Kernel regression, Component Principal Analysis.

1 Introduction

The basic problem in classification is to assign an unknown subject to one of K groups $G_1, \ldots, G_K$ on the basis of a multivariate observation $\mathbf{x} = (x_1, \ldots, x_p)^t$, where p represents the number of variables and t denotes the transpose operation. However, in practice, the form of class-conditional densities is seldom known. To overcome this problem, one can consider a nonparametric classification method, which uses a nonparametric multivariate kernel density estimates instead of the parametric densities.
Indeed, recently much attention has been given to the application of nonparametric procedures in the classification problem, which have been shown to exhibit superior performance over standard parametric methods such as linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA) in a wide variety of problems. The recent book of [Hastie *et al.*, 2001] presents an excellent overview of nonparametric classification methods. A disadvantage of such models may be a lack of parsimony in the final model and a sensitivity to the "curse of dimensionality" when the dimension p is large and the sample sizes are moderate.

⋆ The third author is supported by TWAS Research grant 01-159 RG/MATHS/AF/AC.

Two semiparametric alternative models for classification, which are a generalization of the model assumed by LDA and QDA, are recently proposed by [Cooley and MacEachern, 1998] and [Amato *et al.*, 2003]. This generalization relies upon a transformation of the data based on pseudo-independent variables. Then, the multivariate kernel density estimates are replaced by the univariate product kernel estimators. [Cooley and MacEachern, 1998] used principal component analysis (PCA) to obtain a transformation matrix, while [Amato *et al.*, 2003] considered independent component analysis (ICA) (cf. [Comon, 1994]).
In this paper, we propose a nonparametric discrimination method based on a nonparametric Nadaray-Watson kernel regression type-estimator of the posterior probability that an incoming observed vector is a given class. To overcome the curse of dimensionality we introduce a Cooley and MacEachern's variance stabilizing approach which constructs independent predictor variables. Then, the multivariate kernel density estimates is replaced by product of univariate kernel estimators. Some theoretical result on Bayes risk consistence is discussed.
This article is organized as follows. In Section 2, we briefly review the nonparametric classification rules which product indirect estimation of the conditional group probability (or a posteriori probability). We also recall the classification approach based on univariate product density estimators which is an alternative interpretation of LDA and QDA. Section 3 is devoted to our new variance stabilizing kernel regression classification approach. Some theoretical asymptotic result on Bayes consistency is discussed in the same section. In Section 4, we apply our new classification rule to some simulations data sets and a real example, confirming the usefulness of our approach. Section 5 ends with some conclusions.

2 Nonparametric classification rules

The multiple classification problem is well studied in statistics. Typically, there is a qualitative random variable Y that takes on a finite number K of values which we refer to as groups: $G_1, \ldots, G_K$. To assign an individual to one of K distinct groups, we must build an allocation rule from the training sample $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$, where $\mathbf{x_i} \in I\!R^p$ is the observation vector and $y_i \in \{1, \ldots, K\}$ indicates the a priori group membership of $\mathbf{x_i}$.
As is well known, the optimal classification rule $d(\mathbf{x})$ allocates an observed p-variate vector $\mathbf{x}$ via

$$d(\mathbf{x}) = \text{argmax}_{j=1,\ldots,K} I\!P(Y = y_j|\mathbf{x}), \tag{1}$$

where

$$I\!P(Y = y_j|\mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\sum_{i=1}^{K} \pi_i f_i(\mathbf{x})} \tag{2}$$

is the a posteriori probability of group j (or the conditional group probability), with π_j and $f_j(.)$ the a priori probability and group-conditional density of group j, respectively. In practice classification rules are constructed either by combining (1) with (2) and estimating the group densities f_j or by estimating directly the *a posteriori* probability $P(Y = y_j|\mathbf{x})$ from the given data. The first approach is called generative method, while the latter approach is called discriminative method.

2.1 Generative nonparametric rules

Most important parametric and nonparametric generative classification rules based on the direct estimation of group densities are Gaussian discriminant analysis (GDA) and kernel density classification, respectively. In kernel density classification the group-conditional densities are estimated with multivariate kernel density estimators which have the form

$$\hat{f}_j(\mathbf{x}) = \frac{1}{n_j} \sum_{\ell=1}^{n_j} \mathcal{K}(\mathbf{x} - \mathbf{x}_{j\ell}, H_j),$$

where $n_j = \#\{i : y_i = j\}, \{\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jn_j}\}$ is the training sample of group j, $\mathcal{K}(., H_j)$ denotes a multivariate kernel function from $\mathbb{R}^p$ to $\mathbb{R}$, and H_j is a usually a p-dimensional vector of smoothing parameters that governs the degree of smoothness of the estimate (cf. [Scott, 1992]). The recent book of [Hastie *et al.*, 2001] presents an excellent overview of new nonparametric classification methods. A disadvantage of such models may be a lack of parsimony in the final model and a sensitivity to the "curse of dimensionality" when the dimension p is large and the sample sizes are moderate.

2.2 Kernel univariate product estimators

In order to avoid the biased tail estimation and the curse of dimensionality common to multivariate kernel density estimation, [Cooley and MacEachern, 1998] (see also [Amato *et al.*, 2003]) present an alternative view of QDA and LDA which allows them to extend the nonparametric classification problem. In this alternative rotations of the coordinate axes are employed to obtain an assumed mutual independence among the components of the rotated data. Then, the conditional density of the kth sample group can be written as the product of univariate Gaussian density on the transformed sample, i. e.

$$f_k(\mathbf{x}) = f(H_k\mathbf{x}) = \prod_{j=1}^{p} \frac{1}{\sigma_{jk}} \phi\left(\frac{(H_k\mathbf{x})_j - (H_k\mu_k)_j}{\sigma_{jk}}\right), \tag{3}$$

where $\phi(.)$ denotes the density of a standard normal variable and H_k is the transform matrix obtained from the spectral decomposition of the covariance

matrix Σ_k. Then, a natural generalization of LDA and QDA is to replace the univariate Gaussian densities with univariate kernel density, we called kernel product estimator (KPE) the resulting estimator. From the latter algorithm, QDA and LDA are therefore just affected by the way the H_k are estimated. [Cooley and MacEachern, 1998] consider the principal component analysis (PCA) to estimate H_k, while [Amato *et al.*, 2003] propose to use independent component analysis (ICA).

3 Classification via kernel regression estimator

There are several compelling reasons for using discriminative rather than generative classifiers. The first one is that in real world problems the assumed generative model is rarely exact, and asymptotically a discriminative model should typically be preferred (cf. [Vapnick, 1998]). Moreover, there are many problems in which direct classification does not suffice, and where the precise estimation of the conditional group probabilities is most important. Multiple logistic regression (polychotomous regression) has been used for a long time (cf. [Hosmer and Lemeshow, 1989]) to obtain a direct estimate of all the conditional group probabilities.
On the other hand, little is known about nonparametric kernel discriminative method. One early direct kernel approach was proposed by [Lauder, 1983], which is analogue to kernel density estimation. [Hoti and Holmström, 1999] proposed an analogue Nadaray-Watson type-estimator defined by

$$\hat{r}_n^{(k)}(\mathbf{x}) = \frac{\sum_{i=1}^n T_i^{(k)} K((\mathbf{x}-\mathbf{x}_i)/h_n)}{\sum_{i=1}^n K((\mathbf{x}-\mathbf{x}_i)/h_n)} \tag{4}$$

where $T_i^{(k)} = 1$ if $Y_i = k$ and 0 elsewhere, $\mathcal{K}()$ is a multivariate kernel and $k = 1, \ldots, K$. They further improve the flexibility of the estimator by replacing the constants Y_i^j with locally fitted polynomial functions.

3.1 Regression kernel classification method (RKCM)

We attack the problem of curse of dimensionality of the kernel regression classification method, defined via the Nadaray-Watson type-estimator (4), by adapting the Cooley and MacEachern's variance stabilizing approach to KRCM. It consists to replace in (4) the multivariate kernel density estimator by the product of univariate kernel density estimators, which leads to the new estimator

$$\tilde{r}_n^{(k)}(\mathbf{x}) = \frac{\sum_{i=1}^n T_i^{(k)} \prod_{j=1}^p \hat{f}_{kj}^*\{(\hat{H}_k\mathbf{x})_j - (\hat{H}_k\mathbf{x_i})_j\}}{\sum_{i=1}^n \prod_{j=1}^p \hat{f}_{kj}^*\{(\hat{H}_k\mathbf{x})_j - (\hat{H}_k\mathbf{x_i})_j\}}, \tag{5}$$

where $\hat{f}_{kj}^*(z) = \sum_{\ell:y_\ell=k}^n \mathcal{K}\{(z-(\hat{H}_k\mathbf{X}_{k\ell})_j)/h_{kj}\}/h_{kj}n_k$ is the univariate kernel density estimate in the jth dimension of the transformed space for group

k. We allow the pooling of sample covariance information across K groups to obtain $\hat{H}_1 = \ldots = \hat{H}_K = \hat{H}$. Then the common transformation matrix $\hat{H}$ is estimated via the application of PCA on the pooling sample covariance matrix.
Since many kernel functions are highly efficient, we adopt the Gaussian kernels which are widely used (cf. [Farhmeir and Tutz, 1994], pp. 156-157). For simplicity, we can assume that the smoothing parameter in direction j for group k is constant and equal h. Then, the classical cross-validation of the average squared error criterion is often used for the selection of the smoothing parameter h. But, the cross-validation of the misclassification error rate is more convenient in our context, since it is related to discriminant problem. However, in our experimental study, we fix $h_{kj} = 0.9\sigma_{kj}n^{-1/(p+4)}$ as in [Cooley and MacEachern, 1998] ($k = 1, \ldots, K; j = 1, \ldots, p$). A robust estimation of σ_{kj} can be taken equal to the smaller of the sample standard deviation and (1/1.34) x sample interquartile range. This choice is mainly related to density estimation, but it is simple to compute and seems to work well in our numerical study.

3.2 Consistence and convergence rate

[Cooley and MacEachern, 1998] showed that the rule based on KPE of the density of the kth group is consistent on the set $\mathbb{R}^p - \mathcal{N}_k$, where $\mathcal{N}_k$ is a set of Lebesgue measure 0 ($k = 1, \ldots, K$). Moreover, they established that the rate of convergence of the mean integrated squared error to 0 is $O(n^{-4/5})$, regardless of the dimensionality p.
For our KRCM, we have established that the rule based on the regression kernel product estimator $g_n(.)$ is Bayes risk universally consistent, i.e. $lim_{n\longrightarrow\infty}\mathbb{P}\{g_n(\mathbf{X}|D_n)\} - L^* = 0$ for any distribution of the pair $(\mathbf{X}, Y)$, where L^* is the optimal Bayes error probability and D_n denote the training sample of size n. The proof is based on the verification of the three conditions of the general Stone's theorem (cf. [Devroye *et al.*, 1996], Theorem 6.3 in page 98). For saving space, this proof is not included in this note.

4 Numerical experiments

In this section, we report on some case studies for analyzing the practical behavior of KRCM relative to LDA, QDA and KPE on the basis of training and test error rate, respectively. For purposes of comparison, the smoothing parameter in direction j and group k is fixed equal to $h_{kj} = 0.9\sigma_{kj}n^{-1/(4+p)}$ for KRCM and KPE, and a priori probabilities were taken to be equal. As indicated in Section 3, σ_{kj} is estimated by the smaller of the sample standard deviation and (1/1.34) x sample interquartile range. We first present some Monte Carlo numerical experiments on simulated data sets, then we present numerical experiment on real data set.

4.1 Monte Carlo numerical experiments

The first simulated example was also considered by [Cooley and MacEachern, 1998], and has two groups and two predictors. The final predictors are combination of two initial predictors, generated from the normal mixture for the first initial predictor and the standard normal for the second one. The difference between the groups lies in the means of the normals in the mixture distribution of the first predictor (cf.[Cooley and MacEachern, 1998]).
Two hundred and fifty sets for the training and test samples consisted of 100 and 900 observations, respectively, were run from an equal mixture of the two distributions. Table 1 shows the averaged success rates for the training data set and the test data set over 200 simulations, with the standard error of the average in the parentheses. It appears that KRCM performs well than KPE, QDA and LDA, in both training and test sample respectively.
In the second example the optimal boundaries separating the group are non-additive functions of the predictors. The observations of the two groups are described by 6 predictors, the last four of which are random $\mathcal{N}(0,1)$ noise variables for both groups. The first two predictors of group 1 are independent Uniform$[-5,5]$ random variables, whereas the first two variables of group 2 form bivariate normal vectors with means 0, variance 1 and correlation coefficient 1/2. Similar example appears in [Cooley and MacEachern, 1998], where all relevant discriminatory information is contained in a relatively small dimension.
We select a training sample of size 500 and a test sample of size 3000, both from an equal mixture of the two populations. For both the training data set and the test data set, the averaged success rates and their standard errors over 75 replicates are summarized in Table 1. The behavior of KRCM is similar to that in the first example, where the difference with KPE is more important (15% on test data).
The third example is a well-known *waveform* problem composed of three groups with 21 predictors. The predictors are defined by

$$\begin{aligned} x_i &= uh_1(i) + (1-u)h_2(i) + \epsilon_i \qquad \text{Group1} \\ x_i &= uh_1(i) + (1-u)h_3(i) + \epsilon_i \qquad \text{Group2} \\ x_i &= uh_2(i) + (1-u)h_3(i) + \epsilon_i \qquad \text{Group3,} \end{aligned}$$

where $i = 1, \ldots, 21$, u is uniform on $[0,1]$, $\epsilon_i \sim \mathcal{N}(0,1)$ and the h_i are the shifted triangular forms defined by: $h_1(i) = \max(6 - [i-11], 0), h_2(i) = h_1(i-4)$ and $h_3(i) = h_1(i+4)$.
The training and test sets consisted of 500 and 300 observations, respectively, are selected and their averaged success rates are shown in Table 1 where equal prior are used. Again, KRCM is better than QDA, LDA and KPE.

	Mixture data		Nonadditive boundary		Waveform	
Method	Train	Test	Train	Test	Train	Test
LDA	62.92(.050)	59.23(.015)	84.32(.018)	50.62(.012)	97.72(.005)	97.44(.008)
QDA	61.73(.046))	59.22(.013)	85.13(.021)	74.72(.054)	97.95(.007)	96.25(.017)
KPE	78.14(.043)	76.23(.015)	85.39(.023)	84.75(.012)	91.22(.002)	93.89(.003)
KRCM	83.17(.034)	77.31(.015)	99.92(.001)	99.04(.002)	100(.000)	100(.000)

Table 1. Average success rates and standard deviation in parentheses.

4.2 Real data example

The real data set considered is the Diabetes in Pima Indian Women. It is described for instance in [Ripley, 1996]. It concerns a population of $n = 532$ women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. This women described by 7 predictors and two groups. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The training set contains a randomly selected set of 200 subjects, and the sample test set contains the remaining 332 subjects.
For both the training data set and the test data set, the success rates is summarized in Table 2. Here again, the behavior of KRCM is better than all the other methods.

	Pima	
Method	Train	Test
LDA	76.000	77.108
QDA	76.500	69.879
KPE	85.000	81.626
KRCM	99.000	99.698

Table 2. Success rates corresponding to Pima data set.

5 Discussion

In this paper, we propose a nonparametric discrimination method based on a nonparametric Nadaray-Watson kernel regression type-estimator of the posterior probability that an incoming observed vector is a given class. To overcome the curse of dimensionality we introduce a Cooley and MacEachern's variance stabilizing approach which constructs independent predictor variables. Then, the multivariate kernel density estimates is replaced by product of univariate kernel estimators.

Summarizing results experiments, performance of KRCM is very good compared with KPE, LDA and QDA. Consequently, our study confirms that using discriminative rather than generative classifiers is preferred.

References

[Amato *et al.*, 2003]U. Amato, A. Antoniadis, and Gregoire. Independent component discriminant analysis. *Int. Math. J.*, pages 727–734, 2003.

[Comon, 1994]P. Comon. Independent component analysis, a new concept. *Signal Processing*, pages 187–314, 1994.

[Cooley and MacEachern, 1998]C. A. Cooley and MacEachern. Classification via kernel product estimators. *Biometrika*, pages 823–833, 1998.

[Devroye *et al.*, 1996]L. Devroye, L. Gyorfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New-York, 1996.

[Farhmeir and Tutz, 1994]L. Farhmeir and G. Tutz. *Multivariate statistical modelling based generalized linear models.* Springer-Verlag, New-York, 1994.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer-Verlag, New-York, 2001.

[Hosmer and Lemeshow, 1989]D.W. Hosmer and S. Lemeshow. *Applied logistic regression.* Wiley, New-York, 1989.

[Hoti and Holmström, 1999]F. Hoti and L. Holmström. Reduced kernel regression for fast classification. *P. Brenner, L. Arkeryd, J. Rergh and R. Pettersson*, pages 405–412, 1999.

[Lauder, 1983]I. J. Lauder. Direct kernel assessment of diagnostic probabilities. *Biometrika*, pages 254–256, 1983.

[Ripley, 1996]B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[Scott, 1992]D.W. Scott. *Multivariate density estimation: theory, practice and visualization.* Wiley, New-York, 1992.

[Vapnick, 1998]V. N. Vapnick. *Statistical learning theory.* John Wiley & Sons, 1998.