A Comparative Study of Kernel and Classical Methods in Supervised Learning

Marcelo R. P. Ferreira[1], Getúlio José Amorim do Amaral[2]

[1] Departamento de Estatística, CCEN, UFPB
[2] Departamento de Estatística, CCEN, UFPE

Abstract Methods based on kernel density estimation have been used in a wide variety of real-world discrimination problems. This work reviews some classical statistical methods that are frequently used in supervised learning: logistic regression, k-Nearest Neighbour, normal based linear and quadratic classifiers; and a non-parametric one: the kernel classifier. Applications with real data sets are used to compare the classification methods. Our results show that the kernel method outperforms the classical approachs in many situations.

Keywords: Kernel density estimation, Kernel density classification, Classification, Misclassification rate.

1 Introduction

Consider data $\{\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n\}$, where $\underset{\sim}{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, as a realization of a random sample, and let an element of the set $\{f_j(\underset{\sim}{x}), j = 1, \ldots, J\}$ be the density associated with $\underset{\sim}{x}_i$. Let π_j, $j = 1, \ldots, J$, be the classes' prior probabilities, *i.e.* $\pi_j = P(\underset{\sim}{x}_j \in \Pi_j)$ where Π_j denotes the jth class. Then, using Bayes' Theorem, the posterior probability of the observation $\underset{\sim}{x}_i$ being from the jth class, is

$$P(\underset{\sim}{x}_j \in \Pi_j | \underset{\sim}{x}_i = \underset{\sim}{x}) = \frac{\pi_j f_j(\underset{\sim}{x})}{\sum_{j=1}^{J} \pi_j f_j(\underset{\sim}{x})}.$$

According to Bayes' formula, we allocate an observation to the class with highest posterior probability:

$$\underset{\sim}{x} \text{ is allocated to the class } \Pi_j \text{ if } \Pi_j = \underset{j \in \{1,\ldots,J\}}{\arg\max}\ \pi_j f_j(\underset{\sim}{x}).$$

Often the prior probabilities π_j are known, or simply estimated using $\hat{\pi}_j = n_j/n$, $j = 1, \ldots, J$, with $\sum_{j=1}^{J} n_j = n$. Classical parametric approachs make assumptions about the densities f_j. Usually, the data is assumed to have a normal distribution, however, this assumption is very restrictive. With non-parametric discriminant analysis we relax this assumption and thus are able to tackle more complex cases.

The kernel approach for discrimination is to estimate the density f_j of each class Π_j and allocate an observation according to the rule:

$$\underset{\sim}{x} \text{ is allocated to the class } \Pi_j \text{ if } \Pi_j = \underset{j \in \{1,\ldots,J\}}{\arg\max}\ \hat{\pi}_j \hat{f}_j(\underset{\sim}{x}),$$

where $\hat{f}_j(\underset{\sim}{x})$ is the kernel density estimate corresponding to the jth class.

The kernel density estimator of f at the point $\underset{\sim}{x} \in \mathbb{R}^p$ is (see [1, 3] for further details)

$$\hat{f}(\underset{\sim}{x}) = \hat{f}(\underset{\sim}{x}; \boldsymbol{H}) = n^{-1} \sum_{i=1}^{n} K_{\boldsymbol{H}}(\underset{\sim}{x} - \underset{\sim}{x}_i),$$

where the scale factor $\boldsymbol{H}$ is a symmetric positive definite $p \times p$ matrix called the smoothing parameter or bandwidth matrix, and $K_{\boldsymbol{H}} = |\boldsymbol{H}|^{-1} K(\boldsymbol{H}^{-1}\underset{\sim}{x})$, where $K : \mathbb{R}^p \to \mathbb{R}$ is called the kernel; usually K is a symmetric probability density function.

2 Numerical Results

In this section, we will present some numerical results with real and simulated data sets. The real data set (labelled "salmon data") obtained from [2] contains information on growth ring diameters (freshwater and marine water) of 100 salmon fish coming from Alaskan or Canadian water. A random sample ($n = 50$) was used as training sample and the remaining observations were used as test sample. The simulated data set (labelled "synthetic data") is a two-class classification problem. We generate training and testing samples, both with size $n = 100$, from normal mixture densities:

$$\Pi_1 : f_1 \sim \frac{1}{2} N \left(\begin{bmatrix} -\frac{3}{2} \\ -\frac{3}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right) + \frac{1}{2} N \left(\begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right)$$

$$\Pi_2 : f_2 \sim \frac{1}{2} N \left(\begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right) + \frac{1}{2} N \left(\begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right)$$

The kernel discriminant (KD) was compared with the normal linear (LD) and quadratic discriminants (QD). The misclassification rates are shown in the following table. The results show that the kernel classifier have better performance than the other methods.

Table 1: Misclassification rates on test samples

	Misclassification rate (%)	
Discriminant	salmon data	synthetic data
LD	10	31
QD	10	35
KD	08	19

References

[1] Duong, T.: Bandwidth Selectors for Multivariate Kernel Density Estimation. PhD Thesis, University of Western Australia, School od Mathematics and Statistics. (2004)

[2] Johnson, R. A. & Wichern, D. W.: Applied Multivariate Statistical Analysis. Prentice-Hall, New York. (1998)

[3] Scott, D. W.: Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York. (1992)