Journal of Multivariate Analysis 115 (2013) 108–121

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Comparison of binary discrimination methods for high dimension low sample size data

A. Bolivar-Cime [a,*], J.S. Marron [b]

[a] Department of Probability and Statistics, CIMAT, Jalisco S/N, Col. Valenciana, CP 36240, Guanajuato, Gto, Mexico

[b] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260, USA

ARTICLE INFO

Article history:
Received 10 June 2011
Available online 12 October 2012

AMS 2000 subject classifications:
primary 62H30
secondary 62E20

Keywords:
Asymptotic analysis
Binary discrimination
High dimensional data
Machine learning

ABSTRACT

A comparison of some binary discrimination methods is done in the high dimension low sample size context for Gaussian data with common diagonal covariance matrix. In particular we obtain results about the asymptotic behavior of the methods Support Vector Machine, Mean Difference (i.e. Centroid Rule), Distance Weighted Discrimination, Maximal Data Piling and Naive Bayes when the dimension d of the data sets tends to infinity and the sample sizes of the classes are fixed. It is concluded that, under appropriate conditions, the first four methods are asymptotically equivalent, but the Naive Bayes method can have a different asymptotic behavior when d tends to infinity.

1. Introduction

This work deals with *binary discrimination analysis* in the High-Dimension, Low Sample Size (HDLSS) framework for Gaussian data. We focus on the study of asymptotic behavior of the following methods for two-class discrimination: Support Vector Machine (SVM), Mean Difference (MD), Distance Weighted Discrimination (DWD), Maximal Data Piling (MDP) and Naive Bayes (NB). The HDLSS asymptotics of the first three methods have been previously studied in Hall et al. [7], where the probability of correct classification of a new data point is considered when the dimension d of the training data sets tends to infinity for fixed sample sizes of the classes. The present paper takes a different asymptotic viewpoint, based on the angle between the normal vectors of the separating hyperplane. We find conditions that characterize both consistency and strong inconsistency. Previous comparison of these methods has been done by simulations in Marron et al. [10], and Ahn and Marron [1]. A further contribution of the present paper is theoretical analysis of some empirical phenomena observed in [10,1], by specifically studying the asymptotic behavior of the normal vectors to the separating hyperplanes of these three methods, as the data dimension increases.

We give a description of the methods mentioned above in Section 2. In Section 3 we show that the first four methods have asymptotically the same first order behavior when the dimension d of the data sets tends to infinity, for fixed sample sizes of the classes. Specifically, we see that when the data sets are Gaussian with common diagonal covariance matrix and one set has mean zero and the other has mean v_d, then the normal vectors of the separating hyperplanes tend to be in the same direction as v_d when $\|v_d\| \gg d^{1/2}$, i.e. are *consistent*, and tend to be orthogonal to v_d when $\|v_d\| \ll d^{1/2}$, i.e. are *strongly inconsistent*. This paper also contains the HDLSS analysis of behavior in the interesting boundary case $\|v_d\| \approx d^{1/2}$. Moreover we observe that the NB method may have a different asymptotic behavior from the other four methods and may

* Corresponding author.
E-mail addresses: addy@cimat.mx (A. Bolivar-Cime), marron@email.unc.edu (J.S. Marron).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.10.001

be inconsistent in many situations where the other methods are consistent. We talk about some simulations that we have done to assess the theoretical results of this paper in Section 4. In Section 5 we give a discussion of our results. Finally, in Section 6 we give the proofs of the results presented in Section 3.

2. Binary discrimination methods

In this section we present the linear classification methods treated in this paper, which are based on separating hyperplanes. Suppose that we have the following training data set

$$(x_1, w_1), (x_2, w_2), \ldots, (x_N, w_N), \tag{1}$$

where $x_i \in \mathbb{R}^d$ and $w_i \in \{-1, 1\}$, for $i = 1, 2, \ldots, N$. In particular, we have two classes of data, the classes C_+ and C_- corresponding to the vectors with $w_i = 1$ and $w_i = -1$ respectively. Let $X = [x_1, x_2, \ldots, x_N]$ be the $d \times N$ matrix of training data and $w = (w_1, w_2, \ldots, w_N)^\top$ be the vector of corresponding class labels. The following notation will be used:

- W is the $N \times N$ diagonal matrix with the elements of w on the diagonal,
- X_+ (X_-) is the sub-matrix of X corresponding to the class C_+ (C_-),
- m (n) is the cardinality of the class C_+ (C_-),
- $\mathbf{1}_k$ is the k-dimensional vector of ones.

We say that the training data set (1) is *linearly separable* if there exists a hyperplane for which all the data of the class C_+ are on one side of the hyperplane and all the data of the class C_- are on the other side. In this case a hyperplane with such property is called a *separating hyperplane* of the training data set.

Only the separable case is treated here, because in HDLSS situations the data from continuous probability densities are linearly separable almost surely (see [7]), and we consider multivariate Gaussian data in this paper.

2.1. Support Vector Machine

An introduction to the Support Vector Machine (SVM) method for binary discrimination analysis is given in this section. For more comprehensive and detailed studies see for example [4–6,8,15,16].

The SVM method was proposed by Vapnik in [15,16]. It is one of the most popular binary discrimination methods in the literature. In the linearly separable case, the SVM method consists of finding a separating hyperplane that maximizes the distances of the hyperplane to the nearest vector of each class.

Here we develop SVM from several view points, which will be needed in the following analysis. Suppose that there exist a vector v and a scalar b such that the following inequalities hold:

$$\begin{aligned} &v^\top x_i + b \geq 1, \quad \text{if } w_i = 1, \\ &v^\top x_i + b \leq -1, \quad \text{if } w_i = -1. \end{aligned} \tag{2}$$

In this case the hyperplane

$$v^\top x + b = 0$$

is a separating hyperplane of the training data set. Note that the inequalities in (2) can be written as

$$w_i(v^\top x_i + b) \geq 1, \quad i = 1, 2, \ldots, N. \tag{3}$$

The vectors x_i that satisfy the equality in (3) are called *support vectors*. That is, the support vectors are the training vectors that belong to one of the hyperplanes

$$v^\top x + b = -1 \quad \text{or} \quad v^\top x + b = 1. \tag{4}$$

The set of support vectors will be denoted by SV.

Let d_+ and d_- be the shortest distances from the separating hyperplane to the nearest vector in C_+ and C_-, respectively. Then the *margin* of the separating hyperplane is defined as $d_0 = d_+ + d_-$. Hence, the margin of the separating hyperplane is the distance between the hyperplanes given in (4) which is

$$d_0 = \frac{2}{\|v\|}.$$

In the separable case the *optimal separating hyperplane* or *SVM hyperplane*

$$v_0^\top x + b_0 = 0$$

is the unique separating hyperplane with a maximal margin. Thus the SVM hyperplane maximizes $2/\|v\|$ subject to the conditions (3). Equivalently, the SVM hyperplane solves the optimization problem

$$\begin{aligned} &\text{minimize } \frac{\|v\|^2}{2}, \\ &\text{subject to } w_i(v^\top x_i + b) \geq 1, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{5}$$