Journal of Multivariate Analysis 100 (2009) 2270–2286

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Independent rule in classification of multivariate binary data

Junyong Park *

Department of Mathematics and Statistics, University of Maryland Baltimore County, United States

ARTICLE INFO

Article history:
Received 22 April 2008
Available online 18 May 2009

AMS subject classifications:
primary 62H30
secondary 62F12

Keywords:
Classification
Independent rule
Sparsity
High dimensional multivariate binary data
MLE
Convergence rate

ABSTRACT

We consider the performance of the independent rule in classification of multivariate binary data. In this article, broad studies are presented including the performance of the independent rule when the number of variables, d, is fixed or increased with the sample size, n. The latter situation includes the case of $d = O(n^\tau)$ for $\tau > 0$ which cover "the small sample and the large dimension", namely $d \gg n$ when $\tau > 1$. Park and Ghosh [J. Park, J.K. Ghosh, Persistence of plug-in rule in classification of high dimensional binary data, Journal of Statistical Planning and Inference 137 (2007) 3687–3707] studied the independent rule in terms of the consistency of misclassification error rate which is called persistence under growing numbers of dimensions, but they did not investigate the convergence rate. We present asymptotic results in view of the convergence rate under some structured parameter space and highlight that variable selection is necessary to improve the performance of the independent rule. We also extend the applications of the independent rule to the case of correlated binary data such as the Bahadur representation and the logit model. It is emphasized that variable selection is also needed in correlated binary data for the improvement of the performance of the independent rule.

1. Introduction

High dimensional data is getting more common in recent statistical science and related fields. There have been extensive studies on classification problems in high dimension both empirically and theoretically, however the studies have focused mainly on normal populations. See for example [1–3]. In such high dimensional data, one typical approach is simplifying classification rule such as the independent rule (or naive Bayes rule) which has been successful in classification problem. The independent rule has been widely used especially for the case of classification of normal populations due to the parsimonious model by ignoring off-diagonal terms in covariance matrix. Bickel and Levina [4] studied the performances of the independent rule and showed that the independent rule outperforms Fisher's rule under some structured parameter space. Fan and Fan [3] also investigated the independent rule and provides the adaptive variable selection procedures which are effective especially in the large dimension and small samples. Most studies on the independent rule highlights that the simplified rules such as the independent rule outperform the full model, for example Fisher's rule. Fan and Lv [5] investigated the necessity of variable selection in various problems including regression and classification. In this article, as the classification problem of non-normal populations, we consider multivariate binary data which are commonly used in many applied areas, ranging from DNA fingerprint data to FMRI and bacterial taxonomy etc.

The independent rule has been also widely used in the problem of classification of multivariate binary data and has achieved successful performance. For example, see [2]. As a recent empirical study, Wilbur et al. [6] analyzed DNA fingerprinting data which is high dimensional multivariate binary data and they emphasized that the independent rule combined with a variable selection procedure performs much better than a rule without considering variable selection.

* Corresponding address: Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, 21250 Baltimore, MD, United States.
E-mail addresses: junpark@umbc.edu, junpark@math.umbc.edu.

0047-259X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.05.004

Park and Ghosh [7] studied the performance of the independent rule in multivariate binary data in the case of a growing number of variables depending on the sample size. They showed that variable selection is necessary to achieve the optimal Bayes error, however they did not study the case of correlated binary data and the convergence rate. In this paper, we focus on the rate of convergence such as how fast misclassification error rate converges to the Bayes error rate. We also study the performance of the independent rule for the case of correlated binary data.

Throughout this paper, we define (Y, X) to be a random vector where $Y = 1$ or 0 and X is a d-dimensional multivariate binary vector. Let $f(x) = P(Y = 1|X = x)$ be the posterior probability of $Y = 1$ given $X = x$, then the optimal rule is $\delta(X) \equiv I(f(x) > 1/2)$ where $I(\cdot)$ is indicator function. The error rate of $\delta(X)$ is theoretically the minimum error rate called the Bayes error. However, in practice, $\delta(X)$ is unknown which needs to be estimated based on the observed samples. This estimated rule $\hat{\delta}(X)$ is based on estimates of f, namely $\hat{f}$, therefore $\hat{\delta}(X) = I(\hat{f}(x) > 1/2)$ and its corresponding misclassification error rate is $P(\hat{\delta}(X) \neq Y)$. This $\hat{f}$ is obtained under the assumption that for given Y, the variables are independent even for dependent variables. With this independent rule $\hat{\delta}$, one main issue in this paper is the behavior of $r(f, \hat{f}) \equiv P(\hat{\delta}(X) \neq Y) - P(\delta(X) \neq Y)$ especially when we discuss the convergence rate of $r(f, \hat{f})$. Due to the difficulty of computation of $r(f, \hat{f})$, one alternative approach is to use a well known inequality such that $r(f, \hat{f}) \leq (2R(f, \hat{f}))^{1/2}$ where $R(f, \hat{f}) = E(f(X) - \hat{f}(X))^2$.

In this paper, we investigate the behavior of $r(f, \hat{f})$ considering both $R(f, \hat{f})$ and $r(f, \hat{f})$. First, we present studies on the convergence rate of $r(f, \hat{f})$ where the number of variables is fixed, which may be considered as the classical asymptotics. In such case, we shall see that the convergence rate of $r(f, \hat{f})$ is dramatically different from $R(f, \hat{f})$. On the contrary, for high dimensional data, it is assumed that the number of variables is allowed to increase with the sample size with $d = O(n^\tau)$. This set up covers the large dimension and the small sample size especially when $\tau > 1$, namely the $d \gg n$ case. This growing number of variables expands the model space, called a triangular array framework which is commonly assumed in the area of high dimensional data analysis. See for example [8,9,7]. In such high dimensional data, some restrictions are put on parameter space to assume sparsity and the pre-ordered variables which imply that the variables are ordered such that the early located variables are more important than the latter part of variables. We also discuss the case where the pre-ordered condition is removed but it is assumed that there exists a rearrangement of variables which leads to the pre-ordered case, which is called an unordered case. In this situation, the rearrangement of variables needs to be found out based on the observed data and we show that $\hat{\delta}(X)$ with selected variables after rearrangement achieves similar results as the pre-ordered case.

The above results of the independent rule are obtained when all the variables are also independent, however one main reason that the independent rule has been widely used in high dimensions is its highly successful performance even when variables are not independent. We extend our study to the case of correlated binary data which are modeled for example by the Bahadur representation and the logit model. See [10]. One may consider the independent rule as a sort of regularized rule since we ignore estimation on dependent structure, however, we shall see that in very high dimensional data, the independent rule is not enough regularization to achieve the Bayes error. We also need to consider variable selection as additional regularization to improve the performance of the independent rule, which can be regarded as the same result as [3] in the classification of normal populations.

This paper is organized as follows. In Section 2, we introduce some notations and definitions used in this paper. In Section 3, when the number of variables is fixed, we discuss the convergence rate of $r(f, \hat{f})$. In Section 4, the performance of the independent rule for the case of independent multivariate binary data are presented for the case when variables are selected and when not selected and the corresponding convergence rates of $r(f, \hat{f})$ will be compared. In Section 5, we extend the previous results to correlated multivariate binary cases such as the Bahadur representation and the logit model and present similar results that the independent variable case including the independent rule with selected variables produces better performance than the independent rule with all the variables. We present simulation studies in Section 6 and discussion and future work in Section 7.

2. Notations

Suppose there are d-dimensional multivariate binary vectors $X = (X_1, X_2, \ldots, X_d)$ which are generated from $X|Y = j$ conditioned on jth class($j = 0$ or 1). Conditioned on $Y = j$, the marginal distribution of X, $X_i|Y = j$, is a *Bernoulli*(p_{ji}) random variable with p_{ji}. From the jth class, $X_j^k = (X_{j1}^k, \ldots, X_{ji}^k, \ldots, X_{jd}^k)$ for $1 \leq k \leq n_j$ are observed and the collection of observations is denoted by $D = \{(X_j^k, Y^k), Y^k = j, 1 \leq k \leq n_j, j = 0, 1\}$. The prior probabilities can be $P(Y = 1) = p > 0$ and $P(Y = 0) = 1 - p = q$, however, without loss of generality, we may consider homogeneous prior probabilities, i.e., $p = q = 1/2$ and consequently equal sample size case, $n_1 = n_2 \equiv n$. With the assumption of $n_1/(n_1 + n_2) \to p > 0$, we can easily extend to the non-homogeneous case with the preservation of all asymptotic results presented in this paper. Since each variable is modeled by a Bernoulli random variable such that $X_{ji} \sim$ *Bernoulli*(p_{ji}) for $j = 0$, 1 and $1 \leq i \leq n$, we define θ

$$\theta \equiv \theta_d \equiv (\theta_{1d}, \theta_{2d}) \equiv (p_{01}, p_{02} \ldots, p_{0d}, p_{11}, p_{12} \ldots, p_{1d}). \tag{1}$$

If correlated multivariate binary data are considered, there may be more parameters which determine dependent structure of Bernoulli variables, namely the parameter vector ϱ. The Bahadur representation and the logit model in Section 5 will be