Communications for Statistical Applications and Methods
2022, Vol. 29, No. 5, 603–614

https://doi.org/10.29220/CSAM.2022.29.5.603
Print ISSN 2287-7843 / Online ISSN 2383-4757

Naive Bayes classifiers boosted by sufficient dimension reduction: applications to top-*k* classification

Su Hyeong Yang[a], Seung Jun Shin[1,a], Wooseok Sung[b], Choon Won Lee[b]

[a]Department of Statistics, Korea University, Korea; [b]Coptiq Co., Ltd., Korea

Abstract

The naive Bayes classifier is one of the most straightforward classification tools and directly estimates the class probability. However, because it relies on the independent assumption of the predictor, which is rarely satisfied in real-world problems, its application is limited in practice. In this article, we propose employing sufficient dimension reduction (SDR) to substantially improve the performance of the naive Bayes classifier, which is often deteriorated when the number of predictors is not restrictively small. This is not surprising as SDR reduces the predictor dimension without sacrificing classification information, and predictors in the reduced space are constructed to be uncorrelated. Therefore, SDR leads the naive Bayes to no longer be naive. We applied the proposed naive Bayes classifier after SDR to build a recommendation system for the eyewear-frames based on customers' face shape, demonstrating its utility in the top-*k* classification problem.

Keywords: dimension reduction, soft classification, recommendation system, top-k classification

1. Introduction

Classification is the most crucial task in contemporary statistical applications such as disease detection, pattern recognition, recommendation system, information retrieval, etc. In statistical learning, classifiers can be categorized broadly into two types: hard classification and soft classification (Liu *et al.*, 2011). The hard classification directly tackles class labels by estimating classification boundaries. A well-known example in this category is the support vector machine (SVM) (Vapnik, 1996) that seeks the classification boundary by maximizing the geometric margin of two classes. The hard classification, including SVM, often shows a promising prediction performance since it directly predicts the class label without a concrete understanding of the data generating process of the data.

On the other hand, soft classification tries to uncover the data generating process, which naturally yields a sensible classification rule. Namely, for a given pair of p dimensional predictor $\mathbf{X}$ and corresponding class label Y with K classes, the soft classification seeks the class probability defined as

$$p_k(\mathbf{x}) = P(Y = k \mid \mathbf{X} = \mathbf{x}), \quad k = 1, \ldots, K,$$

which contains complete information about the data generating process of classification problems. The optimal classification rule based on $p_k(\mathbf{x})$ that minimizes misclassification error rate is to classify an example with $\mathbf{X} = \mathbf{x}$ to the k^{*th} class that maximizes the class probability, i.e., $k^* = \operatorname{argmax}_k p_k(\mathbf{x})$.

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Sungbuk-Gu, Seoul 02841, Korea. E-mail: sjshin@korea.ac.kr

Published 30 September 2022 / journal homepage: http://csam.or.kr

Unlike the hard classification, the soft classification can not only predict the class label but also measure the uncertainty of the prediction. A well-known example in this category includes the logistic regression and the naive Bayes classifier, among many others.

The soft classification is more complicated than the hard one since the class probability is a more informative and complex target than the class label itself. To make things simple, soft classification methods often assume rigid assumptions. For example, the logistic regression imposes a linear model to the logit of the class probability, and the naive Bayes classifier requires statistical independence among predictor variables.

Hard classification has become more popular recently since many applications mainly focus on improving prediction accuracy. However, there are various applications for which soft classification is more desired than hard classification. For example, the soft classification is conceptually more straightforward to extend from standard classification problem to more complex ones, such as top-k classification that allows $k(\geq 1)$ predictions and do not penalize as long as the true label belongs to the set of k predictions. The standard multi-class classification can be regarded as the top-k classification with $k = 1$. The top-k classification is quite popular in various applications, including recommending merchandise. For example, you visit an optical store to purchase eyewear. The optician recommends several types of frames, and you decide whether to purchase one of them or not.

The top-k classification can be directly solved by estimating $p_k(\mathbf{x})$ since the best top-k error (a.k.a top-k Bayes error) at a given $\mathbf{X} = \mathbf{x}$ is obtained if and only if

$$\hat{Y} \in \{Y \mid p_Y(\mathbf{x}) \geq p_{[k]}(\mathbf{x}), Y = 1, 2, \ldots, K\},$$

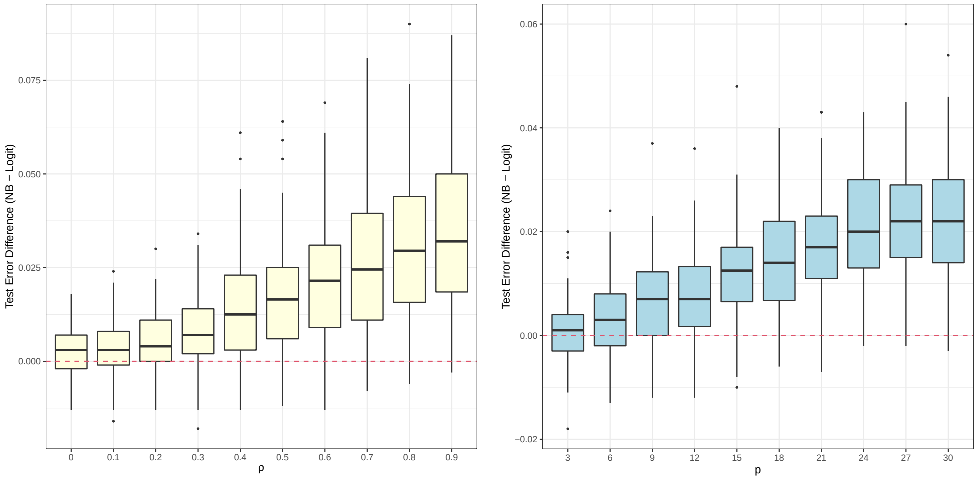
where $\hat{Y}$ denotes the prediction of Y given $\mathbf{X} = \mathbf{x}$, and $p_{[k]}(\mathbf{x})$ denotes the kth largest class probability, i.e., $p_{[1]}(\mathbf{x}) \geq p_{[2]}(\mathbf{x}) \cdots \geq p_{[K]}(\mathbf{x})$ (Lapin *et al.*, 2016). Therefore, the extension to the top-k classification is straightforward in soft classification that directly estimates the class probabilities $p_k(\mathbf{x}), k = 1, \ldots, K$.

In this article, we consider employing the naive Bayes classifier to solve the top-k classification. This work is motivated by Coptiq, a Korean start-up company that manufactures and sells personalized eyewear based on 3D printing technology. The goal of the company is to build a simple but accurate recommendation system to suggest the best five eyewear-frames for a customer's face shape. Although there are numerous more sophisticated alternatives to this problem, we choose the naive Bayes classifier to have a scalable algorithm to handle monthly accumulated data sets that grow in size as the company sells more eyewar.

The naive Bayes classifier is the simplest soft classifier that estimates the class probability under the statistical independence of $\mathbf{X}$.

$$P_k(\mathbf{x}) = \frac{\prod_{j=1}^{p} P\left(X_j = x_j \mid Y = k\right) P(Y = k)}{\sum_{\ell=1}^{L} \prod_{j=1}^{p} P\left(X_j = x_j \mid Y = \ell\right) P(Y = \ell)}.$$

Note that both $P(X_j = x_j \mid Y = k)$ and $P(Y = k)$ for all $j = 1, \ldots, p$ and $k = 1, \ldots, K$ can be readily estimated from the data in a nonparametric way. The naive Bayes classifier is a simple and fast algorithm even for a large data set but heavily relies on the statistical independence assumption among predictors. In addition, its performance also depends on the predictor dimension p. It may suffer from a large p since the estimation errors from $P(X_j = x_j \mid Y = k)$, $j = 1, 2, \ldots, p$ are to be accumulated. In other words, the naive Bayes may not be that naive if the number of predictors is small and they are all independent.



(a) vs $\rho = \{0, 0.1, 0.2, \ldots, 0.9\}$ with $p = 10$

(b) vs $p = \{3, 6, 9, \ldots, 30\}$ with $\rho = 0$

Figure 1: *Relative Performance of Naive Bayes Classifier: Boxplots depict the relative performance of the naive Bayes classifier for different values of* ρ *where* $\text{cov}(X_i, X_j) = \rho^{|i-j|}$*, compared to the logistic regression which can be regarded as an oracle method in this toy example. The naive Bayes performs very well when low-dimensional predictors are independent.*

In order to illustrate this, we consider a toy example simulated from a logistic regression model. We first randomly generate a p-dimensional predictor $\mathbf{X} = (X_1, \ldots, X_p)^T$ from a mean-zero multivariate normal distribution with a covariance with the autoregressive structure, i.e., $\text{cov}(X_i, X_j) = \rho^{|i-j|}$ for a given $\rho \in [0, 1]$. The binary response Y is then generated from Bernoulli distribution with the success probability $p(\mathbf{x}) = P(Y = 1 \mid \mathbf{X} = \mathbf{x})$ being

$$\log\left\{\frac{p\left(\mathbf{X}\right)}{1 - p\left(\mathbf{X}\right)}\right\} = 1 + \boldsymbol{\beta}^T\mathbf{X},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ with $\beta_j \sim \text{Uniform}(-1, 1)$, $j = 1, \ldots, p$.

We generate 2,000 observations independently and randomly split them into equal-sized training and test sets. First, we investigate the performance of the naive Bayes classifier as ρ increases when p is fixed as 10. Next, we check its performance as p increases under the independence assumption holds (i.e., $\rho = 0$). Figure 1 shows the difference in test error rate between the naive Bayes classier and the logistic regression. We remark that the data are generated from the logistic model, and thus the naive Bayes cannot beat the logistic regression. However, the naive Bayes classifier is comparable when ρ is closed to 0, and p is not very large. This simple experiment demonstrates the promising performance of the naive Bayes classifier when the low dimensional predictors are independent. This naturally leads us to reduce the predictor dimension before applying the naive Bayes classifier. For example, the principal component analysis (PCA) can reduce the dimension of $\mathbf{X}$. Moreover, the principal vectors are orthogonal by construction, and the corresponding scores are independent as long as $\mathbf{X}$ is normally distributed. However, PCA is not the best choice for dimension reduction in supervised learning such as regression and classification. In this article, we propose to employ the sufficient dimension reduction (SDR) (Li, 2018) to $\mathbf{X}$ that reduces predictor dimension without loss of information about Y contained $\mathbf{X}$. In this regard, SDR can be viewed as a supervised dimension reduction while PCA is an unsupervised one, and thus better suitable for the classification problem.

2. Naive Bayes with sufficient dimension reduction

The SDR is a popular supervised dimension reduction tool that projects $\mathbf{X}$ to a lower-dimensional subspace without loss of classification (or regression) information contained in $\mathbf{X}$. Given a pair of random variable response Y and p-dimensional predictor $\mathbf{X}$, the SDR seek a matrix $\mathbf{B} \in \mathbb{R}^{p \times d}$ that satisfies

$$Y \perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}, \tag{2.1}$$

where $\perp$ denotes the statistical independence. Under (2.1), once can naturally reduce the predictor by projecting $\mathbf{X}$ to the column space of $\mathbf{B}$, which we call dimension reduction subspace (DRS). However, $\mathbf{B}$ and thus DRS is not unique. If the intersection of all DRS itself is a DRS, it is the DRS with the smallest dimension (i.e., maximal dimension reduction). We call this central subspace denoted by $\mathcal{S}_{Y|\mathbf{X}}$. It is known that $\mathcal{S}_{Y|\mathbf{X}}$ exists under very mild conditions (Yin *et al.*, 2008). In SDR, $\mathcal{S}_{Y|\mathbf{X}}$ is the final target and it is assumed that $\mathrm{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$.

There are tons of methods to estimate $\mathbf{B}$. The sliced inverse regression (SIR) (Li, 1991) is the earliest proposal that is still popular in practice due to its simplicity. The sliced averaged variance estimator (SAVE) (Cook and Weisberg, 1991) and the principal Hessian direction (pHd) (Cook RD, 1998) are also traditional candidate. These methods are available in `dr`-package in `R`. More recent methods include, but not limited to the directional regression (Li and Wang, 2007), cumulative slicing mean estimation (Zhu *et al.*, 2010), and principal support vector machine (Li *et al.*, 2011). Since SDR is firstly developed under the regression context with contiuous Y. It suffers when Y is categorical. In particular, all aforementioned methods fails to identify $\mathbf{B}$ when $d = \dim(\mathcal{S}_{Y|\mathbf{X}}) > K$ if Y is K-class response. The problem is indeed serious in binary classification with $K = 2$. Recently, several methods are developed to tack this issue in binary classification. See Shin *et al.* (2014) and Shin *et al.* (2017) for example.

The SDR assumption (2.1) implies

$$p_k(\mathbf{x}) = P\left(Y = k \mid \mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}\right).$$

Assuming $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_d)$ is orthonormal basis of $\mathcal{S}_{Y|\mathbf{X}}$ without loss of generality, $\mathbf{B}^T\mathbf{X} = (\mathbf{b}_1^T\mathbf{X}, \dots, \mathbf{b}_d^T\mathbf{X})$ called sufficient predictors, i.e., predictors projected on $\mathcal{S}_{Y|\mathbf{X}}$, are always uncorrelated as Cov $(\mathbf{b}_i^T\mathbf{X}, \mathbf{b}_j^T\mathbf{X}) = \mathrm{trace}\{\mathrm{Cov}(\mathbf{X}) \cdot \mathbf{b}_i^T\mathbf{b}_j\} = 0$, $\forall i \neq j$. Therefore the original predictors $\mathbf{X}$ are normally distributed, and the sufficient predictors $\mathbf{B}^T\mathbf{X}$ are independent. We remark that the normality of $\mathbf{X}$ seems very strong in theory but not in practice since it does not assume anything about the relationship between Y and $\mathbf{X}$, but the marginal distribution of $\mathbf{X}$ only.

Now, it is natural to consider a Naive Bayes classifier using the sufficient predictor under (2.1) as follows:

$$P\left(Y = k \mid \mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}\right) = \frac{\prod_{j=1}^{p} P\left(\mathbf{b}_j^T X_j = \mathbf{b}_j^T x_j \mid Y = k\right) P(Y = k)}{\sum_{\ell=1}^{L} \prod_{j=1}^{p} P\left(\mathbf{b}_j\mathbf{X} = \mathbf{b}_j\mathbf{x} \mid Y = \ell\right) P(Y = \ell)},$$

which is equivalent to $p_k(\mathbf{x})$ under (2.1).

We remark that SDR boosts its practical utility and enhances classification performance by cleverly mitigating significant drawbacks of the naive Bayes classifier. This leads us to employ the naive Bayes classifier after the SDR to solve the soft classification problem and then apply it to the top-k classification problem motivated by the eyewear company, Coptiq.

3. Simulation study

We conduct a simulation study in order to demonstrate the performance of the naive Bayes classifier with SDR. We first generate p-dimensional predictors from $\mathbf{x}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), i = 1, \dots, n$ for three different covariance structure $\boldsymbol{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^{p}$ as follows: $\sigma_{ii} = 1$ for all $i = 1, \dots, n$ and

1. Independent: $\sigma_{ij} = 0$,
2. Auto-Regressive: $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.7$,
3. Compound Symmetry: $\sigma_{ij} = \rho$ with $\rho = 0.7$

for $i \neq j$. The naive Bayes classifier deteriorates as the dependency among predictors gets stronger and thus we expect that it becomes inefficient under the auto-regressive or compound symmetry covariance structure. We set $n = 500$ with $p \in \{10, 20, 30\}$.

Next, we consider the logistic regression model to generate y_i given $\mathbf{x}_i$

$$y_i \mid \mathbf{x}_i \sim \text{Multinomial}\{1, p_1(\mathbf{x}_i), \dots, p_K(\mathbf{x}_i)\},$$

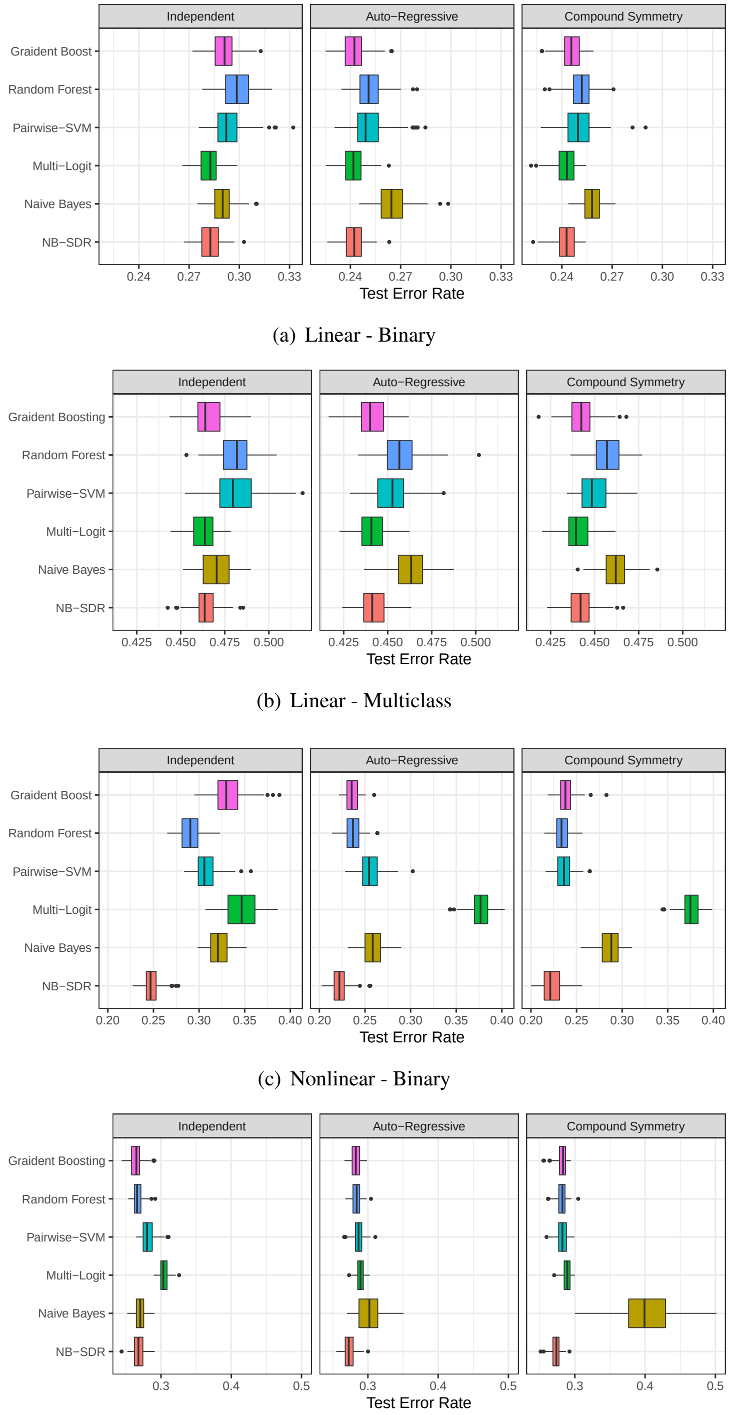
subject to $\sum_{k=1}^{K} p_k(\mathbf{x}) = 1$, where

$$\log\left(\frac{p_k(\mathbf{x}_i)}{1 - p_K(\mathbf{x}_i)}\right) = f\left(\boldsymbol{\beta}_k^T \mathbf{x}_i\right), \quad k = 1, \dots, K-1. \tag{3.1}$$

That is, the true central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is span$\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{K-1}\}$ under (3.1).

We consider two scenarios for the classification function f: $f(u) = u$ and $f(u) = (u-1)^2 - 1$ to represent linear and nonlinear ones, respectively. Note that the linear f yields the conventional logistic regression model where the logistic regression outperforms all others. On the other hand, when f is nonlinear it suffers from model misspecification as we will see in the following. We consider both binary (i.e. $K = 2$) and multi-class problems with $K = 3$, and set $\boldsymbol{\beta}_1 = (1,\ 1,\ \mathbf{0}_{p-2}^T)^T$ for the binary case, and $\boldsymbol{\beta}_1 = (1,\ \mathbf{0}_{p-1}^T)^T$ and $\boldsymbol{\beta}_2 = (-1,\ \mathbf{0}_{p-1}^T)^T$ for the multiclass case.

To implement the proposed method, we first apply SIR to estimate $\mathcal{S}_{Y|\mathbf{X}}$, and then the naive Bayes classifier is trained on the estimated $\mathcal{S}_{Y|\mathbf{X}}$. We denote this with NB-SDR in upcoming figures. As competing methods against NB-SDR, we consider five popular classification tools: naive Bayes without SDR, logistic regression, SVM (with Gaussian kernel), random forest, and gradient (logit-) boosting. Tuning parameters for SVM, random forest and gradient boosting are chosen by five-fold cross-validation. The classification performance is measured by the error rate for the test set with a sample size of 5,000, independent of the training set. Figure 2 depicts the performance of the classifiers for different scenarios with $p = 10$. Under the linear cases, all methods perform quite well. It is not surprising that the logistic regression performs the best. The naive Bayes classifier is not a good choice when predictors are not independent. However, the proposed NB-SDR shows promising performance even with the strongly correlated predictors. Compared to the popular black-box type method, the NB-SDR is still promising. The logistic regression fails for nonlinear cases, while other methods seem okay due to their flexibility. Again, we observe similar patterns as seen in the linear cases. The proposed NB-SDR outperforms existing methods, demonstrating the SDR's dramatic effect on the naive Bayes classifier. We observe nearly identical patterns for $p = 20$ and 30 and thus omitted them to avoid redundancy.



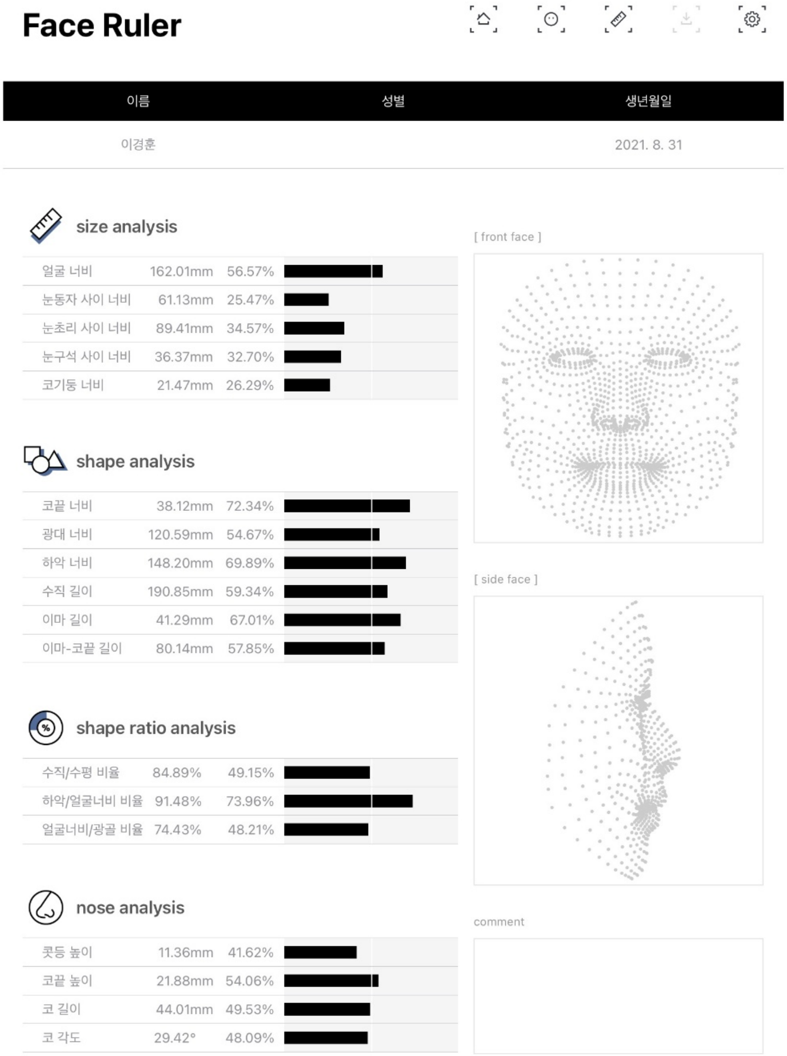
(a) Linear - Binary

(b) Linear - Multiclass

(c) Nonlinear - Binary

(d) Nonlinear - Multiclass

Figure 2: *Simulation Results: The proposed NB-SDR shows promising performance for all scenarios. This clearly demonstrates the dramatic effect of the SDR on the naive Bayes classifier.*



Face Ruler

이름	성별	생년월일
이경훈		2021. 8. 31

size analysis

얼굴 너비	162.01mm	56.57%
눈동자 사이 너비	61.13mm	25.47%
눈초리 사이 너비	89.41mm	34.57%
눈구석 사이 너비	36.37mm	32.70%
코기둥 너비	21.47mm	26.29%

shape analysis

코끝 너비	38.12mm	72.34%
광대 너비	120.59mm	54.67%
하악 너비	148.20mm	69.89%
수직 길이	190.85mm	59.34%
이마 길이	41.29mm	67.01%
이마-코끝 길이	80.14mm	57.85%

shape ratio analysis

수직/수평 비율	84.89%	49.15%
하악/얼굴너비 비율	91.48%	73.96%
얼굴너비/광골 비율	74.43%	48.21%

nose analysis

콧등 높이	11.36mm	41.62%
코끝 높이	21.88mm	54.06%
코 길이	44.01mm	49.53%
코 각도	29.42°	48.09%

[front face]

[side face]

comment

Figure 3: *Captured image of Face Ruler, the software developed by Coptiq to measure the face shape of a customer.*

4. Application to eyewear-frame recommendation

Coptiq is a Korean start-up company that manufactures and sells personalized eyewear based on 3D printing technology. The company developed software (called Face Ruler) that scans customers' faces directly by using a true-depth camera. The facial information is stored as 1221 points in 3-dimensional space, as shown in Figure 3.

The raw data collected from the scanned image is processed and translated into the nineteen features listed below, which are based on the knowledge and expertise of optical technicians. For example, "Face Width" is the distance between the right and left cheek ends when looking straight at the

Table 1: List of predictors, and their normality test results and transformations applied if necessary

Type	Variable Name	Skewness	p-value	Transformation
Base	Age	0.4441	0.0000*	Square Root
Size	Face Width	−0.5402	0.0000*	Yeo-Johnson
	Eye Gap (Center)	−0.1950	0.0081*	Yeo-Johnson
	Eye Gap (Outer)	−0.1462	0.2272	
	Eye Gap (Inner)	−0.0568	0.0389	
Shape	Cheekbone size	−0.4069	0.0000*	Yeo-Johnson
	Mandibular size	−0.4987	0.0000*	Yeo-Johnson
	Vertical Face	−0.7814	0.0000*	Ordered Quantile
	Forehead	−0.2619	0.0000*	Box-Cox
	Forehead-Nose Gap	−0.0567	0.3105	
Ratio	Face-Width/Face-Length	−0.0735	0.5743	
	Mandibular-Size/Face-Width	0.1787	0.0588	
	Face-Width/Cheekbone-Size	−0.8850	0.0517	
Nose	Width (Bridge)	0.1468	0.3483	
	Width (Tip)	−0.1341	0.0371	
	Height (Bridge)	0.1757	0.0942	
	Height (Tip)	−0.0723	0.0887	
	Length	−0.0479	0.9645	
	Angle	−0.0271	0.9396	

face, and "Face Length" is the distance between the tip of the brow and the lower end of the face.

The data set contains the face information of customers who purchased eyewear in the store run by Coptiq for the last three years, from January 2018 to January 2021. The company received informed consent from all customers when purchasing the eyewear. The processed data used for the analysis does not contain any information identifying individuals.

The company produces 67 types of frames, that are too many, and we restrict our attention to 27 types with a sufficient purchase history. We also exclude missing observations and very few customers who bought extreme-sized frames (e.g., XXS or XXXL). Finally, 5,069 customers are included in the analysis. The company wants to develop a recommendation model for recommending top-five frame types that best fit the customer based on the features obtained from the scanned face image. This can be viewed as a standard example of top-k classification problem. Toward this, we need to estimate class probabilities $p_k(\mathbf{x})$ for a given $\mathbf{x}$, i.e., facial information measured by the Face Ruler.

To apply NB-SDR, we first need to check the marginal normality of the predictors. We admit that, in theory, the marginal normality does not necessarily imply the joint normality. Still, we presume that the marginal normality is sufficient to assume joint normality in the application for technical simplicity. For the normality check, We conduct the Kolmogorov test and have that some variables are not normally distributed under the significance level $\alpha = 0.01$. We remark that we use quite small α since the sample size is large. As reported in Table 1, age, face-width, eye-gap (center), eye-gap (inner), cheekbone size, mandibular size, vertical face-length, and forehead length turn out to be non-normal, and we applied various transformation methods, including Box-Cox transformation(Box and Cox, 1964), Yeo-Johnson transformation (Yeo and Johnson, 2000), Ordered Quantile transformation (Beasley *et al.*, 2009) all available in `bestNormalize` package in `R`. After the proper transformation, we confirmed that all variables follow normal distributions.

We then apply SIR to reduce the predictor dimension. We estimate the structure dimension from the sequential χ^2 test which given $d = 3$. Figure 4 display the estimated biases of $\mathcal{S}_{Y|\mathbf{X}}$ by SIR. We observe that three variables related to the nose shape (Nose-Angle, Nose-Length, Nose-Bridge Height) are the most informative for frame type choice. After the dimension reduction, one can train

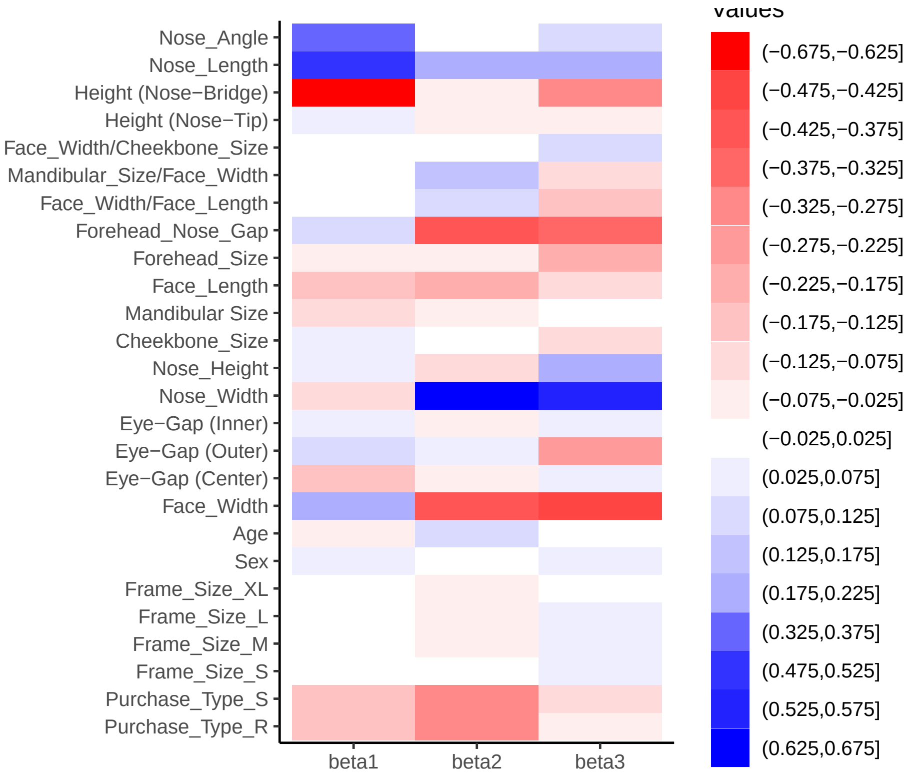


Figure 4: *The estimated directions of the central subspace by SIR for the Coptiq Data.*

Table 2: Test top-five accuracy and computing times of different models applied to Coptiq Data

	NB	NB-SDR	SVM	MLR	RF	GBM
Accuracy	0.556	0.631	0.613	0.630	0.612	0.636
Computing Time (in sec.)	0.018	1.973	377.725	1.617	3.784	100.996

the naive Bayes classifier.

To compare the performance of the proposed method, we split the data into training and test sets as follows. The customers who purchase the eyewear before October 2020 are used as a training set and the ones after October 2020 are used as a test set. We compared the test performance of naive Bayes with sufficient dimension reduction (NB-SDR) to the following five popular classification methods considered in Section 3: naive Bayes classifier (NB), support vector machine (SVM) with the Gaussian kernel, multinomial logistic regression (MLR), random forest(RF), and gradient boosting machine (GBM). Both SVM and GBM are tuned via cross-validation to optimize the cross-validated top-five error.

First, we compare the top-k classification performance for different methods. Table 2 shows the test top-5 accuracy that counts the case when the purchased frame type is included in the recommended five types. We also report the computing time (in seconds) to illustrate the computational efficiency of the proposed method. One can observe that the gradient boosting machine shows the best performance while the proposed NB-SDR and multinomial logistic regression are also compara-

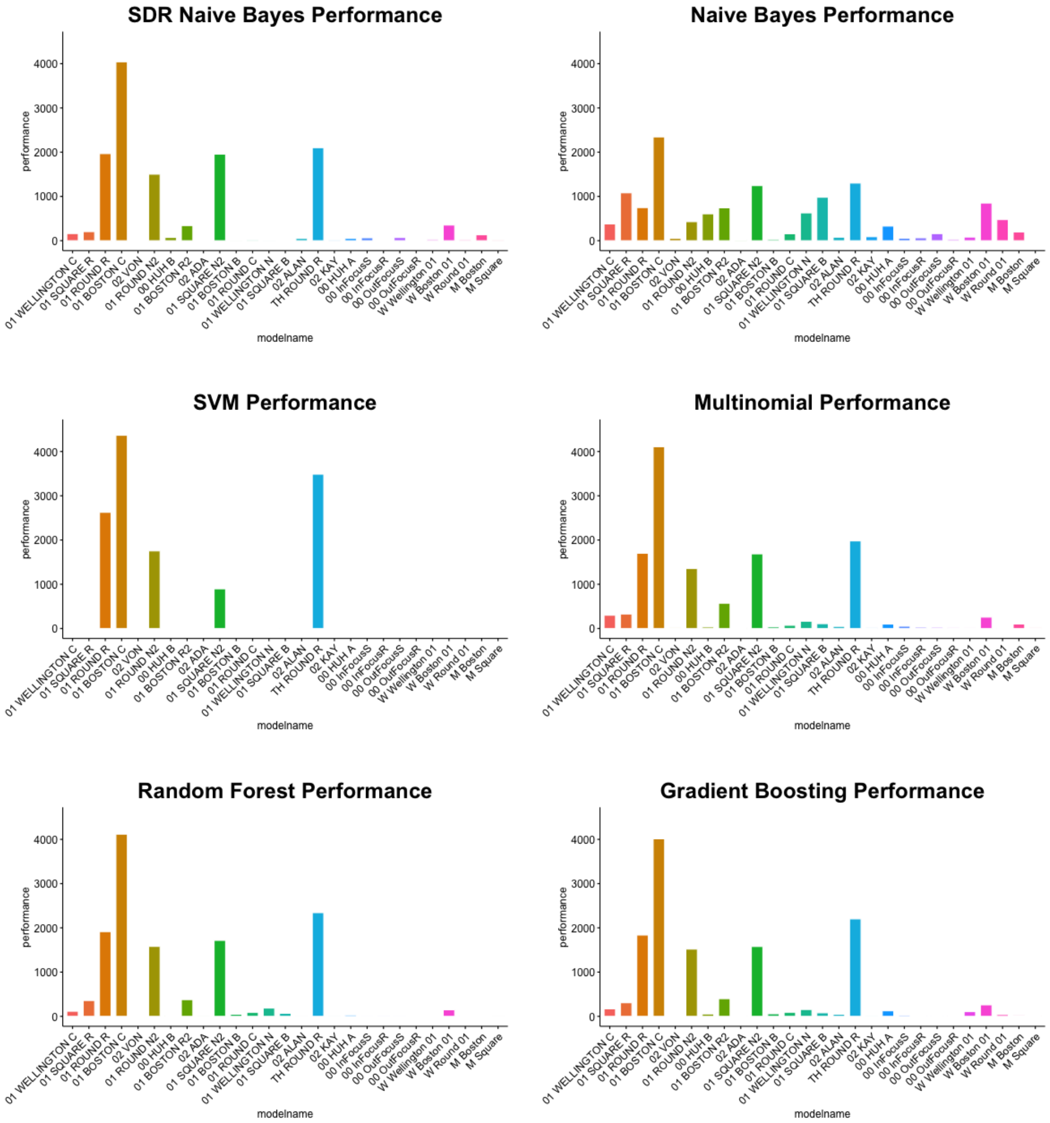


Figure 5: *Bar charts of the recommendation scores for different methods including NB-SDR and five competitors, NB, SVM, MLR, RF and GBM.*

ble. One drawback of GBM is its computational complexity. Both GBM and SVM take much longer than others since they involve additional tuning steps. This is not a minor issue since the size of data will increase as more eyewear are sold. Second, we also compare the results in terms of the diversity of the recommendation. The diversity of the prediction is another important aspect of the recommendation system since it is not desirable to recommend always the most popular models. In fact, the top-*k* accuracy is often maximized in practice when recommending simply the best-selling models. To check the diversity of the recommendation, we calculate a recommendation score for each model and compare its distribution. The score is computed as follows. Five points to the firstly recommended frame, which has the highest probability (or score), four points to the frame with the second-highest probability, three points to the third, two points to the fourth, and one point for the fifth, and combined the scores for each frame type.

Figure 5 depicts the bar chart (i.e., distribution) of the recommendation score for different methods. One can compare the diversity of the recommendation results for different methods by checking the spread of the distribution. We note that the five best-selling models are 01 Boston C, TH ROUND R, 01 SQUARE N2, 01 ROUND R, and 01 ROUND N2, and the SVM tends to recommend only the five best-selling models. On the other hand, NB shows the most diverse recommendation results, but its accuracy is too poor (See Table2). Again, GBM, MLR, and the proposed NB-SDR show a similar diversity with relatively good accuracy; these three methods are all reasonable choices for this particular application. However, MLR and proposed NB-SDR would be a better choice than GBM for large-scale data sets like the Coptiq data, considering the computational complexity.

5. Concluding summary

This article proposes a simple and effective classification tool by combining two popular ideas in statistical learning: the naive Bayes classifier and sufficient dimension reduction. We observe that sufficient dimension reduction can significantly improve the naive Bayes classifier from both simulated and real data analysis. The proposed method is desirable when soft classification is preferred, such as top-*k* classification, as demonstrated in this article.

References

Beasley TM, Erickson S, and Allison DB (2009). Rank-based inverse normal transformations are increasingly used, but are they merited?, *Behavior Genetics*, **39**, 580–595.

Box GE and Cox DR (1964). An analysis of transformations, *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**, 211–243.

Cook RD(1998). Principal hessian directions revisited, *Journal of the American Statistical Association*, **93**, 84–94.

Cook RD and Weisberg S (1991). Discussion of "Sliced inverse regression for dimension reduction", *Journal of the American Statistical Association*, **86**, 28–33.

Lapin M, Hein M, and Schiele B (2016). Loss functions for top-k error: Analysis and insights, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1468–1477.

Li B (2018). *Sufficient Dimension Reduction: Methods and Applications with R*, CRC Press, Florida.

Li B, Artemiou A, and Li L (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction, *The Annals of Statistics*, **39**, 3182–3210.

Li B and Wang S (2007). On directional regression for dimension reduction, *Journal of the American Statistical Association*, **102**, 997–1008.

Li KC (1991). Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association*, **86**, 316–342.

Liu Y, Zhang HH, and Wu Y (2011). Hard or soft classification? large-margin unified machines, *Journal of the American Statistical Association*, **106**, 166–177.

Shin SJ, Wu Y, Zhang HH, and Liu Y (2014). Probability enhanced sufficient dimension reduction in binary classification, *Biometrics*, **70**, 546–555.

Shin SJ, Wu Y, Zhang HH, and Liu Y (2017). Principal weighted support vector machines for sufficient dimension reduction in binary classification, *Biometrika*, **104**, 67–81.

Vapnik V (1996). *The Nature of Statistical Learning Theory*, Cambridge University Press, Cambridge.

Yeo IK and Johnson RA (2000) . A new family of power transformations to improve normality or symmetry, *Biometrika*, **87**, 954–959.

Yin X, Li B, and Cook RD (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression, *Journal of Multivariate Analysis*, **99**, 1733–1757.

Zhu LP, Zhu LX, and Feng ZH (2010). Dimension reduction in regressions through cumulative slicing estimation, *Journal of the American Statistical Association*, **105**, 1455–1466.

Received April 28, 2022; Revised May 30, 2022; Accepted May 31, 2022