A RELATIVE EFFICIENCY OF NAÏVE BAYES CLASSIFIER TO BAYES DECISION RULE

HyunJi Kim¹)

ABSTRACT

This paper investigates the relative efficiency of Naïve Bayes classifier (BD) to Bayes decision rule, which is the optimal classification rule. Under normality, Discriminant Analysis is also optimal. This paper investigates the relative efficiency of NB compared to Bayes decision, Quadratic and Linear discriminant function. It will be seen that the efficiency of NB depends on the correlation along with the signal-to-noise-ratio(S/N) in multiplication.

Key words: Naïve Bayes classifier, Bayes decision rule, discriminant function

1. Introduction

For classification, Bayes decision rule (BD) is the best decision rule which gives the minimum probability of misclassification. However, this requires the knowledge of the distribution which we do not have in practice. Hence, it is difficult to use BD since there are some practical problems of estimating joint distribution (Li and Yen, 2005). It is relatively easy to estimate joint distribution in case of normal distribution. Under normality, BD coincides with the rule of Discriminant Analysis (DA). In this case, DA is also optimal under normality (Rausch and Kelley, 2009). Discriminant function has the form of a quadratic function of a data vector \( z \). Quadratic term is dropped if the covariance matrices are equal across groups. Previous studies have shown that the class-conditional probability density estimation for BD occasionally raises troublesome problems (Davy and Tourneret, 2000) and DA performs poorly under non-normal distribution (Krzanowski, 1977; Gnanadesikan, 1989; Pardoe et al., 2006).

Naïve Bayes classifier (NB) is a well-known classification model for its simplicity and efficiency compared to the other sophisticated classification models (John and Langley, 1995; Domingos and Pazzani, 1997; Rish et al., 2001). NB assumes class-conditional

¹) Ph. D. Student, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, South Korea
E-mail: hyunjikim123@gmail.com
independence assumption that means joint class density is a product of marginal densities (Hastie et al., 2008). Therefore, estimating joint density becomes feasible when a class variable is given since we only need to estimate the conditional marginal probabilities. NB will be close to optimal classification model regardless of underlying distribution if the classification performance of NB is close to that of BD. This paper investigates the properties of BD, QDA, LDA and NB and compares their relative efficiencies under normality.

2. Investigation

NB is based on Bayes’ rule and assumes that the prediction variables are conditionally independent when a class variable is given as in the following.

\[ P(C|X) \propto P(X|C)P(C) = \prod_{i=1}^{p} P(X_i|C)P(C) \]  

(2.1)

where \( X \) is a random \( p \)-vector, \( C \) is the class variables.

For binary classification problem, we need to classify between the two classes \( C = A \) and \( C = B \). For NB, the log ratio of the two densities is given as follows.

\[ L(x) = \log \frac{P(C = A|X)}{P(C = B|X)} = \log \frac{P(C = A)}{P(C = B)} + \sum_{i=1}^{p} \log \frac{f(X_i|C = A)}{f(X_i|C = B)} \]  

(2.2)

We assign an observation \( x \) to \( A \) if the log ratio is greater than zero; that is,

\[ C = \begin{cases} A, & \text{if } L(x) \geq 0 \\ B, & \text{if } L(x) < 0 \end{cases} \]  

(2.3)

Now we assume the conditional class densities for the class \( C = A \) and \( C = B \) are \( f_A(x) \sim N_p(\mu_A, \Sigma_A) \) and \( f_B(x) \sim N_p(\mu_B, \Sigma_B) \). Then the log ratio of the two densities is given by

\[ L(x) = \log \frac{f_A(x)}{f_B(x)} = -\frac{1}{2} \left[ \log \frac{\Sigma_A}{\Sigma_B} + (\mu_A - \mu_B)^T \Sigma_B^{-1} (\mu_A - \mu_B) + (\mu_A - \mu_B)^T \Sigma_A^{-1} (\mu_A - \mu_B) \right] - \frac{1}{2} x^T (\Sigma_A^{-1} - \Sigma_B^{-1}) x \]  

(2.4)

\[ L(x) \] has the form of a quadratic function of \( x \) and BD is \( C = A \) if \( L(x) \geq 0 \) and \( C = B \) if \( L(x) < 0 \). Note that \( L(x) \) is equivalent to the classical discriminant function. The above rule generates a BD boundary which assigns a new incoming observation \( x \) into class A or class B. Our interest is not assigning \( x \) to a class, but investigating how the BD boundary changes with NB rule. Hence for our purpose, it is safe to consider the case where the locations of the two class distributions are shifted in the direction of \(-\mu_A\), since the shape of these two boundaries should not change with location shift. Then the above log ratio (2.4) becomes as follows
\[
L(x) = \log_r \frac{f_A(x)}{f_B(x)} = -\frac{1}{2} \left( \log_{10} \frac{\Sigma_A}{\Sigma_B} \right) - \mu^T \Sigma_B^{-1} \mu - \frac{1}{2} x^T (\Sigma_A^{-1} - \Sigma_B^{-1}) x
\]  

where \( \mu = \mu_B - \mu_A \).

When common covariance is assumed such that \( \Sigma_A = \Sigma_B = \Sigma \) in (2.5), \( L(x) \) is equal to linear discriminant function as follows.

\[
L(x) = \log_r \frac{f_A(x)}{f_B(x)} = -\mu^T \Sigma^{-1} (x - \frac{1}{2} \mu)
\]  

For NB, the above log ratio becomes as follows

\[
L(x) = \log_r \frac{f_A(x)}{f_B(x)} = -\mu^T D^{-1} (x - \frac{1}{2} \mu)
\]

where \( D = \text{diag}(\Sigma) \). Therefore we assign \( x \) into \( A \) if the following is satisfied

\[
\sum_{i=1}^{p} \frac{\mu_i}{\sigma_i^2} \left( x_i - \frac{1}{2} \mu_i \right) \leq 0
\]  

For the special case of \( p = 2 \), \( L(x) \) of equation (2.6) becomes as follows

\[
\sum_{i=1}^{2} \frac{\mu_i}{\sigma_i^2} \left( x_i - \frac{1}{2} \mu_i \right) \left( \frac{1}{\sigma_i^2} - \frac{\rho}{\sigma_i \sigma_2} \right) \leq 0
\]

The above equation shows that the effect of \( \rho \) on the BD decision rule comes with the term \( \left( \frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1} \right) \). In other words, the effect of \( \rho \) on the decision rule comes always with the signal-to-noise-ratio (S/N) \( \mu/\sigma \). This can be seen clearly by considering the case of \( \mu_1 = \mu_2 = \mu \) and \( \sigma_1 = \sigma_2 = \sigma \).

\[
\sum_{i=1}^{2} \left( x_i - \frac{1}{2} \mu_i \right) \frac{\mu_i}{\sigma_i^2} (1-\rho) \leq 0
\]  

From Bickel and Levina (2004), misclassification probabilities for BD and NB under normality are \( P_B = 2 \Phi \left( -\frac{1}{2} \Delta \right) \) and \( P_N = 2 \Phi \left( -\frac{1}{2} \frac{\mu^T D^{-1} \mu}{\sqrt{\mu^T D^{-1} D^{-1} \mu}} \right) \), where \( \Delta^2 = \mu^T \Sigma^{-1} \mu \) and \( P_B \) and \( P_N \) correspond to misclassification probabilities of BD and NB, respectively.

We now consider the misclassification probability for the special case of \( p = 2 \) with \( \mu_1 = \mu_2 \) and \( \sigma_1 = \sigma_2 \). In this case, a simple algebra shows \( P_N = P_B \) and it is as follows

\[
P_N = P_B = 2 \Phi \left( -\frac{1}{\sqrt{2}} \frac{\mu}{\sigma} \sqrt{1 + \rho} \right) = 2 \Phi \left( \frac{1}{\sqrt{2} \sigma} \sqrt{\frac{\mu^2}{1+\rho}} \right)
\]  

The above equation also shows the effect of \( \rho \) on the misclassification probability comes together with the S/N. Figure 1 shows that NB misclassification probability is increasing with \( \rho \) and \( \rho \) has effect on the probability together with S/N. It also shows that S/N has considerable effect on the misclassification probability than \( \rho \). We will further extent our
investigation to multivariate cases for the effect of correlation structures and multivariate 
S/N ratio (mahalanobis-distance) on the relative efficiencies of the four classifiers.

Fig. 1: NB misclassification probabilities for with respect to \( \rho \) and S/N ratio

REFERENCES

Bayes', and some alternatives when there are many more variables than observations, 

Dave, M., Doncarli, C. and Tourneret, J. Y. (2000). Supervised Classification Using MCMC Methods, 
in *Proc. IEEE ICASSP-00* 1, 33-36.


Gnanadesikan, R. (1989). Discriminant analysis and clustering: Panel on Discriminant Analysis, 

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, 


Pardoe, I., Yin, X. and Cook, R. D. (2007). Graphical tools for quadratic discriminant analysis, 


Rish, I., Hellerstein, J. and Thathachar, J. (2001). An analysis of data characteristics that affect Naive 
Bayes performance. *In ICML-01.*