Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 830–837
Pattern Recognition Letters
www.elsevier.com/locate/patrec

On the optimality of Naïve Bayes with dependent binary features

Ludmila I. Kuncheva *

School of Informatics, University of Wales, Dean Street, Bangor, Gwynedd LL57 1UT, UK

Received 20 July 2005; received in revised form 7 November 2005
Available online 30 January 2006

Communicated by Prof. F. Roli

Abstract

While Naïve Bayes classifier (NB) is Bayes-optimal for independent features, we prove that it is also optimal for two equiprobable classes and two features with equal class-conditional covariances. Although strict optimality does not extend for three features, equal covariances are expected to be beneficial in higher-dimensional spaces.

Keywords: Statistical pattern recognition; Naïve Bayes classifier (NB); Optimality of NB; Dependent binary features

1. Introduction

The Naïve Bayes classifier (NB), called also "idiot's Bayes", continues to receive a lot of praise in the literature due to its simplicity and accuracy (Langley et al., 1992; Hand and Yu, 2001; Jamain and Hand, 2005). NB is Bayes-optimal, i.e., guarantees minimum classification error, when the features in the problem are independent. However, it is well documented that NB is consistently good far beyond this optimality condition. The word most used to describe its performance is "surprising". Many studies look for explanations for this phenomenon and try to establish *necessary and sufficient* conditions for the optimality of NB. While important arguments and results have been already formulated (details are given in Section 2 below), there is no generally valid set of such necessary and sufficient conditions.

This study looks for further insights into optimality of NB for binary features. The motivation came from a problem from veterinary medicine where the signs measured on cattle are binary and the diagnosis is one of two possible classes, BSE or not BSE. BSE is a notifiable fatal neurodegenerative disease in cattle which has no known cure. There was a BSE epidemic in Britain in the 1990s and with the first BSE case diagnosed in the USA at the end of 2003, the problem becomes one of global importance. The curious aspect of this problem was that there was no data set as such but only estimates of the class-conditional probabilities by domain experts. Given are only the marginal probabilities, $P(x_i = 1|\omega_k)$, where x_i is the ith sign (feature). Value 1 means that the sign is present in the animal, and ω_k is the class label (BSE or not BSE). As no further information is available, the features must be considered as independent, hence NB can be applied as the optimal classifier. Assuming that the probability estimates are the exact probabilities, the question is how different is NB from the true optimal Bayes classifier? As there is no data set, an experimental verification is not possible. This prompted the question of how much dependence NB can tolerate and still be optimal.

The rest of the paper is organized as follows. Section 2 gives a brief account of some important results from the literature explaining why NB works when the independence assumption does not hold. In Section 3, a two-feature two-class problem is considered. We prove that NB is optimal if the dependencies between the two features are the

* Tel.: +44 1248 383661; fax: +44 1248 361429.
E-mail address: L.I.Kuncheva@bangor.ac.uk

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.12.001

same for both classes. Unfortunately, this result does not extend beyond two features. Section 4 looks into three binary features and brings into the study non-pairwise measures of dependence: Q_{123}, divergence and two distances between probability distributions. Section 5 summarizes the results and outlines other possible explanation routes.

2. Naïve Bayes: why is it so successful?

Let $\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$ be a feature vector. To label it in one of the c classes of the problem, $\omega_1, \ldots, \omega_c$, we use the posterior probability $P(\omega_k|\boldsymbol{x})$. Choosing the class corresponding to the largest posterior probability for the respective $\boldsymbol{x}$ guarantees minimum error across the whole space spanned by the n features. We shall call this the Bayes classifier, and the corresponding error, the Bayes error, E_{B}. The posterior probabilities are calculated as $P(\omega_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_k)P(\omega_k)}{p(\boldsymbol{x})}$, where $p(\boldsymbol{x}|\omega_k)$ is the class-conditional probability density function (pdf) conditioned on ω_k, $P(\omega_k)$ is the prior probability for ω_k and $p(\boldsymbol{x})$ is the unconditional pdf. Naïve Bayes classifier (NB) assumes conditional independence between the features and calculates the class-conditional pdf as a product of n individual pdf's

$$p(\boldsymbol{x}|\omega_k) = \prod_{i=1}^{n} p(x_i|\omega_k). \tag{1}$$

If the independence assumption does not hold, then the approximation of $p(\boldsymbol{x}|\omega_k)$ is inaccurate, which *may* lead to misclassifications. The surprising success of NB has been attributed to various estimation properties (Hand and Yu, 2001; Domingos and Pazzani, 1997).

- NB estimates fewer parameters than other popular models, therefore it is less prone to overtraining, especially for small sample sizes.
- The traditional pre-selection of features tends to eliminate correlated features anyway therefore the independence assumption may nearly hold for the remaining feature subset.

The most important explanation though lies in the fact that the conditional independence is only a sufficient but not a necessary condition for optimality of NB (Hand and Yu, 2001; Domingos and Pazzani, 1996, 1997; Zhang, 2004; Rish, 2001; Rish et al., 2001). Indeed, the *accuracy of approximation is irrelevant* as long as for any $\boldsymbol{x}$ the largest posterior probability corresponds to the same class as with the true posterior probabilities. In fact, if there are more than two classes, even the order of the other posterior probabilities is irrelevant. It seems though that the quest for *quantifying* the degree of dependence which NB can tolerate, started by Langley et al. (1992), is still on-going. Domingos and Pazzani (1996, 1997), Rish (2001), Rish et al. (2001), Zhang (2004) and others have identified cases where NB is optimal and other cases where it is not. Consider for example continuous-valued variables and two Gaussian classes. If the covariance matrices for the classes are diagonal, then the features are independent and NB is the optimal model. The class-conditional pdf's can be decomposed into individual Gaussians and calculated as in (1). If the features are dependent, and the covariance matrices are equal, $\Sigma_1 = \Sigma_2$, then NB will also recover the correct (linear) classification boundary despite of the flawed approximation of the pdf's. There are however, limits on the abilities of NB to reach near-optimal performance. Take for example a linearly separable pair of non-Gaussian classes. A linear classifier trained by the perceptron algorithm is guaranteed to learn the classification boundary while NB is not.

3. Optimality of NB for two binary features

Let $\boldsymbol{x} = [x_1, x_2]^{\mathrm{T}}$ where $x_1, x_2 \in \{0, 1\}$. We assume that we have complete knowledge of the true probabilities in the problem, i.e., all $P(\boldsymbol{x}|\omega_j)$, $j = 1, 2$, are given[1] for all four values of $\boldsymbol{x}$. To facilitate the algebraic manipulations, the eight probabilities are denoted as $a, b, \ldots, h$ as shown in Table 1(a) and (b).

While dependence is not defined in a unique commonly agreed way in the space of categorical variables, for quantitative variables such measures exist. To evaluate dependence, we shall treat the "0" and the "1" as numbers and will calculate covariance between the two binary features, separately for each class. The mean for x_1 given class ω_1 is $\mu_1 = 0 \times (a + b) + 1 \times (c + d) = c + d$. The mean for x_2 is respectively $\mu_2 = b + d$. The covariance is the expectation of $(x_1 - \mu_1)(x_2 - \mu_2)$ (summed across the four values and weighted by the respective probability)

$$\begin{aligned}\mathrm{Cov}(x_1, x_2|\omega_1) = {} & a(0 - (c + d))(0 - (b + d)) \\ & + b(0 - (c + d))(1 - (b + d)) \\ & \times c(1 - (c + d))(0 - (b + d)) \\ & + d(1 - (c + d))(1 - (b + d)) = ad - bc.\end{aligned} \tag{2}$$

Proposition. *Let* $\boldsymbol{x} = [x_1, x_2]^{\mathrm{T}}$ *where* $x_1, x_2 \in \{0, 1\}$, *and let* ω_1 *and* ω_2 *be the classes of interest with* $P(\omega_1) = P(\omega_2) = \frac{1}{2}$. *If* $\mathrm{Cov}(x_1, x_2|\omega_1) = \mathrm{Cov}(x_1, x_2|\omega_2)$, *then the Naïve Bayes classifier (NB) is optimal for this problem.*

Proof. For NB to make a mistake for some $\boldsymbol{x}$, one of the following must be true:

$$P(\boldsymbol{x}|\omega_1)P(\omega_1) > P(\boldsymbol{x}|\omega_2)P(\omega_2) \quad \text{and}$$
$$P(x_1|\omega_1)P(x_2|\omega_1)P(\omega_1) < P(x_1|\omega_2)P(x_2|\omega_2)P(\omega_2)$$

or

$$P(\boldsymbol{x}|\omega_1)P(\omega_1) < P(\boldsymbol{x}|\omega_2)P(\omega_2) \quad \text{and}$$
$$P(x_1|\omega_1)P(x_2|\omega_1)P(\omega_1) > P(x_1|\omega_2)P(x_2|\omega_2)P(\omega_2).$$

[1] We shall denote probability mass functions by capital P.

Table 1
Class-conditional probability mass functions for classes ω_1 and ω_2 for two binary features

		$x_2 = 0$	$x_2 = 1$
(a) Class ω_1			
	$x_1 = 0$	a	b
	$x_1 = 1$	c	d
$a+b+c+d=1$			
(b) Class ω_2			
	$x_1 = 0$	e	f
	$x_1 = 1$	g	h
$e+f+g+h=1$			

Without loss of generality, consider $\boldsymbol{x} = [0,0]^{\mathrm{T}}$ and let

$$a > e. \tag{3}$$

For NB to make a mistake and assign ω_2 to $\boldsymbol{x} = [0,0]^{\mathrm{T}}$, we must have

$$\frac{1}{2}P(x_1 = 0|\omega_1)P(x_2 = 0|\omega_1) < \frac{1}{2}P(x_1 = 0|\omega_2)P(x_2 = 0|\omega_2), \tag{4}$$

$$(a+b)(a+c) < (e+f)(e+g),$$
$$(a+b)(a+c) - (e+f)(e+g) < 0,$$
$$a^2 + ac + ab + bc - e^2 - eg - ef - fg < 0. \tag{5}$$

From the equivalence of the covariances, $ad - bc = eh - fg$,

$$bc = ad - eh + fg. \tag{6}$$

Substituting in (5),

$$a^2 + ac + ab + ad - eh + fg - e^2 - eg - ef - fg < 0,$$
$$a(a + c + b + d) - e(e + f + g + h) < 0,$$
$$a < e. \tag{7}$$

The above result contradicts (3) therefore (4) cannot be true and NB makes the same decision as the Bayes classifier would. The same argument will hold for the other three values of $\boldsymbol{x}$. □

Unfortunately, this optimality argument does not hold even for the simple case when the two classes are not equiprobable. Denote $p = P(\omega_1)$ and respectively $1 - p = P(\omega_2)$. An error will occur for $\boldsymbol{x} = [0,0]^{\mathrm{T}}$ if $pa > (1-p)e$ and $p(a+c)(a+b) < (1-p)(e+g)(e+f)$. The former is equivalent to

$$a - \left(\frac{1-p}{p}\right)e > 0. \tag{8}$$

Developing the latter leads to

$$a - \left(\frac{1-p}{p}\right)e < \frac{2p-1}{p}\underbrace{(eh - gf)}_{\text{the covariance}} \tag{9}$$

In order to force a contradiction, the right-hand side of (9) should be required to be negative. For this to hold, we need either $p > 0.5$ and negative covariance or $p < 0.5$ and positive covariance. Notice that if the covariance is zero (independence), then the contradiction is in place for any p, and NB is optimal.

It is interesting to find out whether there are conditions on p and the sign of the covariance for which NB is guaranteed to be optimal. Suppose that we do require that $p > 0.5$ and the covariance is negative. In this case, for $a > e$, NB is guaranteed to make the correct decision because (8) and (9) cannot hold simultaneously. It is possible that $pb > (1-p)f$ (can be shown by an example), therefore

$$b - \left(\frac{1-p}{p}\right)f > 0. \tag{10}$$

The corresponding inequality for NB is

$$b - \left(\frac{1-p}{p}\right)f < -\frac{2p-1}{p}(eh - gf). \tag{11}$$

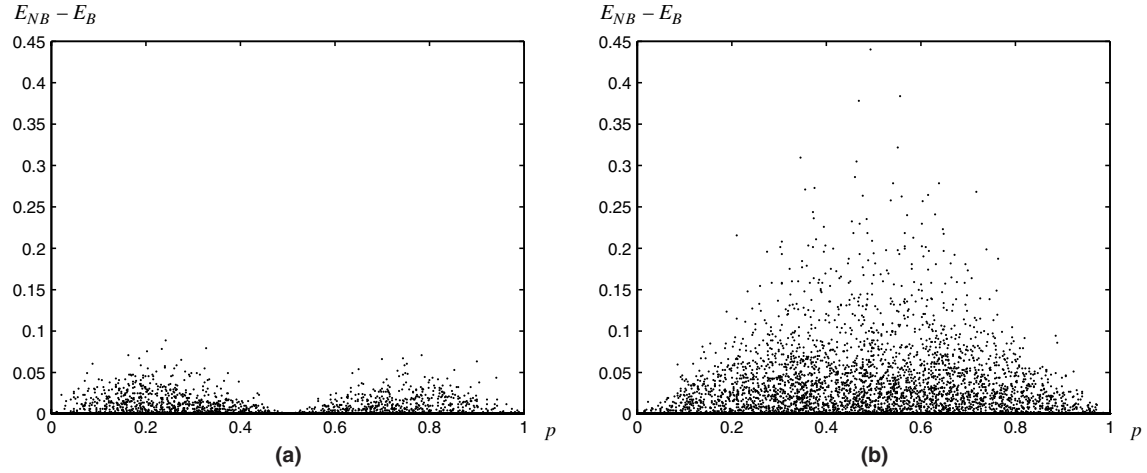


Fig. 1. Scatterplot of the error difference between Naïve Bayes classifier (NB) and the Bayes (optimal) classifier versus the prior probability for class ω_1. (a) Equal covariances and (b) no restrictions.

As the RHS is positive, this inequality may or may not hold together with (10). This argument shows that there is no simple condition that guarantees optimality of NB for two classes, two binary features and equal class-conditional covariances when prior probabilities are not equal.

To evaluate the extent to which prior probability influences the optimality of NB in the two-feature two-class case, simulation experiments were carried out. 10 000 random sets of probabilities $a, b, \ldots, h$ and p were generated so that the class-conditional covariances were equal. Fig. 1(a) shows the scatterplot of the differences between NB error and Bayes error, $E_{\mathrm{NB}} - E_{\mathrm{B}}$, for the 10 000 data points versus the prior probability p. Another set of 10 000 sets of probabilities was generated, this time without the restriction of equal covariances. The scatterplot is given in Fig. 1(b). It is clear that the restriction of equal covariances, although not guaranteeing optimality, brings NB very close to the Bayes (optimal) classifier. For equiprobable classes, $p = 0.5$, NB is indeed optimal, as seen in Fig. 1(a). If covariances are not equal, the largest discrepancies between NB and the optimal classifier occur for p about 0.5.

We also calculated the mean $E_{\mathrm{NB}} - E_{\mathrm{B}}$. For equal covariances, this value was 0.0018 (std 0.0066) and for the unrestricted probabilities, the mean was 0.0135 (std 0.0320). NB was not optimal in about 13% of the cases with equal covariances and in about 32% of the cases with no restriction. This shows that equal covariances are a strong prerequisite for near-optimal performance of NB.

4. Optimality of NB for three binary features

It is interesting whether the results from the previous section carry forward for more than two features.

Let $\boldsymbol{x} = [x_1, x_2, x_3]^{\mathrm{T}}$, where $x_1, x_2, x_3 \in \{0, 1\}$. Consider again the two-class problem with $P(\omega_1) = P(\omega_2) = \frac{1}{2}$. Assume that the pairwise covariances are equal across the two classes for all pairs of features, i.e.,

$$\mathrm{Cov}(x_i, x_j|\omega_1) = \mathrm{Cov}(x_i, x_j|\omega_2) = C_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, 3, \quad i \neq j. \tag{12}$$

The following example demonstrates that NB is *not* optimal for this case. Table 2 shows the class-conditional probability mass functions for the two classes.

The class-conditional pairwise covariances are equal for ω_1 and ω_2 and are as follows: -0.0434 between x_1 and x_2, 0.0230 between x_1 and x_3, and 0.0880 between x_2 and x_3.

The marginal distributions are given as the bottom row in Table 2. Using these, we have

$$P([0,0,0]^{\mathrm{T}}|\omega_1) < P([0,0,0]^{\mathrm{T}}|\omega_2) \quad (0.2706 < 0.2856)$$

and

$$P(x_1 = 0|\omega_1)P(x_2 = 0|\omega_1)P(x_3 = 0|\omega_1) > P(x_1 = 0|\omega_2)P(x_2 = 0|\omega_2)P(x_3 = 0|\omega_2) \quad (0.2226 > 0.1974).$$

While the Bayes classifier will label $\boldsymbol{x} = [0,0,0]^{\mathrm{T}}$ in ω_2, NB will label it in ω_1. This shows that having equal class-conditional covariances is not a sufficient condition for optimality of NB for three features even for equiprobable classes.

To estimate the effect of equal covariances on the optimality of NB, we carried out a simulation study for the problem with three binary features and two equiprobable classes. 10 000 random sets of probability mass functions, $P(\boldsymbol{x}|\omega_1)$ and $P(\boldsymbol{x}|\omega_2)$, were generated where the three covariances C_{12}, C_{13} and C_{23}, were the same for both classes. (The details of the generation procedure are given in Appendix A.) Fig. 2 shows the sorted values of $E_{\mathrm{NB}} - E_{\mathrm{B}}$ for the 20 000 points. The curve with the equal covariance restriction lies underneath the curve where the distributions were generated without the restriction indicating that equal covariances are beneficial. In about 92% of the cases generated without restriction NB makes at least one mistake out of the eight possible values for $\boldsymbol{x}$. For the equal covariances case this figure is about 57%. Also, the mean error differences are 0.0564 (std 0.0486) for the unrestricted case and 0.0157 (std 0.0294) for the equal covariances case.

The results show that equal dependencies of second order are insufficient to guarantee optimality of NB for three features. An additional measure of dependency may therefore be useful. Our first choice was the three-way Q

Table 2
Joint and marginal class-conditional probability mass functions for classes ω_1 and ω_2 for three binary features and equal covariances across the two classes

x_1	x_2	x_3	$P(\boldsymbol{x})$	x_1	x_2	x_3	$P(\boldsymbol{x})$
(a) Class ω_1				*(b) Class ω_2*			
0	0	0	0.2706	0	0	0	0.2856
0	0	1	0.0235	0	0	1	0.0233
0	1	0	0.2435	0	1	0	0.1104
0	1	1	0.2070	0	1	1	0.2464
1	0	0	0.1164	1	0	0	0.0989
1	0	1	0.0428	1	0	1	0.1214
1	1	0	0.0291	1	1	0	0.0654
1	1	1	0.0671	1	1	1	0.0486
0.2554	0.5467	0.3404		0.3343	0.4708	0.4397	
$\underbrace{\qquad\qquad\qquad}_{P(x_i=1\|\omega_1)}$				$\underbrace{\qquad\qquad\qquad}_{P(x_i=1\|\omega_2)}$			

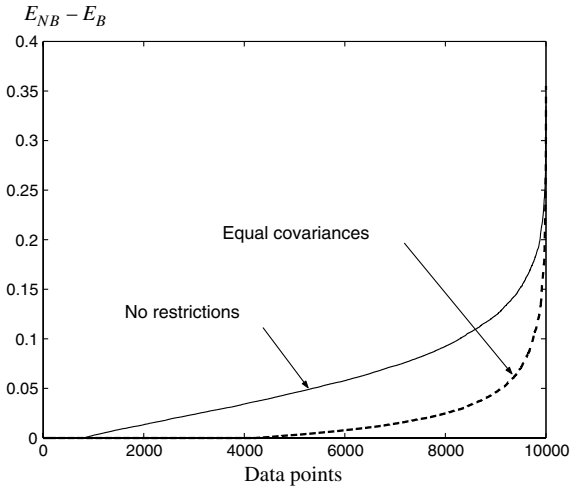


Fig. 2. Sorted values of the error differences for the 10 000 points with and without equal covariances.

statistic (Yule, 1900). Consider the two-way table with probabilities as in Table 1(a). The odds ratio is $r = ad/bc$. Q is meant to serve as a correlation measure varying between -1 and 1 with value 0 corresponding to independence. The transformation which achieves this is $(r-1)/(r+1)$, leading to

$$Q = \frac{ad - bc}{ad + bc}. \tag{13}$$

The numerator of Q is the covariance between x_1 and x_2 as in (2), so for independent variables $Q = 0$. For a three-way table, there are two odds ratios, r_1 and r_2, e.g., for $x_3 = 0$ and $x_3 = 1$, respectively. Their ratio, r_1/r_2 is now normalized to give the three-way Q. Denote by P_{uvw} the probability that $x_1 = u$, $x_2 = v$ and $x_3 = w$, where $u, v, w \in \{0, 1\}$. Then

$$Q_{123} = \frac{P_{111}P_{001}P_{010}P_{100} - P_{000}P_{011}P_{101}P_{110}}{P_{111}P_{001}P_{010}P_{100} + P_{000}P_{011}P_{101}P_{110}}. \tag{14}$$

The information included in Q_{123} is additional to the two-way Q's, which is the reason to include it in this study. The question is whether similar values of Q_{123} for ω_1 and ω_2 correspond to NB being close to optimal. Fig. 3(a) plots the value of the error differences, $E_{\mathrm{NB}} - E_{\mathrm{B}}$, against the difference $Q_{123}(\omega_1) - Q_{123}(\omega_2)$ for the 10 000 data points with equal covariances and (b) gives the plot for the unrestricted case.

The equal covariances case shows a marked pattern whereby similar Q_{123}'s for the two classes bring NB close to optimality (lower error difference). This pattern is not clearly visible for the general case in Fig. 3(b). The pattern in Fig. 3(a) suggests that there might be an optimality condition where the corresponding class-conditional covariances are equal and also $Q_{123}(\omega_1) = Q_{123}(\omega_2)$.

Distances between distributions have been extensively used in pattern recognition for estimating bounds for the Bayes error of classifiers and specifically as criteria for feature selection (Webb, 1999; van der Heijden et al., 2004; Devijver and Kittler, 1982). We carried out simulation experiments as before with three distances in place of Q_{123}: divergence, Bhattacharyya distance and Matusita distance. All three of them estimate how far apart the joint distribution is from the distribution reconstructed through the independence model. The farther apart these distributions are, the larger the dependence between the features. Let M_1 and M_2 be the values of such a measure for ω_1 and ω_2, respectively. Again, we want to find out whether similar values of M_1 and M_2 mean that NB is close to optimality.

Let $\boldsymbol{x}$ be a discrete variable. We consider two probability mass functions:- the true joint distribution, $P^*(\boldsymbol{x})$, and the distribution derived through the independence model, $P_{\mathrm{NB}}(\boldsymbol{x})$. NB uses the latter while Bayes classifier uses the former. Divergence is a popular measure akin to Kullback–Leibler divergence (Webb, 1999; van der Heijden et al., 2004). We used the discrete calculation

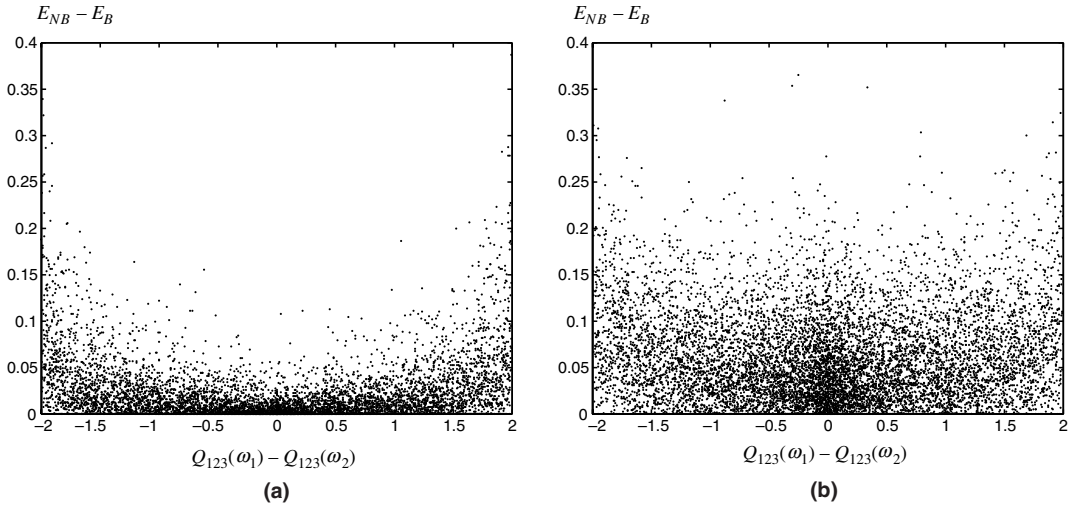
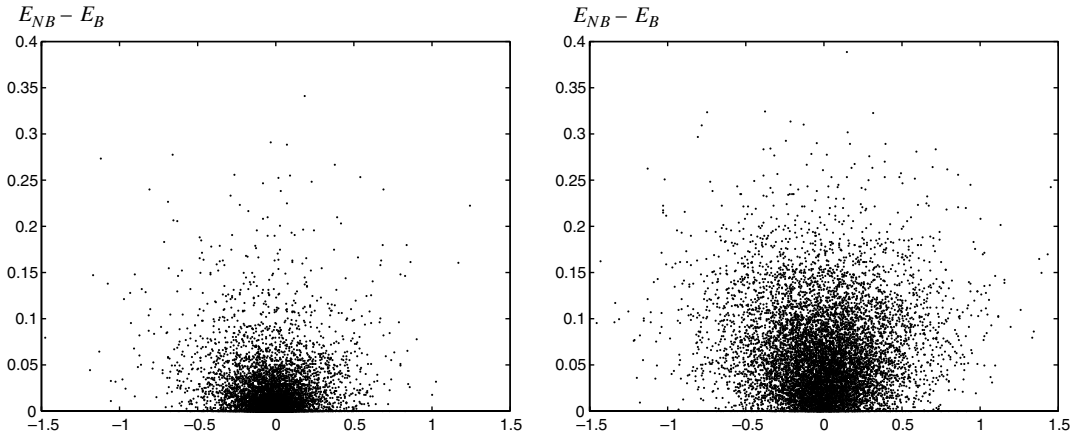


Fig. 3. Scatterplot of the error difference between Naïve Bayes classifier (NB) and the Bayes (optimal) classifier versus $Q_{123}(\omega_1) - Q_{123}(\omega_2)$ with and without equal covariances for the two classes. (a) Equal covariances and (b) no restrictions.



$$\text{Divergence} = \sum_{x} (P^*(\boldsymbol{x}) - P_{\text{NB}}(\boldsymbol{x})) \log\left(\frac{P^*(\boldsymbol{x})}{P_{\text{NB}}(\boldsymbol{x})}\right) \tag{15}$$

The results with the Bhattacharyya and Matusita distances were very similar to these with divergence so we left them out of this paper. Fig. 4 shows the scatterplots of $E_{\text{NB}} - E_{\text{B}}$ versus Divergence(ω_1) − Divergence(ω_2). Interestingly, there is no evidence that similar degrees of feature dependence for the two classes brings NB close to optimality for either the general case or the case of equal covariances. This comes to show that Q_{123} indeed complements the covariances in that it provides additional insight into possible optimality conditions for NB. Divergence, on the other hand, is not indicative in this respect. The span of the divergence differences is only slightly reduced for the distributions with equal covariances. The difference between plots (a) and (b) in Fig. 4 merely shows what we already observed, that NB is closer to the optimal classifier when covariances between features are mirrored for the two classes.

5. Conclusions

This paper looks for optimality conditions for the Naïve Bayes classifier (NB). In Section 3 we prove that for two binary features and two equiprobable classes NB is optimal for dependent features as long as the covariances for the two classes are equal. We also show that this optimality does not hold for different prior probabilities for the classes. Despite not optimal, NB is close to the optimal classifier for the case of equal covariances (Fig. 1). Unfortunately this optimality condition does not carry forward to three features as shown in Section 4. Equal covariances again ensure that NB is closer to optimality than it is in the general case but there is "outstanding dependency" of higher order not accounted for by the pairwise covariances. We pick four measures of (non-pairwise dependency and look into the hypothesis that similar values of these measures for the two classes may complement our optimality condition. Only the three-way Q_{123} showed potential to be considered for the optimality condition (Fig. 3). Disappointingly, the popular divergence measure did not appear to be useful (Fig. 4). The above results are intended as a step in the on-going quest for building a set of optimality conditions for NB (Langley et al., 1992; Domingos and Pazzani, 1997; Hand and Yu, 2001).

One problem with this topic is that, while the notion of *independence* is well defined, there is no agreed measure of dependency between two discrete features. We could have taken the Q statistic or the correlation between the variables, or a myriad of measures available in the statistical literature (Sneath and Sokal, 1973). The problem is even worse when it comes to measuring dependency for three or more features.

There are several issues here. First, the fact that a relationship cannot be proven beyond the case of two features using pairwise correlations does not mean that such a relationship does not exist. A relationship may exist with another measure.

Second, we found an interesting pattern for Q_{123} which may prove to be a way forward in expanding (or rather relaxing) the sufficient optimality conditions for NB. However, this seems to be a dead end as there is no definition of higher order Q's.

Third, divergence (Bhattacharyya and Matusita distances as well) did not show any potential worthy of further exploration. According to these measures NB optimality does not depend on how similar the feature dependencies are in the class-conditional distributions. The lack of direct relationship between feature dependence and classification accuracy of NB has been well documented in the literature through extensive experimental studies. Although it has been hypothesized that the distribution of dependencies is more important than the magnitude of the

dependencies itself (Zhang, 2004), there is little evidence in support of this.

Fourth, as Q is limited for up to three features and divergence-like measures are not particularly responsive, it is not clear how an optimality condition can be formulated in the general case. We can speculate that mirrored covariances for the two classes will be beneficial in the higher-dimensional cases. It is possible that different optimality conditions hold for odd and even number of features.

Defining rigorous and general conditions to explain why NB behaves as the optimal classifier is an intriguing research topic fuelled by curiosity rather than practicality. When a data set is available, it is easier to train and test NB than to calculate measures to estimate the chances of NB being optimal. The intuition is different when we come back to the veterinary problem which inspired this study. Since there is no data set, and NB is the only reasonable option, it will be reassuring to discover further conditions which might hold in reality and under which NB is optimal.

Acknowledgements

I am grateful to Dr. Peter D. Cockcroft, Department of Clinical Veterinary Medicine, University of Cambridge, UK, for introducing the BSE problem and data to our group. I also would like to thank Chris Whitaker, School of Psychology, University of Wales, Bangor, for the very helpful discussion on the Q statistic.

Appendix A

The algorithm below was used to generate randomly class-conditional distributions for 2 classes and 2 binary features such that the pairwise covariances are equal for ω_1 and ω_2. The probability mass function (pmf) for class ω_1 for three binary features is shown in Table 3.

Denote the corresponding pmf values for class ω_2 by capital letters $A,\ldots,H$.

Step 1. Generate $A,\ldots,H$ from a uniform random distribution within the unit interval and normalize so that the sum is 1.

Step 2. Form the two-way tables and find the covariances, C_{11}, C_{12} and C_{13} for ω_2. For example, the table for x_1 and x_2 will have entries $(A+E)$, $(B+F)$, $(C+G)$ and $(D+H)$, and the respective covariance will be

$$C_{12} = (A+E)(D+H) - (B+F)(C+G).$$

Step 3. Find the pmf for class ω_1, i.e., calculate $a,b,\ldots,h$. As these eight parameters are bound by three equations for the covariances and one normalizing equation, there are four parameters out of the eight that can be drawn randomly. Suppose we pick randomly e, f, g and h and scale them so that their sum equals a random number between 0 and 1. This scaling is needed so that there is room for the other four parameters between 0 and 1. a, b, c and d are obtained as the solution of the following system of simultaneous equations:

$$\begin{aligned}
&(a+e)(d+h) - (b+f)(c+g) = C_{12},\\
&(a+b)(g+h) - (c+d)(e+f) = C_{13},\\
&(a+c)(f+h) - (b+d)(e+g) = C_{23},\\
&a+b+c+d+e+f+g+h = 1.
\end{aligned}$$

The solution is

$$\begin{aligned}
Z_1 &= \frac{(1-e-f-g-h)(g+h) - C_{13}}{e+f+g+h},\\
Z_2 &= \frac{(1-e-f-g-h)(f+h) - C_{23}}{e+f+g+h},\\
d &= C_{12} - (1-Z_1-Z_2-f-g-h)h\\
&\quad + Z_1Z_2 + Z_1f + Z_2g + fg,\\
a &= d - Z_1 - Z_2 + 1 - e - f - g - h,\\
b &= Z_2 - d,\\
c &= Z_1 - d.
\end{aligned}$$

Step 4. Being probabilities, $a,\ldots,h$ must be between 0 and 1. This is not guaranteed by the solution of the system at Step 3. Therefore the last step is to check whether all values are in $[0,1]$. If not, the new pmf $(a,\ldots,h)$ is discarded and the procedure starts from Step 1.

Table 3
Class-conditional probability mass functions for class ω_1 for three binary features

		$x_2=0$	$x_2=1$
$x_3=0$			
	$x_1=0$	a	b
	$x_1=1$	c	d
$x_3=1$			
	$x_1=0$	e	f
	$x_1=1$	g	h

References

Devijver, P., Kittler, J., 1982. Pattern Recognition: A Statistical Approach. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Domingos, P., Pazzani, M., 1996. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In: Proc. 13th Internat. Conf. on Machine Learning.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–130.

Hand, D.J., Yu, K., 2001. Idiot's Bayes—not so stupid after all? Internat. Statist. Rev. 69, 385–398.

Jamain, A., Hand, D.J., 2005. The Naïve Bayes mystery: a classification detective story. Pattern Recognition Lett.

Langley, P., Iba, W., Thompson, K., 1992. An analysis of Bayesian classifiers. In: Proc. 10th National Conf. on Artificial Intelligence. pp. 399 – 406.

Rish, I., 2001. An empirical study of the Naïve Bayes classifier. In: Proc. Internat. Joint Conf. on Artificial Intelligence, Workshop on "Empirical Methods in A".

Rish, I., Hellerstein, J., Thathachar, J., 2001. An analysis of data characteristics that affect Naïve Bayes performance. Tech. Rep. RC21993, IBM TJ Watson Research Center.

Sneath, P., Sokal, R., 1973. Numerical Taxonomy. WH Freeman & Co.

van der Heijden, F., Duin, R.P.W., de Ridder, D., Tax, D.M.J., 2004. Classification, Parameter Estimation and State Estimation. Wiley, England.

Webb, A., 1999. Statistical Pattern Recognition. Arnold, London, England.

Yule, G., 1900. On the association of attributes in statistics. Phil. Trans. A 194, 257–319.

Zhang, H., 2004. The optimality of Naïve Bayes. In: Proc. 17th Internat. FLAIRS Conf., Florida, USA.