Error-Dependency Relationships for the Naïve Bayes Classifier with Binary Features

Ludmila I. Kuncheva† and Zoë S.J. Hoare‡

Abstract

We derive a tight dependency-related bound on the difference between the NB error and Bayes error for the case of two binary features and two classes. A measure of feature dependency is proposed for multiple features. Simulations and experiments with 23 real data sets were carried out.

Keywords

Pattern recognition; Naïve Bayes classifier; Dependency between features; Classification error; Q statistic

I. Introduction

Naïve Bayes or also "Idiot's Bayes" [4] is a simple and often surprisingly accurate classification technique. Consider an object represented by a feature vector $\mathbf{x} = [x_1, \ldots, x_n]^T$. The problem is to assign the object to one of c predefined classes, $\omega_1, \ldots, \omega_c$. Minimum classification error is guaranteed if the class with the largest posterior probability, $P(\omega_i \,|\, \mathbf{x})$, is chosen. To calculate posterior probabilities, Bayes formula is used with estimates of the prior probabilities, $P(\omega_i)$, and the class-conditional probability density functions (pdf), $p(\mathbf{x} \,|\, \omega_i)$

$$P(\omega_i \,|\, \mathbf{x}) = \frac{P(\omega_i)p(\mathbf{x} \,|\, \omega_i)}{\sum_{j=1}^{c} P(\omega_j)p(\mathbf{x} \,|\, \omega_j)}, \qquad i = 1, \ldots, c. \tag{1}$$

Obtaining an accurate estimate of the joined pdf is difficult, especially if the dimensionality of the feature space, n, is large. The "naïvety" of the Naïve Bayes model comes from the fact that the features are assumed to be conditionally independent. In this case the joined pdf for a given class is the product of the marginal pdfs

$$p(\mathbf{x} \,|\, \omega_i) = \prod_{j=1}^{n} p(x_j \,|\, \omega_i), \qquad i = 1, \ldots, c. \tag{2}$$

Accurate estimates of the marginal pdfs can be obtained from much smaller amounts of data compared to these for the joint pdf. This makes the Naïve Bayes classifier (NB) so attractively simple.

The assumption of conditional independence among features may look too restrictive. Nonetheless NB has demonstrated robust and accurate performance across various domains, often reported as "surprisingly" accurate, even where the assumption is clearly false [4]. The research on NB in the past 15 years has followed two major ideas. One is developing variants of NB in which the independence assumption is relaxed or partly avoided [3, 5–7, 9, 10, 14]. The second is to find out why NB works so well for problems where features are not independent [2, 4, 12, 13]. We shall call the two research trends 'new-models' and 'new-explanations', respectively. Our study belongs in the 'new-explanations' trend.

As pointed out in various studies [2, 4], the optimality of NB (or any classifier making a decision based on continuous-valued outputs for the c classes) will hold so long as the estimated output for the class with the largest true posterior probability, $P(\mathbf{x} \,|\, \omega_k)$, exceeds all the other outputs. The probability estimates do not have to be correct, they do not have to be completely order-correct, and even do not have to be *probability* estimates. Conditions for optimality or non-optimality of NB in special cases have been identified from which we picked the following four results

- NB is optimal for both complete independence and complete *dependence* between the features [12, 13]. Indeed, when all features can be derived from one another, the problem is, in effect, one-dimensional and NB is optimal anyway. Thus the relationship between dependency and NB optimality is non-monotonic.
- The degree of feature dependencies and the accuracy of NB are not directly correlated [2, 12, 13].
- Dependencies between features may be "cancelled out" so that NB is near optimal. Distributions of dependencies should be taken into account [16, 17].
- NB is optimal for two equiprobable classes and two binary features with equal class-conditional covariances [8].

The lack of a direct relationship between dependency and accuracy of NB does not preclude finding a bound on the accuracy as a function of dependency. A major problem in these analyses is that there is no agreed definition

†School Computer Science, Bangor University, UK, e-mail: l.i.kuncheva@bangor.ac.uk
‡Knowledge Support Systems Limited, UK, e-mail: hoarez@kssg.com

of dependency among more than two features. Hence we start with two features and use the Q statistic to measure dependency. We derive a tight bound on the difference between the NB error and the Bayes error as a function of Q. In search for such a bound in higher-dimensional spaces we look for a pattern of relationship between dependency and accuracy through experiments with real data. In this study a measure of dependency based on Q is proposed to account for the differences in the dependency distributions between the classes.

The rest of the paper is organised as follows. In Section 2 the Q measure is explained and a bound on the difference between the classification error of NB and the Bayes error is derived as a function of Q. Section 3 contains the new measure of dependency and an experimental study looking for a relationship between dependency and accuracy for larger number of features. Section IV offers our final comments and conclusions.

II. Error-dependency relationship for two classes and two binary features

A. Dependencies between features

Studying the relationship between feature dependency and the classification error/accuracy of NB is difficult for many reasons. The critical reason is that there is no agreed concept of dependency between more than two variables, especially when multiple classes are also considered. The trivial approach is to take the average of the pairwise dependencies as the measure of dependency for the whole set. Pairwise dependencies may be measured by Pearson correlation (linear dependency between two continuous-valued random features), mutual information (probabilistic dependency between two random-valued features of any type) or Yule's Q statistic for two binary features [15].

Let $x_1, x_2 \in \{0, 1\}$ be binary features and ω_1 and ω_2 be the two mutually exclusive classes. Denote the two joint class-conditional probability mass functions (pmf) as

$$P(00 \mid \omega_1) = a, \;\; P(01 \mid \omega_1) = b, \;\; P(10 \mid \omega_1) = c, \;\; P(11 \mid \omega_1) = d \tag{3}$$

$$P(00 \mid \omega_2) = e, \;\; P(01 \mid \omega_2) = f, \;\; P(10 \mid \omega_2) = g, \;\; P(11 \mid \omega_2) = h \tag{4}$$

where $a+b+c+d = 1$ and $e+f+g+h = 1$. Without loss of generality we assume that all probabilities belong in the open interval (0,1), i.e., values 0 and 1 are not allowed. The dependency between features x_1 and x_2 is measured by the Q statistic. The two conditional dependencies for classes ω_1 and ω_2, respectively are

$$Q_1 = \frac{ad - bc}{ad + bc}, \quad Q_2 = \frac{eh - fg}{eh + fg}. \tag{5}$$

The Q statistic corresponds intuitively to correlation for continuous-valued features. A value $Q = 0$ means that the two features are statistically independent. Values of Q close to 1 show that the two features tend to take simultaneously the same values whereas values of Q close to -1 show that the two features tend to take simultaneously the opposite values.

In classical NB all features are assumed to be independent for each class. We relax this assumption by letting *some* dependencies to hold. In our example we shall assume that x_1 and x_2 are conditionally independent for class ω_2, i.e., $Q_2 = 0$. This means that $eh - fg = 0$. We are looking for a relationship between the degree of dependency, Q_1, and the NB classification error.

B. Bayes error

Assume that the classes are equiprobable, i.e., $P(\omega_1) = P(\omega_2) = \frac{1}{2}$. To guarantee minimum error, an object $[x_1, x_2]^T$ should be assigned to the class with the largest posterior probability. Since the prior probabilities are equal and the denominators of both posterior probabilities are equal, the class label will be determined by the corresponding values in the two pmfs. For example, suppose that the object to be labelled is $[1, 0]^T$. We compare the corresponding value from the pmf for ω_1, c, with the one from the pmf for ω_2, g, and decide for the class with the bigger value. Let $c > g$, hence we choose class ω_1. The probability of making this error is the probability of simultaneous occurrence of class ω_2 and object $[1, 0]^T$. This probability is calculated as $P([1, 0]^T \mid \omega_2) \times P(\omega_2) = g \times \frac{1}{2}$. The *Bayes error*, E_B, is the total probability of error across the whole feature space, which in our case has 4 elements

$$E_B = \frac{1}{2} \left(\min\{a, e\} + \min\{b, f\} + \min\{c, g\} + \min\{d, h\} \right). \tag{6}$$

The Bayes error is needed as a benchmark. We want to find out how much the classification error of NB deviates from E_B for various degrees of dependency Q_1.

C. Naïve Bayes error

To build the Naïve Bayes classifier, we treat the features as conditionally independent, conditioned separately upon each class label. By definition, x_1 and x_2 are independent for class ω_2. For class ω_1 we will re-construct the pmf. Assuming independence, the probability for $x_1 = 0$ and $x_2 = 0$ given class ω_1 is

$$P(x_1 = 0 \;\text{ and }\; x_2 = 0 \mid \omega_1) \;\; = \;\; P(x_1 = 0 \mid \omega_1) \times P(x_2 = 0 \mid \omega_1) = (a + b) \times (a + c). \tag{7}$$

Thus, NB will label the objects according to a new pair of pmfs

$$\begin{aligned} P(00\,|\,\omega_1) &= (a+b)(a+c), \qquad P(01\,|\,\omega_1) = (a+b)(b+d), \\ P(10\,|\,\omega_1) &= (a+c)(c+d), \qquad P(11\,|\,\omega_1) = (b+d)(c+d) \end{aligned} \tag{8}$$

and given $Q_2 = 0$,

$$P(00\,|\,\omega_2) = e, \;\; P(01\,|\,\omega_2) = f, \;\; P(10\,|\,\omega_2) = g, \;\; P(11\,|\,\omega_2) = h. \tag{9}$$

Denote by E the error of the NB classifier. There are two possibilities for each $\mathbf{x} = [x_1, x_2]^T$: either NB makes the same decision as the (true) Bayes classifier or NB makes the alternative decision. Consider again the example above where $x_1 = 1$ and $x_2 = 0$. If NB chooses the same class label as the Bayes classifier, the corresponding error component in E is the same as the component in E_B (6), $\frac{1}{2}\min\{c, g\}$. If NB chooses the alternative class label, the corresponding error component in E will be $\frac{1}{2}\max\{c, g\}$. The largest possible error of the NB classifier will occur if it makes mistakes for all four objects. Summing across all four objects (assuming mistakes everywhere) and taking out the Bayes error, we obtain

$$\begin{aligned} \Delta E &= E - E_B \\ &= \frac{1}{2}\left(\max\{a, e\} + \max\{b, f\} + \max\{c, g\} + \max\{d, h\}\right) \\ &\quad -\frac{1}{2}\left(\min\{a, e\} + \min\{b, f\} + \min\{c, g\} + \min\{d, h\}\right) \\ &= \frac{1}{2}\left(\,|\,a - e\,| \,+\, |\,b - f\,| \,+\, |\,c - g\,| \,+\, |\,d - h\,|\,\right). \end{aligned} \tag{10}$$

D. Dependency-error relationship

To get an initial impression about the dependency-error relationship, we generated randomly 10000 pairs of pmfs as in (3) and (4) so that $Q_2 = 0$. We calculated Q_1, re-constructed the pmf for ω_1 through (8) and calculated ΔE. The 10000 points $(Q_1, \Delta E)$ are plotted in Figure 1 (a). The figure shows that

- ΔE can only be positive (trivial)
- ΔE is 0 for $Q = 0$ (expected: NB is optimal ($E = E_B$) when the features are conditionally independent for all the classes)
- ΔE may be 0 for any degree of dependency. This shows again that the independence assumption is a sufficient but not a necessary condition for optimality of NB. On the contrary, the bottom left and right corners of the scatterplot are populated with points, indicating that there is no clear-cut pattern to the relationship between ΔE and Q_1. This resonates with previous literature in that the degree of dependency is not directly related to the error.
- The scatterplot is symmetrical about $Q_1 = 0$. This was to be expected because the encoding of the binary features as absent = 0 and present = 1 is arbitrary. If the 0 and the 1 were swapped for one of the features and kept for the other feature, Q_1 will only change its sign. The way of encoding has no effect on E_B or E. Therefore it is sufficient to consider $|\,Q_1\,|$.
- The cloud of points has a pronounced shape which suggests the possibility of finding a rigorous upper bound on ΔE.

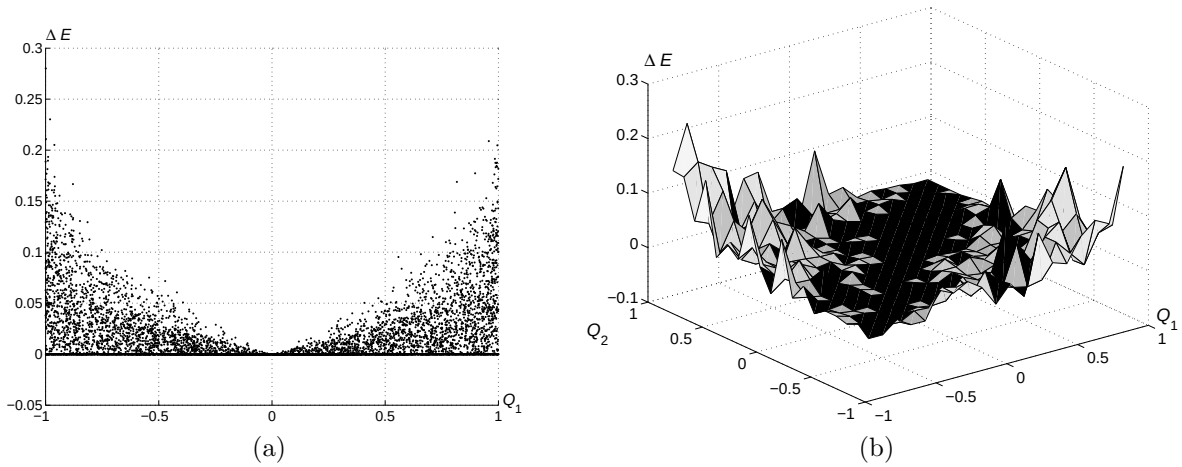
Figure 1 (b) shows the ΔE surface, approximated on 10000 pairs (Q_1, Q_2) obtained from randomly generated pmfs. This time neither of the pmfs has been restricted to correspond to independent features. NB is guaranteed to be optimal for $Q_1 = Q_2 = 0$. The dark region along the $(-1, -1)$–(1,1) diagonal in Figure 1 (b) corresponds to the area where NB is close to the optimal Bayes classifier, i.e., as long as $Q_1 \approx Q_2$, NB is approximately optimal. If we have large values of both Q_1 and Q_2, then the two features are practically identical. They may not be of much value as a pair but will be equally good/bad for both NB and the Bayes classifier. If both Q_1 and Q_2 have large negative values, then a swap of the 0 and the 1 of one of the features will again make the two features almost identical, and the above explanation holds. This finding may be used towards explaining the robust performance of NB when features are known to defy the assumption of conditional independence.

In fact, $Q_1 = Q_2$ may be beneficial but it does not *guarantee* optimality of NB. We note, however, that NB is optimal for equal priors and equal *covariances*, i.e., $ad - bc = eh - fg$ [8].

Many currently used measures of dependency average the class-conditional dependencies weighted by the prior probabilities. If we take the average across the two classes $Q = (Q_1 + Q_2)/2$, the results will be useless. For $Q = 0$, we will trace the whole back diagonal, from (-1,1) to (1,-1), and will get the whole range of values of ΔE. Thus an average measure of pairwise dependency does smooth out important diversity in the data.

E. A bound on the NB error

Consider again the two pmf pairs (3)-(4) and (8)-(9). Recall that NB makes a decision according to (8)-(9). For example, there will be a misclassification for $[1, 0]^T$ if $c > g$ and $(a + c)(c + d) < g$ (Bayes decision is for ω_1 and NB decision is for ω_2).



(a) (b)

Fig. 1. (a) Scatterplot of ΔE versus Q_1 with 10000 randomly generated pmfs where $Q_2 = 0$; (b) ΔE as a function of (Q_1, Q_2) approximated on 10000 randomly generated pmfs with any Q_2.

Theorem 1. Given is a problem of discriminating between two classes, ω_1 and ω_2 using two binary features. Assume that the two features are conditionally independent for class ω_2. Then the Naïve Bayes classifier can at most misclassify 3 of the 4 elements of the feature space.

Proof. For this proof we shall derive the conditions for NB making 4 errors and will prove that these conditions cannot hold together.

Denote the pmfs for class ω_1 and ω_2 as in (3) and (4), respectively. Denote $Q = \frac{ad-bc}{ad+bc}$. For the Naive Bayes classifier to assign the wrong class label for $[0,0]^T$, we have one of the following two situations

$$a > e \quad \text{and} \quad (a+b)(a+c) < e, \text{ or} \tag{11}$$
$$a < e \quad \text{and} \quad (a+b)(a+c) > e \tag{12}$$

Simple algebraic manipulation of (11) leads to

$$Q(ad+bc) \quad > \quad a - e. \tag{13}$$

We require Q to be positive for the error to be possible. Since $a > e$, we can write $Q(ad+bc) > |\, a - e \,|$. Starting from (12), we arrive at $-Q(ad+bc) > e - a$, which can only hold for negative Q because $e > a$. As argued in the previous section, the bound will be symmetrical about $Q = 0$, therefore we can consider only the case $Q > 0$.[1] To guarantee errors for the remaining 3 elements of the feature space, we follow (13)

$$b < f \text{ and } (a+b)(b+d) > f \quad \rightarrow \quad Q(ad+bc) > |\, b - f \,| \tag{14}$$
$$c < g \text{ and } (a+c)(c+d) > g \quad \rightarrow \quad Q(ad+bc) > |\, c - g \,| \tag{15}$$
$$d > h \text{ and } (b+d)(c+d) < h \quad \rightarrow \quad Q(ad+bc) > |\, d - h \,| \tag{16}$$

Thus, for NB to make 4 misclassifications, the following system of simultaneous equations and inequalities must hold

$$0 < a, b, c, d, e, f, g, h < 1 \qquad \text{probability restriction} \tag{17}$$
$$a + b + c + d = 1 \qquad \text{pmf restriction for class } \omega_1 \tag{18}$$
$$e + f + g + h = 1 \qquad \text{pmf restriction for class } \omega_2 \tag{19}$$
$$eh - fg = 0 \qquad \text{conditional independence for } \omega_2, \text{ assumed} \tag{20}$$
$$D = ad - bc, D > 0 \qquad \text{note that } D = Q(ad+bc) \tag{21}$$
$$a > e, \;\; D > a - e \qquad \text{necessary and sufficient for an error at } [0,0]^T \tag{22}$$
$$\text{inequalities (14) to (16)}$$

Next we show that the system has no solution. Denote

$$\epsilon_1 = a - e, \quad \epsilon_2 = f - b, \quad \epsilon_3 = g - c, \quad \epsilon_4 = d - h. \tag{23}$$

[1] The derivation for $Q < 0$ follows the same logic and leads to the same conclusion.

Note that all ϵ_i are strictly positive. Substituting in $eh - fg = 0$,

$$(a - \epsilon_1)(d - \epsilon_4) - (b + \epsilon_2)(c + \epsilon_3) = 0. \tag{24}$$

As b and c are both positive, by replacing ϵ_2 and ϵ_3 with D, which is strictly greater than both, the left-hand side becomes strictly negative. By replacing ϵ_1 and ϵ_2 with D, which is strictly greater than both, the left-hand side becomes even smaller. We must only make sure that $(a - D)$ and $(d - D)$ are not both negative, because in this case the logic will not hold. Take first $(a - D)$ and recall that $D = ad - bc$. Assume that $a \leq D$. Then $ad \leq Dd$, and also $ad - bc \leq Dd - bc$. Then $D(1 - d) \leq -bc$. The left-hand side is strictly positive and the right-hand side is strictly negative, hence the assumption is invalid and so $a > D$. The same logic leads to $d > D$. Therefore, substituting D for all ϵ_i makes the left-hand side of (24) *strictly* negative

$$(a - D)(d - D) - (b + D)(c + D) \quad < \quad 0. \tag{25}$$

Opening the brackets and cancelling D^2, we arrive at

$$D(1 - a - b - c - d) = 0 \quad < \quad 0. \tag{26}$$

This contradiction completes the proof. ∎

Theorem 2. Given is a problem of discriminating between two classes, ω_1 and ω_2 using two binary features. Denote the pmf for class ω_1 as in (3) and let $Q = \frac{ad-bc}{ad+bc}$. Assume that the two features are conditionally independent for class ω_2. Then the difference between NB error and Bayes error is bounded from above as

$$\Delta E = E - E_B \leq \frac{3Qab(1 - a - b)}{a(1 - Q) + b(1 + Q)}. \tag{27}$$

Proof. Denote the pmf for class ω_2 as in (4). As found in (13)-(16), $Q(ad + bc)$ is greater than any of $|\, a - e \,|$, $|\, b - f \,|$, $|\, c - g \,|$, and $|\, d - h \,|$. Therefore each term in the error difference ΔE (10) is at most $Q(ad + bc)$. According to Theorem 1, maximum 3 mistakes are possible, therefore

$$\Delta E = E - E_B < \frac{3}{2} Q(ad + bc). \tag{28}$$

The bound is actually nonlinear on Q because a, b, c and d are related. First, $a + b + c + d = 1$, and second, $Q = \frac{ad-bc}{ad+bc}$. Hence, we can pick two of the values, e.g., a and b, and express c and d. Through simple algebraic manipulations (28) is transformed to

$$\Delta E < \frac{3Qab(1 - a - b)}{a(1 - Q) + b(1 + Q)}. \tag{29}$$

∎

We note that this bound has a simpler version if we use the *covariance* between the two features, i.e., $Cov_1 = ad - bc$. In this case, assuming that $Cov_2 = eh - fg = 0$, $\Delta E \leq \frac{3}{2} Cov_1$. The covariance, however did not generalise well in the empirical study with multiple features, therefore we continue with Q.

Figure 2 (a) shows 10000 simulated $(Q, \Delta E)$ points with $a = 0.5$, $b = 0.4$ and Figure 2 (b) shows 10000 points with $a = 0.6$, $b = 0.1$. The bound (29) is plotted with a solid line.

The plots give a further insight into the quality of the NB classifier. Most of the points lie within the 1-error area, where ΔE is small, i.e., NB is nearly optimal. Only a small fraction of points reach the 3-error area for both choices of a and b.

Figure 2 also reinforces the observation made earlier that a large difference between the dependency patterns for the classes leads to poorer results. In these figures, Q_2 is kept at 0, and only Q_1 is varied. The further away Q_1 is from 0, the larger the discrepancies are between the two Qs, and so is ΔE. The first derivative of the bound (29) is, positive for any probabilities a and b $(a + b < 1)$

$$\frac{\partial}{\partial Q} \left[\frac{3Qab(1 - a - b)}{a(1 - Q) + b(1 + Q)} \right] = \frac{3ab(1 - a - b)(a + b)}{\left(a(1 - Q) + b(1 + Q)\right)^2} > 0, \tag{30}$$

indicating that ΔE grows with dependence. We also note that Theorem 1 and 2 hold only when $Q_2 = 0$. If this restriction is not in place, it is possible that NB makes errors for all four points of the feature space. Even though the result is limited to the very simple case of two classes and two binary features, it conveys an important message. The degree of dependency between the features may not be a good indicator of the NB error *per se* but may be used to construct bounds thereof.

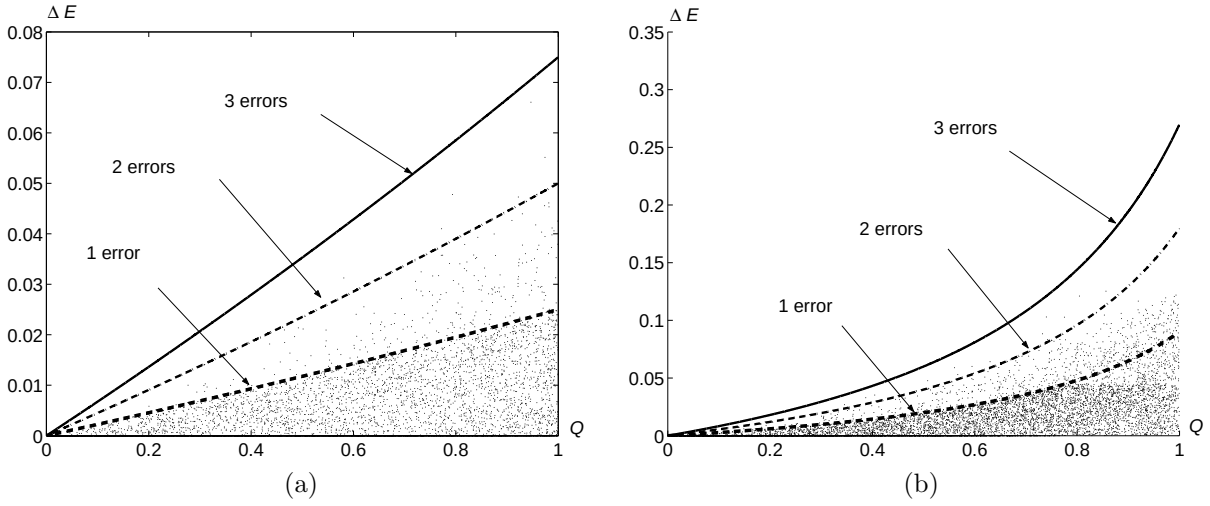


Fig. 2. ΔE as a function of Q and the number of errors for two choices of (a, b): subplot (a) $a = 0.5$, $b = 0.4$; subplot (b) $a = 0.6$, $b = 0.1$

III. An experimental study for multiple binary features

In the light of the results from the previous section distributions of the dependencies rather than their magnitude should be taken into account. Interestingly, the observation that similar dependencies for the two classes ($Q_1 \approx Q_2$) is beneficial for NB (see Figure 1) appears to be limited to the case of two features. The experimental results below show rather the opposite:- large discrepancies between dependencies are generally better than equal dependencies.

A. Measuring dependencies for multiple binary features

Here we use three measures from the literature [2, 12, 13] and propose a new one. Each measure has a value for every pair of features, (i, j), and every class. Thus a notation $M_{i,j}^{(k)}$ will be the value of measure 'M' for features (i, j), $i, j = 1, \ldots, n$, $i \neq j$, and class ω_k, $k = 1, \ldots, c$.

The measures can be organized in a dependency matrix $\mathbf{M}$ of size $n(n-1)/2 \times c$ with entry $M_{i,j}^{(k)}$. In the prefect case scenario, all features are conditionally independent which means that $\mathbf{M}$ is a zero matrix (if we use Q). The idea so far has been to measure how far from this "independence pattern" $\mathbf{M}$ is, hoping that the deviation from independence is related to either E or ΔE.

A measure based on entropy and mutual information has been used before [2, 12, 13]

$$M_{i,j}^{(k)} \equiv I_{i,j}^{(k)} = H_{i,j}^{(k)} - H_i^{(k)} - H_j^{(k)}, \tag{31}$$

where H denotes entropy. $I_{i,j}^{(k)}$ measures how much information is lost if features i and j are used individually rather than together. Denote the joint pmf of pair (i, j) for class ω_k as in (3). Then

$$H_{i,j}^{(k)} = -(a \log_2 a + b \log_2 b + c \log_2 c + d \log_2 d) \tag{32}$$

$$H_i^{(k)} = -((a+b) \log_2 (a+b) + (c+d) \log_2 (c+d)) \tag{33}$$

$$H_j^{(k)} = -((a+c) \log_2 (a+c) + (b+d) \log_2 (b+d)) \tag{34}$$

To find an overall measure of deviation from independence for the pair of features (i, j), the pairwise $I_{i,j}^{(k)}$ are *averaged* across the classes weighted by the prior probabilities, i.e.

$$I_{i,j} = \sum_{k=1}^{c} P(\omega_k) I_{i,j}^{(k)}. \tag{35}$$

To arrive at one final value for the whole data set, we can take maximum or mean across all pairs of features [2]

- $$I_{\max} = \max_{i,j} I_{i,j} \tag{36}$$

- $$I_{\text{mean}} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I_{i,j} \tag{37}$$

TABLE I

Description of the data sets, NB accuracy and the four dependency measures

Data set	n	Objects	Largest prior	NB accuracy	$I_{\max}$	I_{mean}	$H_{\max}$	$Q_{\max}^{\text{diff}}$
crabs	7	200	0.50	0.62	0.71	0.32	0.85	0.51
ecoli	7	336	0.57	0.91	0.15	0.02	0.76	0.60
german	24	1000	0.70	0.74	0.36	0.01	0.91	0.75
glass	9	214	0.67	0.74	0.21	0.04	0.79	0.84
heart	13	303	0.54	0.84	0.08	0.01	0.85	0.81
image	19	210	0.86	0.85	0.59	0.15	0.74	0.86
ionosphere	34	351	0.64	0.77	0.67	0.11	0.89	0.73
iris	4	150	0.67	0.86	0.23	0.11	0.59	0.67
letters	16	1555	0.51	0.94	0.51	0.07	0.85	0.93
liver	6	345	0.58	0.65	0.15	0.03	0.82	0.23
phoneme	5	5404	0.71	0.76	0.11	0.03	0.90	0.71
pima	8	768	0.65	0.70	0.22	0.02	0.89	0.29
satimage	36	6435	0.76	0.69	0.44	0.18	0.74	0.57
sonar	60	208	0.53	0.74	0.44	0.03	0.84	0.66
soybean	35	266	0.68	0.98	0.48	0.03	0.79	1.00
spam	57	4601	0.61	0.89	0.27	0.01	0.88	1.00
spect	22	267	0.79	0.81	0.35	0.03	0.72	0.84
vehicle	18	846	0.75	0.65	0.62	0.13	0.77	0.50
votes	16	232	0.53	0.92	0.21	0.04	0.83	1.00
vowel	11	990	0.91	0.91	0.13	0.03	0.72	0.76
wbc	30	569	0.63	0.92	0.46	0.04	0.81	0.54
wine	13	178	0.67	0.96	0.66	0.13	0.77	0.88
zoo	16	101	0.59	0.98	0.81	0.19	0.81	1.00
Correlation with NB accuracy					−0.2297	0.0555	−0.2181	**0.7185**

Rish (2001) proposed that a better error-dependency relationship may be found using the maximum entropy of the marginal pmfs. Hence, the third measure in our experiment is

$$\bullet \quad H_{\max} = \max_{i=1}^{n} \left(\sum_{k=1}^{c} P(\omega_k) H_i^{(k)} \right). \tag{38}$$

The three measures above take the weighted average across classes. As argued above, similarity or dissimilarity of the dependencies may give a better prediction of E or ΔE. Therefore we propose the following measure using Q. For every pair of features (i, j) and for every pair of classes (k, s), we calculate the absolute difference in the measures as

$$Q_{i,j}^{(k,s)} = \left| P(\omega_k) Q_{i,j}^{(k)} - P(\omega_s) Q_{i,j}^{(s)} \right|. \tag{39}$$

The measure proposed here is

$$\bullet \quad Q_{\max}^{\text{diff}} = \max_{i,j,k,s} Q_{i,j}^{(k,s)}, \quad i, j = 1, \ldots, n, \quad k, s = 1, \ldots, c. \tag{40}$$

B. Results with real data

Table I shows the summary of the 23 data sets used in this experiment.[2] We discretised all the features into binary using the median as the threshold. For all data sets which had more than two classes, classes from 2 to c were grouped into one class and relabelled as "class 2".[3]

NB was built and tested on each data set using a random split of the data into 90% for training and 10% for testing, this process repeated 100 times. The average testing accuracy was taken to be an estimate of $1 - E$. The 4 measures were calculated once on each data set. Table I shows the NB accuracy and the four measures for the 23 data sets. The correlation between NB accuracy and the measures is displayed in the bottom row.

$Q_{\max}^{\text{diff}}$ shows that there is a possible relationship between NB accuracy and the degree of dependency between the features. This correlation seems counterintuitive as it was observed that NB benefited from similar dependencies, and therefore low $Q_{\max}^{\text{diff}}$, for the case of 2 features. While this may still be true for multiple features, we did not encounter

[2]The "crabs" data set is from [11] while all the other data sets are from UCI ML Repository [1].

[3]The two exceptions were the soybean data and the letters data. For the soybean data, class 1 was too small therefore we joined classes 1 to 6 to be class ω_1 and classes 7 to 15 to be ω_2. For the letters data we only used letters 'A' and 'B' as the two classes.

$Q^{\text{diff}}_{\max}$ turns from a measure of discrepancies between feature dependency more into a measure of discrepancy between the class-conditional pmfs. Therefore it becomes a good indicator of the possibility to separate the classes by any classifier, including NB, hence the reasonable correlation shown in Table I. On the other hand, I_{mean}, $H_{\max}$ and $I_{\max}$ behave as intuitively expected showing weak correlations with NB.

As in any experimental study, the results should not be taken as a dogma because the selection of data sets was random. The important finding here is that the difference of the dependencies across the classes is perhaps more relevant than the magnitude of the dependencies themselves.

IV. Conclusions

We view this study as a step towards finding a stronger predictor of the NB error/accuracy and explaining why NB is so successful. The theoretical bound established through Theorems 1 and 2 complies with the current understanding that large deviation from independence allows larger NB error. However, the experimental results show that deviation from independence may not be as important on its own when multiple features are concerned. The proposed feature dependency measure in this case serves more as a measure of discrepancy between the class-conditional pmfs. Thus $Q^{\text{diff}}_{\max}$ becomes a good indicator of the possibility to separate the classes by any classifier, and this explains its correlation with the NB error.

References

[1] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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

[3] N. Friedman, D. Geiger, and M. Goldszmid. Bayesian network classifiers. *Machine Learning*, 29(2):131–163, 1997.

[4] D. J. Hand and K. Yu. Idiot's Bayes - not so stupid after all? *International Statistical Review*, 69:385 – 398, 2001.

[5] R. Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996.

[6] R. Kohavi, B. Becker, and D. Sommerfield. Improving simple Bayes. Technical report, Data Mining and Visualization Group, Silicon Graphics Inc, California, 1997.

[7] I. Kononenko. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7:317 – 337, 1993.

[8] L. I. Kuncheva. On the optimality of Naïve Bayes with dependent binary features. *Pattern Recognition Letters*, 27:830–837, 2006.

[9] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 399 – 406, 1992.

[10] P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 399–406, 1994.

[11] B. D. Ripley. *Pattern Recognition and Neural Networks*. University Press, Cambridge, 1996.

[12] I. Rish. An empirical study of the naive Bayes classifier. In *Proceedings of the International Joint Conference on Artificial Intelligence, Workshop on "Empirical Methods in A"*, 2001.

[13] I. Rish, J. Hellerstein, and J. Thathachar. An analysis of data characteristics that affect Naive Bayes performance. Technical Report RC21993, IBM TJ Watson Research Center, 2001.

[14] G. I. Webb, J. Boughton, and Z. Wang. Not so naive Bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5–24, 2005.

[15] G.U. Yule and M.G. Kendall. *An introduction of the Theory of Statistics*. Griffin Co. Ltd., 1940.

[16] H. Zhang. The optimality of Naive Bayes. In *Proceedings of the 17th International FLAIRS conference*, Florida, USA, 2004.

[17] H. Zhang and C. X. Ling. A fundamental issue of Naive Bayes. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 591 – 595, 2003.