


When is the Naive Bayes approximation not so naive?

Christopher R. Stephens^{1,2}  · Hugo Flores Huerta^{1,3} · Ana Ruíz Linares^{3,4}

Received: 30 September 2015 / Accepted: 30 June 2017 / Published online: 21 July 2017
© The Author(s) 2017

Abstract The Naive Bayes approximation (NBA) and associated classifier are widely used and offer robust performance across a large spectrum of problem domains. As it depends on a very strong assumption—independence among features—this has been somewhat puzzling. Various hypotheses have been put forward to explain its success and many generalizations have been proposed. In this paper we propose a set of “local” error measures—associated with the likelihood functions for subsets of attributes and for each class—and show explicitly how these local errors combine to give a “global” error associated to the full attribute set. By so doing we formulate a framework within which the phenomenon of error cancelation, or augmentation, can be quantified and its impact on classifier performance estimated and predicted a priori. These diagnostics allow us to develop a deeper and more quantitative understanding of why the NBA is so robust and under what circumstances one expects it to break down. We show how these diagnostics can be used to select which features to combine and use them in a simple generalization of the NBA, applying the resulting classifier to a set of real world data sets.

Keywords Classification · Naive Bayes approximation · Generalized Bayes approximation · Performance prediction · Error analysis

Editor: Johannes Fürnkranz.

✉ Christopher R. Stephens
stephens@nucleares.unam.mx

¹ C3 Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Circuito Exterior, A. Postal 70-543, 04510 Mexico, D.F., Mexico

² Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Circuito Exterior, A. Postal 70-543, 04510 Mexico, D.F., Mexico

³ IIMAS, Universidad Nacional Autónoma de México, Circuito Exterior, A. Postal 70-543, 04510 Mexico, D.F., Mexico

⁴ Instituto Tecnológico de Minatitlán, Minatitlán, Ver., Mexico

1 Introduction

Although superseded by more sophisticated classifiers, the Naive Bayes classifier (NBC) is still widely used in a multitude of different areas of application [see, for example, Wang et al. (2007), Turhan and Bener (2009), Wei et al. (2011), Broos et al. (2011), Panda and Patra (2007), Farid et al. (2014), Bermejo et al. (2014), Ng et al. (2014), Mohamad et al. (2014), as a small, but representative sample]. It has been shown to be remarkably robust, both in its range of applicability and its performance. Although it is well known that it is a high-bias/low-variance classifier, and therefore might be expected to have relatively better performance on small data sets, the robustness of its performance across so many problems with widely differing characteristics remains somewhat of a puzzle, especially given its strong independence assumption on the features, and has led to several papers trying to understand and explain why it seems to be “unreasonably” successful.

There have also been many papers associated with different proposals for generalizing the NBC so as to circumvent its strong independence assumption. However, these generalizations are invariably more complicated to implement and much more resource intensive than the NBC. It is therefore important to develop diagnostics with which, for a given problem, one can gauge if and when the NBC will lead to significant error relative to a more sophisticated classifier that tries to account for potential feature correlations, while at the same time developing a deeper intuitive and theoretical understanding of how and why the NBC is so robust. In this paper, we analyze under what circumstances the NBA (Naive Bayes approximation) and associated NBC can be expected to be suboptimal and develop general diagnostics with which a problem can be examined a priori to ascertain if the NBA is adequate, or if a more sophisticated generalization is required. We then use those diagnostics to determine which features should be combined and from that construct a simple generalization of the NBC.

As far as explaining the robust performance of the NBC, Domingos and Pazzani (1996) have argued that it is largely due to its being applied to classification problems, where errors are counted with respect to whether the classification was correct (yes/no) for a given prediction, not whether the corresponding probability estimate was accurate. Further support for this type of hypothesis comes from the work of Frank et al. (2000), who showed that the performance of the NBA is substantially worse when applied to regression type problems. Pure classification accuracy, however, is only a single, global measure of classifier performance and there are others that can be more appropriate. For instance, it has been demonstrated in Ling et al. (2003) that the area under the ROC curve is a more discriminating measure than pure classification accuracy. Often of more interest are relative risk scores. This is especially the case in problems such as healthcare costs where a very small ($\sim 1\%$), but high risk group, can generate a large fraction of healthcare costs ($\sim 30\%$). In this circumstance it is very unlikely that a classifier is sufficiently accurate that it would place anyone in such a small group. Rather, what is of interest is the relative degree of risk from one patient to another (Stephens et al. 2005). This reasoning, that beyond pure classification the NBA can be more rigorously judged, is circumstantially supported by the fact that the NBA can produce poor probability estimates (Bennett 2000; Monti and Cooper 1999), though in Lowd and Domingos (2005) it was shown that NB models can be as effective as more general Bayesian networks for general probability estimation tasks.

However, even in the case of classification, as noted by Zhang (Zhang and Ling 2003; Zhang 2004), Domingos and Pazzani’s argument does not explain why it is not possible to have situations where the inaccurate probability estimates flip the classification. Zhang has proposed that it is not just the presence of dependencies between attributes that affects performance, it is how they distribute between different classes that plays a crucial role in the

performance of the NBC, arguing that the effect of dependencies can partially cancel between classes and, further, that dependencies can potentially cancel between different subsets of feature values. The question then is, if this is possible, under what conditions will it occur and can we quantify it and therefore predict a priori when the NBA might be inadequate?

In this paper, we will investigate the relationship between attribute dependence and error by analysing a set of model problems, providing tools to investigate the set of attribute dependencies and showing how they affect both probability estimates and classification accuracy. We will thereby provide statistical diagnostics that allow both insight and predictive capacity to estimate when the NBA will breakdown and how to improve it. As accounting for attribute correlation is a question of model bias not model variance, as in [Rish \(2001\)](#), we will consider first a set of “infinite sample” artificial probability distributions, chosen in order to be able to tune the degree of correlation between different attributes while ignoring finite sampling effects. Indeed, it is precisely the existence of sampling error in real world problems that is associated with the superior performance of the NBA in spite of attribute dependencies ([Friedman 1997](#)).

The second major question revolves around how to improve the NBA and NBC. There have been many generalizations ([Friedman et al. 1997](#); [Keogh and Pazzani 1999](#); [Kohavi 1996](#); [Kononenko 1991](#); [Langley 1993](#); [Langley and Sage 1994](#); [Pazzani 1996](#); [Sahami 1996](#); [Singh and Provan 1996](#); [Webb and Pazzani 1998](#); [Webb 2001](#); [Webb et al. 2005, 2012](#); [Xie et al. 2002](#); [Zheng et al. 1999](#); [Zheng and Webb 2000](#); [Liangxiao et al. 2009](#)) of the NBA. Some, such as Lazy Bayesian Rules ([Zheng and Webb 2000](#)), Super Parent TAN ([Keogh and Pazzani 1999](#)) and Hidden Naive Bayes ([Liangxiao et al. 2009](#)), have been shown to have very good performance, with significant improvements over the NBA but at substantial computational cost. A good overview of many of these algorithms can be found in [Zheng and Webb \(2000\)](#), [Liangxiao et al. \(2009\)](#). As it is not the primary purpose of this paper to introduce a new, competing algorithm, we will restrict ourselves to some general comments: All these generalizations seek to discover sets of attribute values that have dependencies such that they should either be combined together, or with the class variable. In general, they are such that the improvement associated with combining a set of features is judged a posteriori through the relative performance of the algorithm with or without that combination. As there are a combinatorially large number of possible attribute value combinations that might be considered however, the process of attribute selection can be intensive, so that, generally, studies have been restricted to considering only pairs of attributes with an exhaustive search of those combinations being performed.

The effectiveness of these generalizations of the NBA is generally judged by comparing the performance of the proposed algorithm against the NBA, and, potentially, a chosen set of other algorithms, on a set of canonical test problems, more often than not taken from the UCI repository. The effectiveness of the new algorithm is then inferred globally across the set of considered problem instances. We know from the No-free lunch theorems ([Wolpert 1996](#); [Wolpert and Macready 1997](#)) that no algorithm is better than any other across all problem instances. The question is: can we infer a priori which algorithm will perform better on a given problem instance? This is especially important if performance enhancements are dominated by only a small number of instances. It also requires detailed insight as to how and why a given generalization outperforms the NBA on some data sets and not on others. Also, as comparatively complicated, “black box” type algorithms there is no transparent theoretical underpinning with which to understand their relative performance. A by product of the development of generalizations of the NBA has been the construction of diagnostics to determine the degree of attribute dependence and therefore detect which features to potentially combine. The most used diagnostic has been that of conditional mutual information ([Rish](#)

2001; Friedman et al. 1997; Zhang and Ling 2003). However, as Rish has pointed out, this measure does not correlate well with NBA performance, arguing that a better predictor of accuracy is the loss of information that features contain about the class when assuming the NB model. What is required is a measure of attribute dependence that relates directly to the appearance of a corresponding error in the NBA and, further, how these errors combine to yield an overall error for a given feature set or classifier.

The structure of the paper is as follows: In Sect. 2 we will compare and contrast the NBA with a particular class of generalizations—the semi-Naive Bayesian classifier (Kononenko 1991). We also introduce the metrics that we use for comparing the relative performance of the different approximations. In Sect. 3, we construct our basic error diagnostics— $\delta(\xi|C)$, the difference between the likelihood functions for a given class, C , and for a given feature combination, ξ , relative to the NBA to the likelihood function. In Sect. 4 we introduce the set of 12 two-feature probability distributions that we use to test our error diagnostics. In Sect. 5 using our 12 distributions we show that our error metrics are natural measures of when features should be combined, showing explicitly in the case of two features how the NBA can be valid even in the presence of strong attribute dependence due to cancelations between errors in the likelihoods of the different classes. In Sect. 6 we extend our analysis to three features to show how the GBA is sensitive to the particular factorization of the likelihoods and to the fact that different classes may have different factorizations. In Sect. 7 we show how our error diagnostics can determine the most appropriate factorizations in this three feature case and how they relate to and predict classifier performance. In Sect. 8 we extend our analysis to four, six and eight features, generalizing the previous analysis and, crucially, showing how errors can cancel between different sets of features for the same classifier. In Sect. 9 we apply the formalism to a set of UCI data sets, showing how our diagnostics can be used to determine which features to combine and constructing a simple generalization of the NBC and comparing its performance to the NBC. Finally, in Sect. 10 we summarize and draw our conclusions.

2 Comparing classifiers

2.1 Naive Bayes approximation

In trying to understand under what circumstances we might expect the NBA and the NBC to break down we will start with Bayes' theorem

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})} \quad (2.1)$$

for a class C and a vector of N features $\mathbf{X} = (X_1, X_2, \dots, X_N)$, where $P(C)$ is the prior probability, $P(\mathbf{X}|C)$ the likelihood function and $P(C|\mathbf{X})$ the posterior probability. Unfortunately, when \mathbf{X} is of high dimension, there are too many different probabilities, $\hat{P}(C|\mathbf{X})$, to estimate. Related to this is the fact that $N_{C\mathbf{X}}$, the number of elements in \mathbf{X} and C , is generally so small that statistical estimates, $\hat{P}(C|\mathbf{X})$ of $P(C|\mathbf{X})$ are unreliable due to large sampling errors.¹ Although $\hat{P}(\mathbf{X}|C)$ suffers from the same problem, if there is statistical independence of the X_i in the class C , then $P(\mathbf{X}|C) = \prod_{i=1}^N P(X_i|C)$, where $P(X_i|C)$ is the marginal conditional probability. Generally, this is not the case. However, one may

¹ Indeed, for a sufficiently large set of discriminatory features so that every combination is unique we will have $N_{C\mathbf{X}} = 0, 1$, with the vast majority of combinations being zero.

make the assumption that it is approximately true, taking $P_{NB}(\mathbf{X}|C) = \prod_{i=1}^N P(X_i|C)$ to find

$$P_{NB}(C|\mathbf{X}) = \frac{\prod_{i=1}^N P(X_i|C)P(C)}{\left(\prod_{i=1}^N P(X_i|C)P(C) + P(\mathbf{X}|\bar{C})P(\bar{C})\right)} \tag{2.2}$$

where \bar{C} is the complement of C . Of course, if we were to calculate $P(C|\mathbf{X})$ using (2.2) we would have to also estimate $P(\mathbf{X}|\bar{C})$, which would present the same problems as estimating $P(\mathbf{X}|C)$. The same naive approximation can be used in this case too, writing $P(\mathbf{X}|\bar{C}) = \prod_{i=1}^N P(X_i|\bar{C})$ to find

$$P_{NB}(\mathbf{X}) = \prod_{i=1}^N (P(X_i|C)P(C) + P(X_i|\bar{C})P(\bar{C})) \tag{2.3}$$

Rather than constructing $P(C|\mathbf{X})$ directly, from (2.1), usually a score function, $S(\mathbf{X})$, that is a monotonic function of $P(C|\mathbf{X})$ itself, is constructed, by considering the odds ratio of the class C and another class, usually its complement, \bar{C}

$$S(\mathbf{X}) = \log \frac{P(C|\mathbf{X})}{P(\bar{C}|\mathbf{X})} = \log \frac{P(C)}{P(\bar{C})} + \log \frac{P(\mathbf{X}|C)}{P(\mathbf{X}|\bar{C})}$$

The NBA to this score function, $S_{NB}(\mathbf{X})$, is given by

$$S_{NB}(\mathbf{X}) = \log \frac{P(C)}{P(\bar{C})} + \sum_{i=1}^N \log \frac{P(X_i|C)}{P(X_i|\bar{C})} \tag{2.4}$$

As a simple sum this form of the approximation is transparent. Another advantage of this form is that it is not necessary to have in hand $P(\mathbf{X})$ as the idea of such a score function is just to discriminate between the classes C and \bar{C} . This is a different task from estimating the probability $P(C|\mathbf{X})$ directly. As a classifier, (2.4) is such that if $S_{NB}(\mathbf{X}) > 0$ then the instance defined by \mathbf{X} is assigned to the class C and if $S_{NB}(\mathbf{X}) < 0$ to the complement \bar{C} .

2.2 Generalized Naive Bayes approximation

The NBA and NBC are based on a maximal factorization of the likelihood function $P(\mathbf{X}|C)$. Generalizations of the NBA have been associated with introducing dependencies between the features and seeking an alternative factorization of the likelihood functions. In order to have a concrete theoretical framework in which to examine the validity of the NBA, the framework of Bayesian networks (Friedman et al. 1997) is particularly appropriate.² Some generalizations, such as Lazy Bayesian Rules (Zheng and Webb 2000) and Super Parent TAN (Keogh and Pazzani 1999), consider factorizations of the form $\prod_{i=1}^N P(X_i|C, \xi)$, where ξ represents some subset of attribute values. In the case of TAN or SP-TAN, ξ is restricted to be one other variable. Other generalizations, such as the semi-Naive Bayesian classifier (Kononenko 1991), consider factorizations of the form $\prod_{i=1}^m P(\xi_i|C)$. Both types of generalization consider attribute dependencies. Here, we will focus on the semi-Naive Bayesian classifier.

To define generalizations of the NBA and NBC associated with alternative factorizations we will first introduce a “schema” based notation for marginal probabilities familiar

² Factorization of probability distributions and their graphical representations also occur in several other related fields, such as Markov Random Fields (Kindermann and Snell 1980) where cliques represent groups of related variables.

from genetic algorithms (Holland 1975; Poli and Stephens 2014), where for any feature vector, $\mathbf{X} = \{X_1, X_2, \dots, X_N\}$, a marginal probability $P(X_{i_1} X_{i_2} \dots X_{i_m} | C)$ can be written $P(*^{i_1-1} X_{i_1} *^{i_2-1} X_{i_2} \dots X_{i_m} | C)$, where $*^n$ signifies $*$ repeated n times and an $*$ in the i th position means that X_i has been marginalized. The order of the schema is just the number of non-marginalized variables. The NBA uses only order-one schemata, i.e., order-one marginals. We can then denote an arbitrary m -feature value combination by a schema $\xi = (\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_m})$, of order $m < N$.

With this notation we can define the GBA, and a Generalized Bayes Classifier (GBC), via generalizations of Eqs. (2.2) and (2.4) that correspond to alternative factorizations of the likelihood functions. Unlike the NBA the GBA is not unique. For N features there are B_N partitions, where B_N is the Bell number. Worse, there is this number for every feature value set and there are $\mathcal{R} = \prod_{i=1}^n a(i)$ feature value combinations, where $a(i)$ is the cardinality of the i th feature.

Of course, as estimates of probabilities or as classifiers the question is: out of all the possible factorizations which one is optimal and how do we define optimal? For instance, for the case of three variables, out of the 4 possible factorizations, which gives the best approximation to $P(X_1 X_2 X_3 | C)$? Much of this paper will be concerned with the question of how to determine better factorizations using diagnostics. We will denote the GBA associated with a particular factorization by a set of N_ξ schema $\xi^{(i)} = \cup_{\alpha=1}^{N_\xi} \xi^\alpha$. Note that any factorization must correspond to a partition of the set of N feature values. So, for any given feature vector, $\mathbf{X} = (X_1, X_2, \dots, X_N)$, every X_i must be a member of one and only one schema $\xi^\alpha \in \xi^{(i)}$. Thus, we have the GBA for the likelihood functions

$$P_{GB}(\mathbf{X} | C) = P(\xi^{(i)} | C) = \prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha | C) \tag{2.5}$$

where $N_{\xi^{(i)}}^C$ is the number of independent marginals used in the GBA for the likelihood function for the class C , and we take C to abstractly represent either the class C of interest or its complement \bar{C} . In the NBA, $N_{\xi^{(i)}}^C = N_{\xi^{(j)}}^{\bar{C}} = N$. Note that at this level of generality we make no a priori assumption that the optimal factorization of the likelihood function of C is the same as that of \bar{C} , although it is usually assumed that the factorization of the likelihood functions for C and \bar{C} are the same. We note again that our factorizations are a particular subset of all possible factorizations in that we will not consider factorizations where a particular attribute value may occur in more than one schema.

For the posterior probability we have

$$P_{GB}(C | \mathbf{X}) = P(C | \xi^{(i)}) = \frac{\prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha | C) P(C)}{\left(\prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha | C) P(C) + \prod_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} P(\xi^\alpha | \bar{C}) P(\bar{C}) \right)} \tag{2.6}$$

and finally, for the score function

$$S_{GB}(\mathbf{X}) = \ln \frac{P(C)}{P(\bar{C})} + \sum_{\alpha=1}^{N_{\xi^{(i)}}^C} S^C(\xi^\alpha) - \sum_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} S^{\bar{C}}(\xi^\alpha) \tag{2.7}$$

where we define

$$S^C(\xi^{(i)}) = \sum_{\alpha=1}^{N_{\xi^{(i)}}^C} \ln P(\xi^\alpha|C) \quad S^{\bar{C}}(\xi^{(j)}) = \sum_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} \ln P(\xi^\alpha|\bar{C}) \tag{2.8}$$

as the contributions to the overall score from the likelihood of C and \bar{C} respectively.

If we make the simplifying assumption that the optimal factorization of the likelihood functions for C and \bar{C} are the same then (2.7) simplifies to

$$S_{GB}(\mathbf{X}) = \ln \frac{P(C)}{P(\bar{C})} + \sum_{\alpha=1}^{N_{\xi}} S(\xi^\alpha) = \ln \frac{P(C)}{P(\bar{C})} + \sum_{\alpha=1}^{N_{\xi}} \ln \frac{P(\xi^\alpha|C)}{P(\xi^\alpha|\bar{C})} \tag{2.9}$$

which is a natural generalization of the score function in the NBA. The GBC is then such that a feature vector \mathbf{X} belongs to the class C if $S_{GB}(\mathbf{X}) > 0$ and to \bar{C} if $S_{GB}(\mathbf{X}) < 0$. Thus, the NBA and GBA will lead to the same classification if and only if $sign(S_{GB}(\mathbf{X})) = sign(S_{NB}(\mathbf{X}))$, $\forall \mathbf{X}$.

2.3 The difference

Any difference between the NBA and the GBA is due to correlations between the features X_i in the likelihoods for C and \bar{C} . In terms of model bias, the most appropriate factorization of the likelihood functions should be that which most respects the existence of such correlations. For a given factorization we can determine the differences between the GBA and the NBA from Eqs. (2.5), (2.6) and (2.7). For the likelihoods we have

$$\Delta_{P_{GB}}(\mathbf{X}|C) = P(\xi^{(i)}|C) - P_{NB}(\mathbf{X}|C) = \prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha|C) - \prod_{i=1}^N P(X_i|C) \tag{2.10}$$

where $\mathcal{C} = C$ or \bar{C} and for which we do not assume the same factorization. An important property of the (2.10) is that they satisfy the equations

$$\sum_{\mathbf{X}} \Delta_{P_{GB}}(\mathbf{X}|C) = \sum_{\mathbf{X}} \Delta_{P_{GB}}(\mathbf{X}|\bar{C}) = 0 \tag{2.11}$$

where the sum is over all feature vectors $\mathbf{X} = (X_1, X_2, \dots, X_N) = (\xi_1, \xi_2, \dots, \xi_{N_{\xi^{(i)}}^C})$, the decompositions over individual features or over schemata simply corresponding to two distinct partitions. The result (2.11) is a consequence of the conservation of probability as $\Delta_{P_{GB}}$ is composed of the difference between two probabilities, each of which is normalized. Thus, the differences between the GBA and NBA cannot all be of the same sign. As we will see, this plays an important role in understanding under what circumstances the NBA is a good one.

For the posterior probability we have

$$\Delta_{P_{GB}}(C|\mathbf{X}) = \frac{\prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha|C)P(C)}{\left(\prod_{\alpha=1}^{N_{\xi^{(i)}}^C} P(\xi^\alpha|C)P(C) + \prod_{\alpha=1}^{N_{\xi^{(j)}}^{\bar{C}}} P(\xi^\alpha|\bar{C})P(\bar{C}) \right)} - \frac{\prod_{i=1}^N P(X_i|C)P(C)}{\prod_{i=1}^N (P(X_i|C)P(C) + P(X_i|\bar{C})P(\bar{C}))} \tag{2.12}$$

and, finally, for the score function,

$$\begin{aligned} \Delta_{S_{GB}}(C|\mathbf{X}) &= \sum_{\alpha=1}^{N_{\xi^{(i)}^C}} \ln P(\xi^\alpha|C) - \sum_{\alpha=1}^{N_{\xi^{(j)}^{\bar{C}}}} \ln P(\xi^\alpha|\bar{C}) - \sum_{i=1}^N \ln \frac{P(X_i|C)}{P(X_i|\bar{C})} \\ &= \sum_{\alpha=1}^{N_{\xi^{(i)}^C}} \ln \frac{P(\xi^\alpha|C)}{P_{NB}(\xi^\alpha|C)} - \sum_{\alpha=1}^{N_{\xi^{(j)}^{\bar{C}}}} \ln \frac{P(\xi^\alpha|\bar{C})}{P_{NB}(\xi^\alpha|\bar{C})} \end{aligned} \tag{2.13}$$

where $P_{NB}(\xi^\alpha|C) = \prod_{i=1}^m P(\xi_i^\alpha|C)$, m being the number of features, ξ_i^α , in the feature schema ξ^α . In the case of identical factorizations for C and \bar{C}

$$\Delta_{S_{GB}}(C|\mathbf{X}) = \sum_{\alpha=1}^{N_\xi} \ln \frac{P(\xi^\alpha|C)}{P(\xi^\alpha|\bar{C})} - \sum_{i=1}^N \ln \frac{P(X_i|C)}{P(X_i|\bar{C})} \tag{2.14}$$

2.4 Performance metrics

In determining the relative merits of the GBA versus the NBA we will consider several performance metrics by which to judge them. In this paper, as emphasized, the idea is to understand the errors introduced by the intrinsic biases arising from the different approximations. Hence, in determining the impact of correlations on the validity of the NBA, and the improvement of the GBA, this can and, we would argue, should be considered first in an infinite sample setting, where sampling errors can be ignored. Hence, we will first consider various definite probability distributions in a setting with a small number of features. The advantage is that we then have in hand the exact probability distributions and therefore can measure errors in both the NBA and GBA relative to the exact distribution $P_e(C|\mathbf{X})$. Explicitly, we will consider distributions defined by the likelihood functions $P_e(\mathbf{X}|C)$ and $P_e(\mathbf{X}|\bar{C})$, where \mathbf{X} is a N -dimensional vector representing N binary features $X_i = 0, 1$. These likelihood functions will be specified in order to model different degrees of correlation between the features to better understand how the NBA and GBA perform as a function of these correlations.

As we have in hand the exact distributions, we will consider directly as a performance measure the error in the posterior probability distribution $\Delta_i(C|\mathbf{X}) = P_e(C|\mathbf{X}) - P_i(C|\mathbf{X})$, where i denotes the corresponding approximation—NBA or GBA. Secondly, we will consider the classification accuracy by considering the sensitivity S_i of each classifier, $S_i(\mathbf{X})$, which will correspond to the number, or fraction, of feature combinations classified correctly.

Thirdly, we will consider the relative ranking of the full set of feature combinations in terms of the relevant score function $\{S_i(1), S_i(2), \dots, S_i(2^n)\}$, where $S_i(1) \geq S_i(2) \geq \dots \geq S_i(2^n)$. We will then consider the distance

$$D_{ij} = \left(\sum_{m=1}^{2^n} (r_i(m) - r_j(m))^2 \right)^{1/2} \tag{2.15}$$

where $r_i(m)$ is the relative rank of the feature combination m of the classifier S_i and similarly $r_j(m)$ for the classifier $S_j(m)$. Here, if $i = e$, corresponding to the exact classifier, then Eq. (2.15) measures how faithful the ranking of the classifier is relative to the exact one. Finally, in the case of the real world data sets we will consider classification error and AUC as relevant metrics.

3 Coping with correlations

The differences between the NBA and GBA in Sect. 2.3 depend on the particular factorization chosen and this factorization is composed of components—schemata. In determining the relative merits of the NBA and GBA there are then two basic questions: Firstly, what criteria should be used to determine those features that should be considered together rather than independently? Secondly, once we have determined which features to combine, we must ask how the NBA should then be modified. We will first consider how to identify feature sets that should be considered together starting with the simple example of only two features.

3.1 Two features

Taking two features, X_1 and X_2 , treating them as independent potentially leads to errors in $P(\mathbf{X}|C)$, $P(\mathbf{X}|\bar{C})$ and $P(\mathbf{X})$, and therefore in $P(C|\mathbf{X})$ and $S(\mathbf{X})$. Denoting these errors as $\delta(X_1 X_2|C)$, $\delta(X_1 X_2|\bar{C})$ and $\delta(X_1 X_2)$ we have

$$\delta(X_1 X_2|C) = P(X_1 X_2|C) - P_{NB}(X_1 X_2|C) \tag{3.1}$$

$$\delta(X_1 X_2) = P(X_1 X_2) - P_{NB}(X_1 X_2) \tag{3.2}$$

where $\mathcal{C} = C$, or \bar{C} , that satisfy

$$\sum_{X_1 X_2} \delta(X_1 X_2|C) = \sum_{X_1 X_2} \delta(X_1 X_2|\bar{C}) = \sum_{X_1 X_2} \delta(X_1 X_2) = 0 \tag{3.3}$$

which is a consequence of the conservation of probability, as $\sum_{X_1 X_2} P(X_1 X_2|C) = \sum_{X_1 X_2} P_{NB}(X_1 X_2|C) = 1$. For the simple case of binary features we have $\delta(X_1 X_2|C) = \delta(\bar{X}_1 \bar{X}_2|C) = -\delta(X_1 \bar{X}_2|C) = -\delta(\bar{X}_1 X_2|C)$, where \bar{X}_i is the bit complement of X_i , implying that there are only two independent errors and that they have the same magnitude and opposite sign. We can also normalize any of these error terms, $\delta \rightarrow \delta'$; e.g. by dividing by the NBA.

These errors imply a corresponding error in the posterior probability

$$\begin{aligned} \delta(C|X_1 X_2) &= \left(\frac{P(X_1 X_2|C)P(C)}{P(X_1 X_2)} - \frac{P_{NB}(X_1 X_2|C)P(C)}{P_{NB}(X_1 X_2)} \right) \\ &= \frac{\delta(X_1 X_2|C)P_{NB}(\bar{C}|X_1 X_2)P(C) - \delta(X_1 X_2|\bar{C})P_{NB}(C|X_1 X_2)P(\bar{C})}{P(X_1 X_2)} \end{aligned} \tag{3.4}$$

where, in passing to Eq. (3.4), we have used the definitions of the errors (3.1) and (3.2) and (2.3). The quantities $\delta(X_1 X_2|C)$ and $\delta(X_1 X_2|\bar{C})$ offer a complete description of the errors of the NBA for the case of two variables. Interestingly, we can see that, as $\delta(C|X_1 X_2)$ involves the difference of the errors in the two likelihood functions, it is possible to have large errors in these without this necessarily leading to a significant error in the posterior probability itself. We can also see that, all else being equal, the error will be greater when the error in the likelihoods for the class and its complement are of opposite sign. In other words, that the variables are positively correlated in C/\bar{C} and negatively correlated in \bar{C}/C .

With regard to the score function (2.4) there are two potential sources of error—in $P(X_1 X_2|C)$ and in $P(X_1 X_2|\bar{C})$. The exact score is

$$\begin{aligned}
 S(X_1 X_2) &= S^C(X_1 X_2) - S^{\bar{C}}(X_1 X_2) \\
 &= \ln \frac{P(C)}{P(\bar{C})} + \ln \frac{P(X_1 X_2|C)}{P(X_1 X_2|\bar{C})}
 \end{aligned}
 \tag{3.5}$$

Given that in Sect. 2.2 we indicated that the optimal factorizations of the likelihood functions for C and \bar{C} may be distinct, it is convenient to introduce separate measures for the errors in the corresponding score functions

$$\delta_s(X_1 X_2|C) = \ln \left(1 + \frac{\delta(X_1 X_2|C)}{P_{NB}(X_1 X_2|C)} \right)
 \tag{3.6}$$

where, again, $C = C$ or \bar{C} . A consequence of (3.3) is that $\delta_s(X_1 X_2|C)$ cannot have the same sign for all $X_1 X_2$ thereby providing the basis by which deviations from the NBA can cancel between different feature combinations.

Comparing with Eq. (2.4) the difference between the GBA and NBA that accrues from correlation between X_1 and X_2 in C or \bar{C} is given by

$$\delta_s(C|X_1 X_2) = \ln \left(\frac{1 + \frac{\delta(X_1 X_2|C)}{P_{NB}(X_1 X_2|C)}}{1 + \frac{\delta(X_1 X_2|\bar{C})}{P_{NB}(X_1 X_2|\bar{C})}} \right) = S(X_1 X_2) - S(X_1) - S(X_2)
 \tag{3.7}$$

3.2 General case: more than two features

For the case of two features there is only one possible factorization. For more than two features, however, there are two related problems to be confronted: how many feature combinations (schemata) appear in a given factorization and which features appear in a given feature combination (schema)?

In terms of the error between the NBA for a given feature combination—schema ξ —the two-feature analysis generalizes quiet readily. Using our schema notation, errors (3.1) and (3.2) have simple generalizations

$$\delta(\xi|C) = P(\xi|C) - P_{NB}(\xi|C) = P(\xi|C) - \prod_{i=1}^m P(\xi_i|C)
 \tag{3.8}$$

$$\delta(\xi) = P(\xi) - P_{NB}(\xi) = \delta(\xi|C)P(C) + \delta(\xi|\bar{C})P(\bar{C})
 \tag{3.9}$$

where m is the number of features in the schema ξ . As with Eq. (3.3) we have

$$\sum_{\xi} \delta(\xi|C) = \sum_{\xi} \delta(\xi|\bar{C}) = \sum_{\xi} \delta(\xi) = 0
 \tag{3.10}$$

For the error in the posterior probability the generalization is

$$\delta(C|\xi) = \left(\frac{P(\xi|C)P(C)}{P(\xi)} - \frac{P_{NB}(\xi|C)P(C)}{P_{NB}(\xi)} \right)
 \tag{3.11}$$

Finally, for the score function, in distinction to the two feature case, where the factorizations of $P(X_1 X_2|C)$ and $P(X_1 X_2|\bar{C})$ are, by definition, the same, here it is appropriate to consider separately the errors in the contributions to the score from the likelihood functions for C and \bar{C} . From Eq. (2.8) we can then define

$$\delta_s(\xi|C) = \ln \left(1 + \frac{\delta(\xi|C)}{P_{NB}(\xi|C)} \right)
 \tag{3.12}$$

Assuming the same schema appears in the factorizations of $P(\mathbf{X}|C)$ the error is

$$\delta_s(C|\xi) = \delta_s(\xi|C) - \delta_s(\xi|\bar{C}) = S(\xi) - \sum_{i=1}^m S(\xi_i) \tag{3.13}$$

As with the two feature case, the constraints (3.10) imply that not all the errors $\delta_s(\xi|C)$ and $\delta_s(\xi|\bar{C})$ can be of the same sign.

3.3 General case: more than two schemata

The preceding error functions are useful for determining the impact of attribute dependence in a subset, $\xi \subset \mathbf{X}$, of feature values. However, when there are multiple sets, it is not clear how errors may combine to influence the total difference between the NBA and GBA as illustrated by Eqs. (2.10), (2.12) and (2.13). As might have been anticipated, the difference between the likelihood functions and the posterior probabilities for the GBA and NBA do not seem to be simple functions of the errors $\delta(\xi|C)$. However, for the score function we may write

$$\begin{aligned} \Delta_{S_{GB}}(C|\mathbf{X}) &= \sum_{\alpha=1}^{N_{\xi}^C} \ln P(\xi^{\alpha}|C) - \sum_{\alpha=1}^{N_{\xi}^{\bar{C}}} \ln P(\xi^{\alpha}|\bar{C}) - \sum_{i=1}^N \ln \frac{P(X_i|C)}{P(X_i|\bar{C})} \\ &= \sum_{\alpha=1}^{N_{\xi}^C} \delta_s(\xi^{\alpha}|C) - \sum_{\alpha=1}^{N_{\xi}^{\bar{C}}} \delta_s(\xi^{\alpha}|\bar{C}) \end{aligned} \tag{3.14}$$

In (3.14) we can see how the constraints (3.10) may play a role in error cancellation. In the error associated with C , as for each schema, ξ^{α} , there is at least one particular feature combination, $\xi_{i_1 i_2 \dots i_m}^{\alpha}$, with an error $\delta_s(\xi^{\alpha}|C)$ that has a different sign to the others, then some degree of error cancellation is inevitable between different schemata when considering different feature combinations for those schemata.

When the factorizations are the same, Eq. (3.14) simplifies even further,

$$\Delta_{S_{GB}}(C|\mathbf{X}) = \sum_{\alpha=1}^{N_{\xi}} \delta_s(C|\xi^{\alpha}) \tag{3.15}$$

Equations (3.14) and (2.13) hold an important lesson: Just as for a single schema one can have cancelations in likelihood errors between C and \bar{C} , i.e., intra-schema cancelations, so, one can have cancelations between different schemata, i.e., inter-schemata cancelations. What is more, the likelihood errors for the components of the factorizations of C and \bar{C} are what determine the overall error of the NBA.

4 What difference does it make?

The impact of correlations on the validity of the NBA should be understood from the point of view of model bias. As argued, this should be considered first in an infinite sample setting. In Appendix A we introduce a set of 12 different two-feature probability distributions with binary feature values and also the parity function. The distributions are uniquely characterized by specifying values for the likelihoods $P(X_1 X_2|C)$ and $P(X_1 X_2|\bar{C})$ for $X_i = 0, 1$ for each distribution.

These probability distributions have been constructed to exhibit a diversity of different correlation structures that can occur. For instance, distributions 1–3 all use the same set of probabilities for the likelihoods for C and \bar{C} and differ only in how the likelihoods for \bar{C} are assigned to the different feature value combinations. All three distributions exhibit strong correlations between the features. How these correlations affect the validity of the NBA, however, is quite distinct. Distribution 4 is chosen as it exhibits only weak correlations between the features in both likelihoods. Distributions 5–7 show moderate correlations but with different characteristics. For example, distributions 6 and 7 have correlations in the likelihoods that are all of equal magnitude, differing only in the sign between C and \bar{C} . Distribution 8 shows strong correlations in the likelihood for C , but weak correlations for \bar{C} . Distribution 9 is the inverse of distribution 7, with C and \bar{C} interchanged. Distribution 10 is the analog to distribution 9 as distribution 2 is to distribution 1, i.e., the likelihoods of C are the same but the likelihoods for \bar{C} have been permuted. Finally, distributions 11 and 12 are two more that exhibit strong correlations in all likelihood functions but differ in how the correlations are distributed among the features. In all these cases the class probability was taken to be $P(C) = 0.6$. We add in as distribution 0 the well known parity function, where $P(C|X_1X_2) = 0, 1$ according to if $X_1 + X_2$ is even or odd. This serves as a test case where correlations are strongest.

We will consider the errors at two levels: Firstly, in estimating the $P(C|X_1X_2)$; and, secondly, as a classifier, where we consider the difference between the exact classifier $S_e(\mathbf{X}) = \ln(P(C|X_1X_2)/P(\bar{C}|X_1X_2)) + \ln(P(C)/P(\bar{C}))$ and the NBC, Eq. (2.4). We will consider the percentage errors between the exact classifiers and posterior probability and their NBA counterparts, as well as the distance function (2.15). In Table 9 we see the results of the calculation for the exact posterior probability for the twelve probability distributions and the parity function, the NBA to the posterior probability and the percentage difference to the exact expression. There are four feature combinations, $X_1 = 0, 1, X_2 = 0, 1$ for each class $C = 1, \bar{C} = 0$ and vice versa. Also in that table are the exact scores for each classifier and its NBA and the percentage difference between them.

So what can we glean from these results? Clearly, the performance of the NBA is very mixed over the different probability distributions, with mean absolute errors relative to the exact posterior probabilities over a given distribution of between 10 and 100% and, in the case of score differences, errors of between 10 and 5000%, and infinite in the case of the parity function, where the NBA gives zero score. In the case of estimating posterior probabilities the distributions where the NB error is greatest are 11, 3, 1, 8 and 10 and the least in 4, 7, 12, 9 and 2 and, of course, the parity function. What are distinguishing factors of better/worse performance? In Table 9 we also see the errors in the likelihood functions, (3.1), and the errors in the score functions, (3.6). Note that errors in the likelihood functions themselves are not necessarily good indicators of errors in the posterior probabilities or classification errors. Rather, what is of most importance is the relative sign of the errors for C and \bar{C} . The distributions with the largest errors in the posterior probabilities or score—0, 11, 3, 1, 8 and 10—all have sign differences between $\delta(X_1X_2|C)$ and $\delta(X_1X_2|\bar{C})$ for every value of X_1 and X_2 . On the other hand, four of the five distributions with least error—2, 7, 9 and 12—do not have any sign differences between $\delta(X_1X_2|C)$ and $\delta(X_1X_2|\bar{C})$ for any value of X_1 and X_2 . The low error distribution, 4, has sign differences, but in this case the magnitude of the errors is very low $\approx 10^{-2}$.

5 Diagnostics for when the NBA is valid

When considering the validity of the NBA we have to ask: The NBA of what? and, also, With respect to what benchmark? To answer the NBA of what—two important quantities

to calculate are posterior probabilities and classifier scores. For benchmarking we cannot benchmark to the exact answer in real world problems, so we must benchmark to another algorithm. Here, the benchmarking will be to the GBA. Generically, the validation of an algorithm, comes from a performance metric on a set of test problems, i.e., the choice of algorithm is determined *a posteriori*, after testing the algorithm on the set of problems. As emphasised, an important goal here is to infer *a priori* whether a GBA type algorithm will be better than the NBA. We have seen that the differences between the two approximations stem from two different factorisations of the likelihood functions for the class of interest and its complement. We have also seen that we can infer potential errors at the level of the individual factors—schemata—which are combinations of features. This provides a huge simplification as it allows one to infer which approximation will be better on a given problem before applying the full algorithm.

5.1 Diagnostics for when to combine features

So, what is a good diagnostic to determine the relative accuracy of the approximations? In the case of two features we have deduced that significant errors in the likelihood functions are a *necessary*, but not *sufficient*, condition in order to have significant errors in the estimation of posterior probabilities or in classification performance. With this in mind, we propose (3.1), or the associated functions, (3.6), as direct measures of when it is necessary to combine the features $X_1 X_2$ in the context of a given likelihood. Specifically, we introduce

$$\Delta_C(X_1 X_2) = \delta(X_1 X_2|C) / P_{NB}(X_1 X_2|C) \tag{5.1}$$

$$\Delta_s^C(X_1 X_2) = \delta_s(X_1 X_2|C) / S_{NB}^C(X_1 X_2) \tag{5.2}$$

or their unnormalized equivalents, as measures of when to potentially combine variables in the likelihoods of C and \bar{C} . We can then ask to what degree significant errors in these functions lead to significant errors in the distinct performance measures. However, as the difference between the GBA and NBA depends on the relative signs of the errors in the likelihood functions we propose as further diagnostics

$$\Delta(X_1 X_2) = \left(\frac{\delta(X_1 X_2|C)}{P_{NB}(X_1 X_2|C)} - \frac{\delta(X_1 X_2|\bar{C})}{P_{NB}(X_1 X_2|\bar{C})} \right) \tag{5.3}$$

$$\Delta_s(X_1 X_2) = \left(\frac{\delta_s(X_1 X_2|C) - \delta_s(X_1 X_2|\bar{C})}{S_{NB}(X_1 X_2|C)} \right) \tag{5.4}$$

Thus, if $\Delta(X_1 X_2)$ or $\Delta_s(X_1 X_2)$ is large, then the NBA will be expected to give relatively poor performance for predicting, say, $P(C|X_1 X_2)$. In Fig. 1 we see a graph of the error in $P(C|X_1 X_2)$ as a function of $\Delta(X_1 X_2)$ for the twelve probability distributions introduced in Sect. 4. We have differentiated in the graph the distinct distributions, wherein, from the analysis of Sect. 4, we can distinguish the 5 distributions with greatest error, the 5 with least error and the two intermediate—“neutral”—distributions. We see that significant errors in the likelihood functions for C or \bar{C} separately are not sufficient to predict errors in the corresponding posterior distributions. Of the five distributions with smallest errors—2, 4, 7, 9 and 12—only one, 4, has errors in the likelihoods which are small in magnitude. The others are associated with distributions where the errors in the individual likelihoods are large, but of the same sign for C and \bar{C} , thus leading to a partial cancelation between the two in their contributions to the posterior distributions. In distinction, the distributions with the greatest error in $P(C|X_1 X_2)$ have large likelihood errors, but of opposite sign between C and \bar{C} , thus leading to a reinforcement of the individual errors. As can be seen, the correlation is

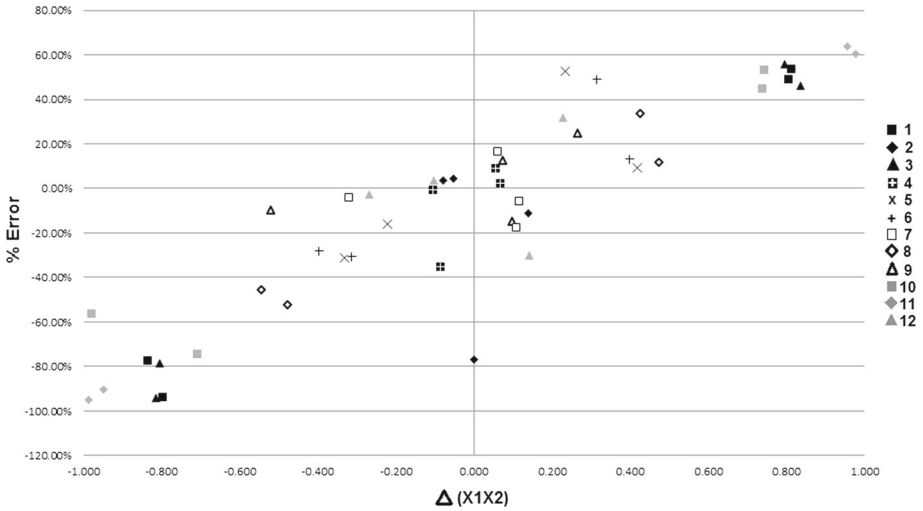


Fig. 1 Graph of % error in NBA of the posterior probability $P(C|X_1X_2)$ as a function of $\Delta(X_1X_2)$ for each feature combination of the 12 probability distributions of Appendix A

impressive—coefficient of correlation 0.90. The analogous coefficients of correlation for Δ_C and $\Delta_{\bar{C}}$ are 0.77 and -0.59 respectively. Thus, we see that error cancellation between the different likelihoods makes $\Delta(X_1X_2)$ a better indicator of performance difference between the NBA and GBA for calculating posterior probabilities than Δ_C and $\Delta_{\bar{C}}$, though the latter also show substantial correlation.

In the case of more than two variables we propose equations (3.8) as diagnostics for when the error in the likelihood function is sufficient to warrant using the GNB approximation for that schema. As with the case of two variables however, significant errors in the likelihood functions are not a sufficient condition for significant errors in the posterior distribution or in classification accuracy. Hence, one may be tempted to take analogs of (5.3) and (5.4), with X_1X_2 replaced by an arbitrary schema ξ , as a natural diagnostic for when features should be combined into a schema of more than two variables. Going beyond the case of two features however, we must confront the possibility that the optimal factorization of the likelihoods for C and \bar{C} may be different. In this case, we propose as diagnostics

$$\Delta_C(\xi) = \frac{\delta(\xi|C)}{P_{NB}(\xi|C)} \tag{5.5}$$

$$\Delta_{\bar{C}}^C(\xi) = \frac{\delta_s(\xi|C)}{S_{NB}^C(\xi)} \tag{5.6}$$

or their unnormalized equivalents. So, we have a conundrum: we can identify measures for when the likelihood functions, or score contributions, for a given schema should not be factorized, but this is not sufficient to guarantee significant errors in the posterior probabilities, or in classification performance. On the other hand, we can identify measures, Eqs. (5.3), (5.4) that directly speak to errors in these latter quantities, but that are associated with a symmetric, not necessarily optimal, factorization of the likelihoods $P(\mathbf{X}|C)$ and $P(\mathbf{X}|\bar{C})$.

When the likelihood factorizations are not symmetric, the above discussion indicates that the validity of the NBA is a “global” as opposed to a “local” question; i.e., for a non-symmetric factorization, the validity should be adjudged *after* calculating the contribution and

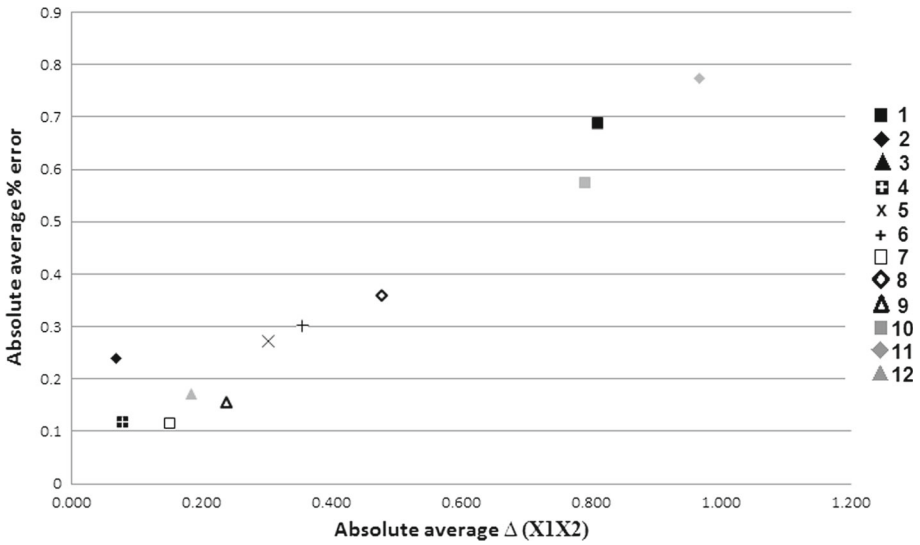


Fig. 2 Graph of average absolute error in $P(C|X_1X_2)$ as a function of average value of D_2 for the 12 probability distributions of Appendix A

corresponding errors of all factors together. On the other hand, for a symmetric factorization, the evaluation of errors can be broken down into a set of separate components, and this is how the generalizations of the NBA discussed in the introduction function. We will consider these two cases in more depth in the next Section, beginning with the symmetric case.

First, however, rather than an analysis for every feature variable combination individually, as in real world problems there may be many of them, we are likely to be more interested in the validity of the NBA when averaged over them, and, as $\Delta(X_1X_2)$ can change sign from one set of feature values to another, as a measure of the validity of the NBA over the whole set of possible feature values we take

$$D_2 = \sum_{i,j} |\Delta(X_iX_j)| \quad \text{or} \quad D_{2s} = \sum_{i,j} |\Delta_s(X_iX_j)| \tag{5.7}$$

In Fig. 2 we see a graph of the average absolute error in the posterior probability versus the average value of D_2 over the feature value combinations 00, 01, 10, 11, and the average value of $\delta(X_1X_2|C)/P_{NB}(X_1X_2|C)$ over those combinations for each of the 12 probability distributions seen in Appendix A. We clearly see the high degree of correlation for Δ , with correlation coefficient of 0.97 (the corresponding correlation coefficient using δ is 0.43), thus demonstrating once again, at least at this simple level, its value as a predictor of when the NBA will break down. This also indicates the importance of considering \bar{C} in understanding the effect of the NBA on the calculation of posterior probabilities.

Besides the impact of the NBA on the calculation of posterior probabilities there is the question of how it impacts classification accuracy. We could use Eq. (5.3) as a diagnostic. However, we can also use (5.4), which shows a high degree of correlation with (5.3). In the case of only two features, if we take the exact score as our target then Eq. (5.4) is somewhat tautological. The correlation between a pure classification performance measure, such as sensitivity, and a diagnostic such as (5.4) is weaker than for that of posterior probability. Although the relation shows a clear tendency there is also a high degree of dispersion. The

reason for the high variance is that a misclassification is not dependent on the difference in magnitude between $S(X_1 X_2)$ and $S_{NB}(X_1 X_2)$ for a given $X_1 X_2$ but, rather, only on if there is a difference in sign between them. To account for this, we can introduce the diagnostic

$$S(X_1 X_2) = \text{sign}(S(X_1 X_2)) - \text{sign}(S_{NB}(X_1 X_2)) \quad (5.8)$$

If $S(X_1 X_2) = 0$ then the GBC and NBC will place the instance $X_1 X_2$ in the same class. If $S(X_1 X_2) \neq 0$ on the other hand the GBC and NBC will place the instance in different classes. From this we can clearly intuit the robustness of the NBC given that it is only when the exact classifier and the NBC are in disagreement that there can be differences between them. This will preferentially occur near a score threshold, S^* , generically $S^* = 0$, that marks the cutoff between one class and another. In other words, S and S_{NB} can be substantially different and still have agreement over the assigned class. This is the logic of the argument of Domingos and Pazzani (1996) about the robustness of the NBC.

A complementary diagnostic is the ratio of the variance in the exact score to the variance in the NB score, the average value of this ratio being 429 for the 5 distributions with highest error and 29 for those with lowest error. This reflects the fact that there is much more dispersion in the exact posterior probabilities than in their NB approximations and that this dispersion is particularly associated with correlation between the features as opposed to the contributions of the individual features themselves. For distribution 4 the NBA performs very well, and we can clearly see why. The variance in the NB scores is 5.94, far larger than the other distributions, and very similar to the exact variance. This variance is the raw material on which the validity of the NBA rests. The larger this variance, and the more similar it is to the exact variance, the less likely it is that sampling errors will lead to an erroneous ranking of the NBC relative to the exact one.

6 Factorization dependence of the performance of the GBA versus the NBA

We will now concentrate on the effect of the GNB and NB approximations in terms of the corresponding classifiers $S_{GNB}(\mathbf{X})$ and $S_{NB}(\mathbf{X})$ rather than on the calculation of posterior probabilities. The reason for this is two-fold, first of all the vast majority of work on the NBA and the GBA is in terms of classification, and, secondly, the analysis of errors is much simpler due to their purely additive nature in the score functions. Analyzing Eq. (3.14), two fundamental observations are pertinent: Firstly, the error in the score function for a given schema ξ may be small, even though the errors in the constituent likelihood functions is large, due to a cancelation between the differences $\delta_s(C|\xi)$ and $\delta_s(\bar{C}|\xi)$. Secondly, the overall difference in the GBC and the NBC may be small due to cancelations between the differences $\delta_s(C|\xi)$ and/or $\delta_s(\bar{C}|\xi)$ with those, $\delta_s(\bar{C}|\xi')$ and $\delta_s(C|\xi')$, of other schemata ξ' .

To develop more intuition for this, we can turn to the philosophy of the two-variable case, where we posited specific distributions for the likelihood functions. Now, however, we are interested in more than two features. We will consider in this Section three-feature distributions $P(X_1 X_2 X_3 | C)$, before passing in later Sections to four-, six- and eight-feature distributions, all with analogous expressions for the likelihoods for \bar{C} . Of course, one can consider correlations in the exact likelihoods such that they do not possess any exact factorization at all. However, any real world problem will, at the very least, have approximate factorizations. What is more, in an algorithmic implementation of the GBA, where a vast space of possible factorizations could be searched, it is natural to concentrate on correlations

between pairs of features as these will tend to be the combinations that have the biggest sample size and therefore the most statistical significance. We will therefore use the probability distributions of Appendix A, creating distributions for more than two features by concatenating the two-feature distributions we have constructed. In this way we will not consider potential correlations between more than two features, although the following analysis generalizes very simply to more than two features. Moreover, all our qualitative observations about the validity of the NBA are independent of the number of correlated features.

We will proceed by considering different possible scenarios: first, we may consider that the correlations in the underlying probability distributions of the likelihoods for C and \bar{C} may be symmetric, i.e., appear in the same feature combinations in the two likelihoods, or asymmetric, i.e., that the correlated feature combinations in the likelihoods for C and \bar{C} are distinct; secondly, in applying the GBA we may choose a factorization of the likelihoods that is symmetric, i.e., the same for both, or asymmetric, i.e., distinct for both. We will observe, naturally, that the optimal factorization, in terms of our performance metrics, is that which best respect the correlations in the underlying probability distributions.

6.1 Symmetric correlations, symmetric factorizations: three features

We will begin with the case where correlations in the likelihoods for C and \bar{C} are associated with the same features and, moreover, the factorizations of the likelihoods of C and \bar{C} are symmetric in that the combined features in the GBA are the same for both likelihoods. In schema language, symmetric factorizations are such that the schema partitions, $\xi^{(i)}$, of both likelihoods are the same. We will fix the exact probability distribution to be a product

$$P_e(\mathbf{X}|C) = P(X_1X_2X_3|C) = P(X_1X_2|C)P(X_3|C) \tag{6.1}$$

where each $P(X_1X_2|C)$ can be chosen from an independent distribution; for example, from our set of twelve distributions. In this case, the likelihoods are a product of one order-two schema and one order-one schema. The exact score function/classifier is

$$S_e(X_1X_2X_3) = \ln \frac{P(X_1X_2X_3|C)}{P(X_1X_2X_3|\bar{C})} = \ln \frac{P(X_1X_2|C)}{P(X_1X_2|\bar{C})} + \ln \frac{P(X_3|C)}{P(X_3|\bar{C})} \tag{6.2}$$

Unlike the NBA, for the GBA there are many possible distinct factorizations. For three binary features, we denote the NBA by the schema partition $\xi^{(0)} = (\xi_1 = X_1, \xi_2 = X_2, \xi_3 = X_3)$. The three possible two-schema partitions are: $\xi^{(1)} = (\xi_1 = X_1X_2, \xi_2 = X_3)$, $\xi^{(2)} = (\xi_1 = X_1X_3, \xi_2 = X_2)$ and $\xi^{(3)} = (\xi_1 = X_2X_3, \xi_2 = X_1)$. The only order-three schema $\xi^{(4)} = (\xi_1 = X_1X_2X_3)$ corresponds to the exact probability distribution itself. As, by construction, there are no dependencies between the schemata $\xi_1 = X_1X_2$ and $\xi_2 = X_3$, the GBA, if it is based on the two schemata $\xi_1 = X_1X_2$ and $\xi_2 = X_3$, should be exact. In other words,

$$\begin{aligned} P_e(X_1X_2X_3|C) &\equiv P(X_1X_2|C)P(X_3|C) \\ &= P(\xi^{(1)}|C) \equiv P_{GB}^e(X_1X_2X_3|C) \end{aligned} \tag{6.3}$$

where the superscript e on P_{GB}^e signifies that this particular factorization of the GBA is exact as it respects the correlation structure of the exact probability distribution. The factorizations of the GBA we will consider here are $(\xi^{(0)}, \xi^{(0)})$, $(\xi^{(1)}, \xi^{(1)})$, $(\xi^{(2)}, \xi^{(2)})$ and $(\xi^{(3)}, \xi^{(3)})$ which are all symmetric, in that the combined features are the same in the likelihoods of both C and \bar{C} .

In Table 1 we see the performance of these different factorizations relative to the exact classifier, and also to the NBA, $(\xi^{(0)}, \xi^{(0)})$, for the cases where the correlations in the under-

Table 1 Performance measures for symmetric GBA factorizations in three feature problem with symmetric correlations in the likelihood functions

F	SS		WW		WS		SW		S S'	
	S	D	S	D	S	D	S	D	S	D
00	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
11	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
22	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898
33	0.625	7.211	1.000	1.414	0.875	4.242	0.625	6.633	0.750	4.898

F is factorization, S is sensitivity and D is distance

lying probability distribution are strong, *s*, (we use distribution 1 of Appendix A) or weak, *w*, (where we use distribution 4 of Appendix A). As performance measures we consider the sensitivity of the different classifiers and the distance function, Eq. (2.15). We consider the cases *ss*, *ww*, *sw* and *ws* for the strength of correlation in the likelihoods for *C* and \bar{C} respectively. Thus, *ss* refers to strong correlation in both likelihoods (distribution 1 for the likelihoods of *C* and \bar{C}), whereas *ws* corresponds to a weak correlation in the likelihood for *C* (distribution 4) but a strong correlation in the likelihood for \bar{C} (distribution 1) and vice versa for *sw*. Finally, we will consider the correlation strength distribution *ss'*, which corresponds to distribution 2 for the likelihoods of both *C* and \bar{C} . The important distinction between distributions 1 and 2 is that, although both are associated with strong correlations in the likelihoods for *C* and \bar{C} , distribution 1 results in a reinforcement of errors between the likelihoods of *C* and \bar{C} , whereas distribution 2 is associated with an intra-schema cancellation.

For the case where there are strong correlations in either or both likelihood functions, we see that the GBA is better or equal to the NBA in every case, independently of the factorization. For the factorization $\xi^{(1)}$ for both *C* and \bar{C} it is strictly better. This is understandable, as in this case the factorization captures precisely the correlation structure of the underlying exact probability distributions. For the factorizations $\xi^{(2)}$ and $\xi^{(3)}$, the performance of the GBA is equivalent to that of the NBA, because there are no correlations between the features in the pairs X_1X_3 or X_2X_3 . For example, for $\xi^{(2)}$, which considers $P(\xi_1|C) = P(X_1X_3|C)$, we have $P(X_1X_3|C) = P(X_1|C)P(X_3|C)$, which is equivalent to the NBA. Note also that the performance of the NBA in the case of *ss'* is substantially better than for the distribution *ss*, showing how the quality of the NBA is improved when there is a cancellation effect among large errors in the likelihoods. We can also conclude that there is no performance cost relative to the NBA of choosing an inappropriate symmetric factorization, 22 or 33, but neither is there an improvement.

6.2 Asymmetric correlations, all factorizations: three features

We now consider the case where the correlations in the two likelihoods are asymmetric with

$$P_e(\mathbf{X}|C) = P(X_1X_2X_3|C) = P(X_1X_2|C)P(X_3|C) \tag{6.4}$$

$$P_e(\mathbf{X}|\bar{C}) = P(X_1X_2X_3|\bar{C}) = P(X_1X_3|\bar{C})P(X_2|\bar{C}) \tag{6.5}$$

with an exact score function/classifier

Table 2 Performance measures for all GBA factorizations in three-feature problem with asymmetric correlations in the likelihood functions, where F is factorization, S is sensitivity and D is distance and we added in the NBA for easy comparison

F	SS		WW		WS		SW		S S'	
	S	D	S	D	S	D	S	D	S	D
00	0.625	8.367	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
01	0.625	8.367	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
02	0.500	6.782	1.000	0.000	1.000	0.000	0.500	6.928	0.625	6.164
03	0.625	8.367	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
10	1.000	3.464	1.000	0.000	0.875	4.899	1.000	0.000	0.875	3.162
11	1.000	3.464	1.000	0.000	0.875	4.899	1.000	0.000	0.875	3.62
12	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
13	1.000	3.464	1.000	0.000	0.875	4.899	1.000	0.000	0.875	3.262
20	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
21	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
22	0.500	6.782	1.000	0.000	1.000	0.000	0.500	6.928	0.625	6.164
23	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
30	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
31	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348
32	0.500	6.782	1.000	0.000	1.000	0.000	0.500	6.928	0.625	6.164
33	0.625	8.366	1.000	0.000	0.875	4.899	0.500	6.928	0.750	7.348

$$\begin{aligned}
 S_e(X_1 X_2 X_3) &= \ln \frac{P(X_1 X_2 X_3 | C)}{P(X_1 X_2 X_3 | \bar{C})} \\
 &= \ln P(X_1 X_2 | C) + \ln P(X_3 | C) - \ln P(X_1 X_3 | \bar{C}) - \ln P(X_2 | \bar{C}) \quad (6.6)
 \end{aligned}$$

Once again, there are three possible factorizations: $\xi^{(1)} = (\xi_1 = X_1 X_2, \xi_2 = X_3)$, $\xi^{(2)} = (\xi_1 = X_1 X_3, \xi_2 = X_2)$ and $\xi^{(3)} = (\xi_1 = X_2 X_3, \xi_2 = X_1)$. However, distinct to the symmetric case, here the optimal GBA factorization is different for the two likelihoods. For $P(X_1 X_2 X_3 | C)$, the factorization $\xi^{(1)}$ will be optimal, while for $P(X_1 X_2 X_3 | \bar{C})$ the factorization $\xi^{(2)}$. As in the previous Section, we consider the cases *ss*, *ww*, *sw*, *ws* and *ss'* for the strength of correlation in the likelihoods for *C* and \bar{C} respectively. We use distribution 1 for *s* and distribution 4 for *w*. For *ss'* we use distribution 2. The difference now is that *s* in *ss* is associated with distribution 1 for $P(X_1 X_2 | C)$ and $P(X_1 X_3 | \bar{C})$, $P(X_1 X_2 | C)$ for *s* in *sw* and $P(X_1 X_3 | \bar{C})$ for *s* in *ws*. Similarly, for *w*.

In Table 2 we see the performance of the different factorizations relative to the exact classifier and the NBA (factorization 00) for the cases *ss*, *ww*, *sw*, *ws* and *ss'*. The first immediate observation is that the factorization 12, corresponding to $\xi^{(1)}$ for $P(X_1 X_2 X_3 | C)$ and $\xi^{(2)}$ for $P(X_1 X_2 X_3 | \bar{C})$, has optimal performance, with 100% classification accuracy and perfect ranking. In general, we see that the GBA for most factorizations is better than or equal to the NBA with respect to both metrics. For the distribution *ss* the GBA is strictly better or equal to the NBA for all factorizations and both performance measures. Moreover, it is strictly better in those cases, 1* and *2, where one of the combined feature pairs captures the underlying correlations in the probability distributions; 1* capturing the strong correlation in $P(X_1 X_2 | C)$ and *2 the underlying strong correlation in $P(X_1 X_3 | C)$. Interestingly, for the case *ww*, the NBA is optimal with respect to both sensitivity and distance. This is in

distinction to the case of symmetric correlations. In the mixed distributions, ws and sw , we see the same pattern as for the case of symmetric correlations: i.e., the GBA is strictly better for sw in the cases 10, 11 and 12 and 13, where the GBA captures the strong correlation in the likelihood for C ; for the distribution ws it is strictly better in the cases 02, 12, 22 and 32, where the GBA accounts for the strong correlation in the likelihood for \bar{C} . For the distribution ss' , interestingly, in no case is the GBA worse than the NBA in terms of the distance metric. For the classification metric it is worse for 02, 22 and 32. These are precisely the cases where features are combined in the strongly correlated likelihood for \bar{C} but not for the strongly correlated likelihood for C .

7 Error analysis: choosing the right factorization

In the above we showed how the relative performance of the GBA when compared to the NBA was sensitive to the factorization used for the GBA relative to the distribution of correlations inherent in the underlying probability distributions. However, to determine the relative efficacy of the GBA in the above we exhaustively considered every factorization. This is not practicable for real world problems with many features, hence, the importance of a priori diagnostics. We will consider the distribution of errors associated with all two-feature combinations using the error diagnostics, $\Delta_C(X_i X_j)$, $\Delta_{\bar{C}}(X_i X_j)$ and $\Delta(X_i X_j)$, and show how these indicate which features should be combined and, therefore, when and why the GBA should be used. We will work first in the context of the three-feature problem, considering first symmetric and then asymmetric correlations in the underlying probability distributions.

For symmetric correlations, an analysis of the errors for every pair of feature values yields several noteworthy observations. First, that $\Delta_C(X_i X_j) = \Delta_{\bar{C}}(X_i X_j) = \Delta(X_i X_j) = 0$ for $X_i X_j = X_1 X_3$ or $X_i X_j = X_2 X_3$, for all five correlation strength distribution types. In other words, our diagnostics can clearly identify combinations of features where there are no correlations, that can then be well approximated by the NBA. The only non-zero values are associated with the feature combination $X_1 X_2$, where their values depend on the degree of correlation in the underlying probability distributions. For the distribution ss , both $\Delta_C(X_1 X_2)$ and $\Delta_{\bar{C}}(X_1 X_2)$ are large—ranging from 64 to 96% of the corresponding probabilities in the NBA for the four different feature value combinations—while the error in $\Delta(X_1 X_2)$ is approximately double that, due to the reinforcement of the errors from C and \bar{C} . For the distribution ww , on the other hand, both $\Delta_C(X_1 X_2)$ and $\Delta_{\bar{C}}(X_1 X_2)$ are relatively small, only about 1–36% of the corresponding NBA probabilities. The errors in $\Delta(X_1 X_2)$ are only of the order of 10–39% in spite of the fact that there is no cancellation between the likelihood errors for C and \bar{C} . For sw and ws , errors are dominated by the corresponding strongly correlated distribution. Thus, for sw , the likelihood error in C is 83–95%, while that of \bar{C} is only 1–17%. The resultant error $\Delta(X_1 X_2)$ is thus dominated by the error of the likelihood in C . A similar result holds for the distribution ws , where now the error in the likelihood for \bar{C} dominates the overall error. For the distribution ss' , although the errors in $\Delta_C(X_1 X_2)$ and $\Delta_{\bar{C}}(X_1 X_2)$ are large, 83–95% in C and 64–78% in \bar{C} , the error in $\Delta(X_1 X_2)$ is relatively small, 6–54%, due to the cancellation in errors between the two likelihoods.

In the case where the correlations in the underlying probability distributions for the likelihood functions are asymmetric, with a correlation between $X_1 X_2$ for C and between $X_1 X_3$ for \bar{C} , the most noteworthy difference to the symmetric case is that, as expected, the Δ_{ij} in the case ss' do not exhibit any cancellation, as the strong correlations are in likelihoods associated with distinct feature sets.

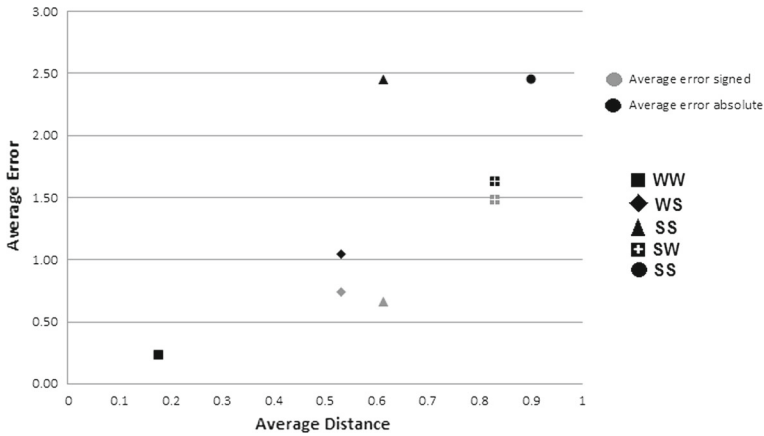


Fig. 3 Graph of average distance as a function of average error for symmetric GBA factorizations in three-feature problem with symmetric correlations in the likelihood functions

7.1 Relating errors to the GBA: three features

We see then that our diagnostics clearly indicate which combination of features should *potentially* be treated together rather than independently and therefore which factorization of the GBA is the most appropriate. We also see that it may be necessary to combine features in only one likelihood as opposed to both. In order to determine whether the features $X_i X_j$ should be combined together and, further, whether they should be combined together for the likelihoods of both C and \bar{C} or only one or none, implies setting a threshold for the errors above which the features will be combined. In this infinite population setting the natural threshold is zero as there are no sampling errors affecting the Δ , which are then a pure measure of the error due to model bias. We will consider using Δ_C to determine when to combine features in the likelihood for C , $\Delta_{\bar{C}}$ for combining features in \bar{C} and Δ to determine if we are likely to see error cancelation or error reinforcement. In this three-feature setting we can consider classifier performance as a function of Δ , where performance is measured in terms of our distance function or classifier accuracy. In Fig. 3 we graph average error against average distance for each correlation distribution ww, ws, sw, ss' and ss (a graph of average error against classifier accuracy is very similar). We show results for both the average signed normalized error and the average absolute error. The difference between the two errors is a measure of the degree of cancelation in the errors between C and \bar{C} . For both performance measures we see a clear correlation between the average error and performance, with, as expected, the distribution ww showing the best performance and ss the worst. Note for the distributions ws, sw and, particularly, the distribution ss' , the difference between the average signed and absolute errors, once again indicating that large errors in the individual likelihoods is not sufficient to predict performance but indicating that our diagnostics do correlate with performance.

8 Cancelations between correlations in different feature combinations

By a systematic analysis of the case of two and three features, we may understand almost all of the principle elements—correlation type (symmetric, asymmetric), correlation strength

(weak, strong) and “correlation correlation” (reinforcing, canceling)—that explain the relative difference in performance between the GBA and the NBA. Moreover, we saw that our diagnostics correlate very well with this difference allowing us to predict a priori when the NBA might be expected to break down and what factorization of the GBA should be used. Our error diagnostics allow us to identify which features should be combined and therefore which factorization of the GNB is better. Furthermore, the analysis also allows us to understand why the NBA is such a robust and versatile performer in spite of its very strong assumption of independence between features. Essentially, the only element missing when passing to more than three features is that of error reinforcement or cancelation between different feature combinations.

8.1 Cancelations for four features

The simplest illustration of the cancelation between different feature combination occurs with two binary schemata, considering concatenations of two two-feature distributions taken from Appendix A. Explicitly, we consider a probability distribution $P_e(\mathbf{X}|C)$ for four features, $X_1, X_2, X_3,$ and $X_4,$ of the form

$$P_e(\mathbf{X}|C) = P(X_1 X_2 X_3 X_4|C) = P(X_1 X_2|C)P(X_3 X_4|C) \tag{8.1}$$

where each $P(X_i X_{i+1}|C)$ can be chosen from an independent distribution. Note that the correlation structure is symmetric in the likelihoods for C and \bar{C} . We will restrict attention to symmetric correlations as for the purpose of studying reinforcement or cancelation of errors between feature sets there is nothing new in the asymmetric version. As, by construction, there are no dependencies between the features $X_i X_j$ for $ij = 13, 14, 23, 24$ the schema partition $\xi^{(e)} = (\xi_1, \xi_2)$ with $\xi_1 = X_1 X_2$ and $\xi_2 = X_3 X_4,$ will be exact and, hence, the GBA based on these two schemata should be an exact approximation. In other words,

$$\begin{aligned} P(X_1 X_2 X_3 X_4|C) &\equiv P(X_1 X_2|C)P(X_3 X_4|C) \\ &= P(\xi^{(1)}|C) = P(\xi_1|C)P(\xi_2|C) \equiv P_{GB}^e(X_1 X_2 X_3 X_4|C) \end{aligned} \tag{8.2}$$

and analogously for $P(X_1 X_2 X_3 X_4|\bar{C})$. In contrast, the NBA gives for the likelihood function

$$P(\mathbf{X}|C) = P(X_1|C)P(X_2|C)P(X_3|C)P(X_4|C) \tag{8.3}$$

while the error in the score function relative to the optimal (exact) factorization is given by

$$\begin{aligned} \Delta_{S_{GB}}(C|X_1 X_2 X_3 X_4) &= \delta_s(\xi_1|C) + \delta_s(\xi_2|C) - \delta_s(\xi_1|\bar{C}) - \delta_s(\xi_2|\bar{C}) \\ &= \delta_s(C|\xi_1) + \delta_s(C|\xi_2) \\ &\equiv \ln\left(\frac{P(X_1 X_2|C)}{P(X_1|C)P(X_2|C)}\right) - \ln\left(\frac{P(X_1 X_2|\bar{C})}{P(X_1|\bar{C})P(X_2|\bar{C})}\right) \\ &\quad + \ln\left(\frac{P(X_3 X_4|C)}{P(X_3|C)P(X_4|C)}\right) - \ln\left(\frac{P(X_3 X_4|\bar{C})}{P(X_3|\bar{C})P(X_4|\bar{C})}\right) \end{aligned} \tag{8.4}$$

We will consider different concatenations of the probability distributions of Appendix A, as discussed in 3.1. We saw there, that for distributions 1, 3, 8, 10 and 11 the NBA was particularly bad, while for distributions 2, 4, 7, 9 and 12 the NBA was better. However, we also saw that there were different reasons why the NBA might work well in the two-feature examples. First, as in distribution 4, that the correlations were weak in both the likelihoods for both C and \bar{C} ; and, alternatively, as in distribution 2, that the correlations were strong in the likelihoods for both C and \bar{C} but that there were cancelations in the errors between

Table 3 Table of errors for four-feature distribution SS

Features	$\Delta S_C(X_1 X_2)$	$\Delta S_{\bar{C}}(X_1 X_2)$	$\Delta S_C(X_3 X_4)$	$\Delta S_{\bar{C}}(X_3 X_4)$
1111	-3.15	0.49	-3.15	0.49
1110	-3.15	0.49	0.67	-1.50
1101	-3.15	0.49	0.61	-1.01
1100	-3.15	0.49	-1.80	0.58
1011	0.67	-1.50	-3.15	0.49
1010	0.67	-1.50	0.67	-1.50
1001	0.67	-1.50	0.61	-1.01
1000	0.67	-1.50	-1.80	0.58
0111	0.61	-1.01	-3.15	0.49
0110	0.61	-1.01	0.67	-1.50
0101	0.61	-1.01	0.61	-1.01
0100	0.61	-1.01	-1.80	0.58
0011	-1.80	0.58	-3.15	0.49
0010	-1.80	0.58	0.67	-1.50
0001	-1.80	0.58	0.61	-1.01
0000	-1.80	0.58	-1.80	0.58

them. Hence, as building blocks for the concatenations we will use distribution 4, denoted as W , which has small errors in the NBA for both likelihoods; distribution 1, S , which exhibits large errors in both likelihoods and of opposite sign; distribution 2, W' , which has large errors for both likelihoods, but with cancelations between them; and, finally, distribution 1, but where C and \bar{C} have been interchanged, S' . This latter artifice has the effect of giving errors for each specific feature combination of different sign to that of the corresponding error of distribution 1. Hence, WW is a concatenation of distribution 4, SW of distributions 1 and 4; SS of distribution 1; WW' of distributions 4 and 2; $W'W'$ of distribution 2 and SS' of distribution 1 with distribution 1 where C and \bar{C} are inverted.

With these in hand we can now examine how errors cancel both at the intra- and inter-schemata level. In Table 3 we see the different errors in the likelihoods for C and \bar{C} for each schema $\xi_1 = X_1 X_2$; $\xi_2 = X_3 X_4$ for the distribution SS . The most notable feature is that, for any schema, the signs of the errors vary between the different feature combinations 11, 10, 01, 00. In fact, as previously stated, for binary features the errors in the likelihood functions for $X_1 X_2$ and $\bar{X}_1 \bar{X}_2$ must be the same and opposite to those of $\bar{X}_1 X_2$ and $X_1 \bar{X}_2$. Although for features with higher cardinality and for schemata of more than two features the situation is more complicated, it is still true that the errors for different feature values cannot all be of the same sign and therefore when different feature value combinations are considered in different schemata it is inevitable that there will be cancelations.

This phenomenon can be seen clearly in Table 3: considering $\delta_s(C|\xi) = \delta_s(C|X_1 X_2) + \delta_s(C|X_3 X_4)$, 8 configurations, $X_1 X_2 X_1 X_2$ and $X_1 X_2 \bar{X}_1 \bar{X}_2$ lead to an enhanced error while another 8, $X_1 X_2 X_1 \bar{X}_2$ and $X_1 X_2 \bar{X}_1 X_2$, are associated with a cancelation in errors. The same is true for $\delta_s(\bar{C}|\xi)$. For example, 1111, 1010 etc. are associated with error enhancement, while 1110, 1101 etc. are associated with error cancelation. This pattern of error enhancement and error cancelation is equally valid for any four binary feature distributions. Hence, in order to analyse the different possibilities for cancelations between the four different likelihood

Table 4 Table of average error for different four-feature distributions

	ΔS_1	$ \Delta S_1 $	ΔS_2	$ \Delta S_2 $	ΔS_C	$ \Delta S_C $	$\Delta S_{\bar{C}}$	$ \Delta S_{\bar{C}} $	ΔS_T	$ \Delta S_T $	S_{GNB}	S_{NB}	[%]
WW	0.228	0.228	0.228	0.228	0.263	0.304	0.121	0.152	0.323	0.456	2.511	2.305	50
SW	2.453	2.453	0.228	0.228	1.558	1.709	0.896	0.972	2.453	2.681	2.757	2.091	121
SS	2.453	2.453	2.453	2.453	2.476	3.115	1.258	1.792	3.010	4.907	3.008	0.136	23,956
WW'	0.228	0.228	0.662	2.453	1.558	1.709	0.896	0.972	0.754	2.681	2.288	2.091	37
SS'	2.453	2.453	2.403	2.403	1.859	2.403	1.867	2.453	2.681	4.857	2.731	0.134	1667
W'W'	0.662	2.453	0.662	2.403	2.476	3.115	1.258	1.792	1.218	4.907	1.216	0.136	6595

functions in this four-feature problem we consider the following quantities: ΔS_i , $i = 1, 2 = \delta_s(C|\xi_i) = \delta_s(\xi_i|C) - \delta_s(\xi_i|\bar{C})$ is the sum of the signed errors for each feature combination, while $|\Delta S_i| = |\delta_s(\xi_i|C)| + |\delta_s(\xi_i|\bar{C})|$ is the sum of the absolute errors in the two likelihoods. Similarly, $\Delta S_C = \delta_s(\xi_1|C) + \delta_s(\xi_2|C)$ is the sum of the signed errors for the likelihoods of C summed across the two schemata ξ_1 and ξ_2 . $\Delta S_{\bar{C}} = \delta_s(\xi_1|\bar{C}) + \delta_s(\xi_2|\bar{C})$ is the analogous quantity for the likelihoods of \bar{C} . Finally, $\Delta S_{total} = \delta_s(\xi_1|C) - \delta_s(\xi_1|\bar{C}) + \delta_s(\xi_2|C) - \delta_s(\xi_2|\bar{C})$ is the signed error for the full feature set, while $|\Delta S_{total}| = |\delta_s(\xi_1|C)| + |\delta_s(\xi_1|\bar{C})| + |\delta_s(\xi_2|C)| + |\delta_s(\xi_2|\bar{C})|$ is the sum of the absolute errors across all four likelihood functions.

Table 4 shows the absolute values of these different diagnostics averaged over the 16 different feature combinations 1111, 1110, . . . , 0000 for each concatenation. For the homogeneous distributions WW and SS , we see that $\Delta S_i = |\Delta S_i|$, $i = 1, 2$, indicating that there are no cancelations between the errors in the likelihoods for C and \bar{C} in a given schema. In other words, the errors in the likelihoods for C and \bar{C} reinforce one another rather than cancel. On the other hand, in both cases, $\Delta S_i < |\Delta S_i|$, $i = C, \bar{C}$, which indicates that there are cancelations between schemata, for both likelihood functions, as illustrated above for the case SS . In the case of the distribution SW , once again there are no cancelations between the errors for C and \bar{C} in a given schemata but there are between schemata. For the distributions WW' , SS' and $W'W'$, all three contain distributions, S' or W' , where there is a cancelation of errors between the likelihoods for C and \bar{C} within a given schemata, i.e., that $\Delta S_i < |\Delta S_i|$ for $i = 2$ for WW' and SS' , or both in the case of $W'W'$. Additionally, there are also cancelations between schemata for all three distributions. By comparing ΔS_{total} with $|\Delta S_{total}|$ we can see the extent of the overall error cancelation for the score function. By far the largest reductions are associated with those distributions, WW' , SS' and $W'W'$, where there are cancelations at both the intra- and inter-schemata level. The least cancelation is for the distribution SW . This is due to the fact that it is a concatenation of a distribution, S , with large errors, with another, W , with small errors. However, although the greatest error cancelation is associated with those distributions where there are both intra- and inter-schemata cancelations they do not correspond necessarily to those distributions where the average absolute difference between the NB score and the exact score is highest. Rather, the largest differences in score are associated with those distributions where the NB scores are small. As mentioned previously, the NBA has inadequate raw material with which to work.

8.2 Cancelations for more than four features

We can concatenate the two-feature distributions as many times as we like to see how things change as a function of the number of features and as a function of the mix of correlations. For instance, for six and eight features we will take the probability distributions for the likelihood functions to be

Table 5 Table of average error for different six- and eight-feature distributions

	ΔS_C	$ \Delta S_C $	$\Delta S_{\bar{C}}$	$ \Delta S_{\bar{C}} $	ΔS_{total}	$ \Delta S_{total} $	Score NBG	Score NB	[%]
WWW	0.366	0.470	0.158	0.235	0.410	0.706	3.435	3.237	15.90
SWW	1.625	1.921	0.925	1.082	2.533	3.003	3.557	2.379	506.64
SSW	2.598	3.372	1.311	1.928	3.139	5.300	3.702	2.158	154.54
SSS	3.338	4.823	1.559	2.774	3.838	7.598	3.861	0.175	5708.10
WW'W	1.625	1.921	0.925	1.082	0.863	3.003	2.867	2.379	156.64
SS'S	2.772	4.140	2.177	3.457	3.799	7.598	3.799	0.175	4471.94
WWWW	0.436	0.607	0.179	0.304	0.464	0.912	3.691	3.377	114.07
SWWW	1.601	2.013	0.896	1.124	2.453	3.137	3.879	3.136	110.82
SSWW	2.558	3.419	1.282	1.944	3.076	5.363	4.084	2.306	942.91
SSSW	3.296	4.824	1.530	2.764	3.738	7.588	4.296	2.091	185.62
SSSS	4.021	6.230	1.888	3.584	4.514	9.814	4.511	0.195	8574.92
WW'WW'	2.558	3.419	1.282	1.944	1.351	5.363	3.057	2.306	453.01
SS'SS'	2.993	4.907	2.993	4.907	3.093	9.814	4.023	0.193	2281.02

$$\begin{aligned}
 P_e(\mathbf{X}|C) &= P(X_1 X_2 X_3 X_4 X_5 X_6 | C) \\
 &= P(X_1 X_2 | C) P(X_3 X_4 | C) P(X_5 X_6 | C)
 \end{aligned}
 \tag{8.5}$$

$$\begin{aligned}
 P_e(\mathbf{X}|C) &= P(X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 | C) \\
 &= P(X_1 X_2 | C) P(X_3 X_4 | C) P(X_5 X_6 | C) P(X_7 X_8 | C)
 \end{aligned}
 \tag{8.6}$$

with analogous expressions for \bar{C} , so that the correlation structure is symmetric.

For the 6-feature case there are three order-two schemata, $\xi_1 = X_1 X_2$, $\xi_2 = X_3 X_4$, $\xi_3 = X_5 X_6$, and for the 8-feature case four, $\xi_1 = X_1 X_2$, $\xi_2 = X_3 X_4$, $\xi_3 = X_5 X_6$ and $\xi_4 = X_7 X_8$, with no dependencies in either case between features that are in different concatenated two-feature blocks. The schema partitions $\xi^e = (\xi_1, \xi_2, \xi_3)$ and $\xi^e = (\xi_1, \xi_2, \xi_3, \xi_4)$, with $\xi_1 = X_1 X_2$, $\xi_2 = X_3 X_4$, $\xi_3 = X_5 X_6$, $\xi_4 = X_7 X_8$ will be exact as will be the GBA based on this factorization. The corresponding errors in the score function can be calculated from Eq. (3.14). In these examples, we have constructed all the dependencies between different variables explicitly. By choosing as schemata $\xi_1 = X_1 X_2$ and $\xi_2 = X_3 X_4$, for four features, $\xi_1 = X_1 X_2$, $\xi_2 = X_3 X_4$, $\xi_3 = X_5 X_6$ for 6 features and $\xi_1 = X_1 X_2$, $\xi_2 = X_3 X_4$, $\xi_3 = X_5 X_6$ and $\xi_4 = X_7 X_8$ for eight features we have chosen a GBA that matches the exact factorization of the likelihood function and, hence, the GBA is exact. The implication of this is that the error of the NBA is just the sum of the errors associated with the four order-two dependencies arising from the four schemata ξ_1 , ξ_2 , ξ_3 and ξ_4 .

We will consider concatenations of the distributions W , W' , S and S' of Sect. 8.1, for six features considering the distributions WWW , SWW , SSW , SSS , $WW'W$ and $SS'S$ and for eight features the distributions $WWWW$, $SWWW$, $SSWW$, $SSSW$, $SSSS$, $WW'WW'$ and $SS'SS'$. In Table 5, for each distribution we see our error measures— ΔS_C , $\Delta S_{\bar{C}}$, ΔS_{total} , $|\Delta S_C|$, $|\Delta S_{\bar{C}}|$ and $|\Delta S_{total}|$, as well as the GNB and NB scores. We see manifest the phenomenon of error cancellation at both the intra- and inter-schemata level. Note that the examples with the greatest degree of cancellation are those with feature pairs associated with the strongly correlated distributions W' and S' that exhibit important error cancellations between the likelihoods for C and \bar{C} , leading to cancellations of up to 75% of the absolute error with the main contribution coming from cancellations of the errors in ΔC and $\Delta \bar{C}$. The symmetric

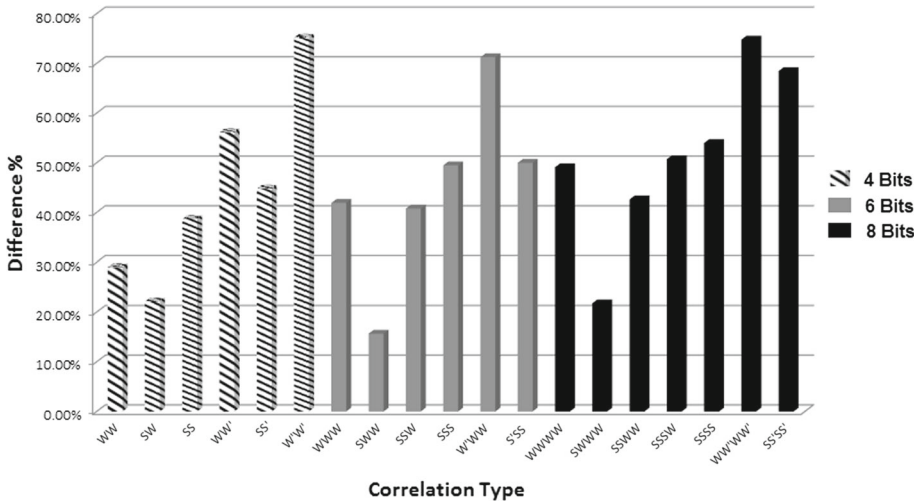


Fig. 4 Graph of relative score cancellation versus correlation type

distributions *WWW*, *SSS*, *WWW* and *SSSS* illustrate very well the existence of cancelations across different feature value combinations given that not all can have the same sign of error. Moreover, this type of cancelation increases as the number of features increases. Indeed, we see that, even if there are very strong correlations with errors that reinforce between the likelihoods of *C* and \bar{C} , i.e., without intra-schemata cancelation, the overall error cancelation due to inter-schemata cancelation is more than 50% of the absolute error and this reduction is about the same for the weakly correlated case *WWW*.

These effects are summarized in Fig. 4: For a given correlation type—(*WW*, *WWW*, *WWW*), (*SS*, *SSS*, *SSSS*)—the relative degree of cancelation is an increasing function of the number of features. This repeated concatenation of the same distribution shows and isolates the effect of inter-schemata error cancelation between different feature values combinations in different modules due to the fact that the error function cannot be of the same sign over all feature value combinations. We also see the enhanced cancelation for distributions with *W'* and *S'* modules due to the addition of intra-schemata cancelation.

8.3 Performance as a function of number of features and degree of correlation

So, how do our error measures relate to performance in these multi-feature cases? In Figs. 5 and 6 we see the performance of the NBA as a function of our signed score error measures, ΔS_{total} , averaged over all feature value combinations for the four- six- and eight-feature distributions considered in the previous Sections (the results are very similar for the absolute error measure $|\Delta S_{total}|$). The most notable feature of the graphs is the good degree of correlation between the error measure and the corresponding performance measure, with an R^2 of approximately 0.7, clearly showing that our error diagnostics do predict relative performance. We can also note from these graphs that the correlations between points associated with a fixed number of features are stronger than when considering the full set of distributions.

We have used two different performance metrics as they represent two different measures of classifier performance. As emphasized by Domingos and Pazzani (1996), the all or nothing nature of classification should explain some of the robustness of the NBA. Indeed, we can

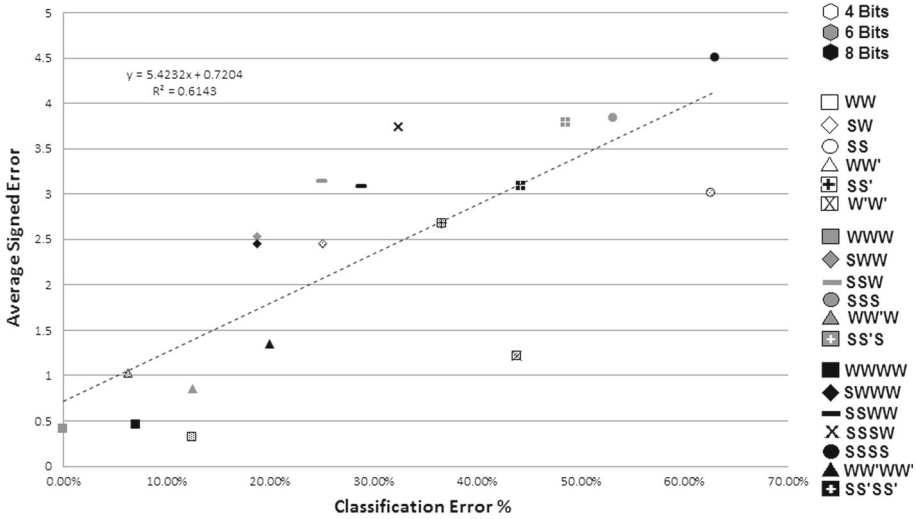


Fig. 5 Graph of classification error versus ΔS_{total} for different 4, 6 and 8 feature distributions

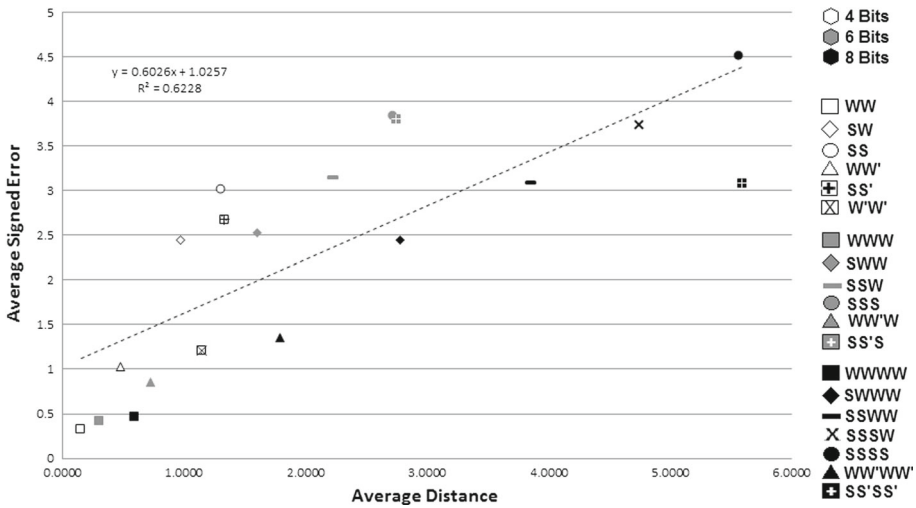


Fig. 6 Graph of distance versus ΔS_{total} for different 4, 6 and 8 feature distributions

confirm this very well using the present analysis. Indeed, there is a low degree of correlation ($R^2 = 0.32$) between the two measures for the 19 concatenated distributions we have considered. Why the low correlation? Well, the distance measure is a metric that determines the similarity between the NBA ranking and the GBA (in this case, exact) ranking and is global over the full set of predictions. On the other hand, classification performance is particularly sensitive to the NB score in the vicinity of the score threshold, i.e., $S_{NBA} = 0$. This means that even large errors are relatively unimportant if the NB score is far from the threshold and, on the contrary, the effect of small errors may be significantly amplified in the vicinity of the threshold.

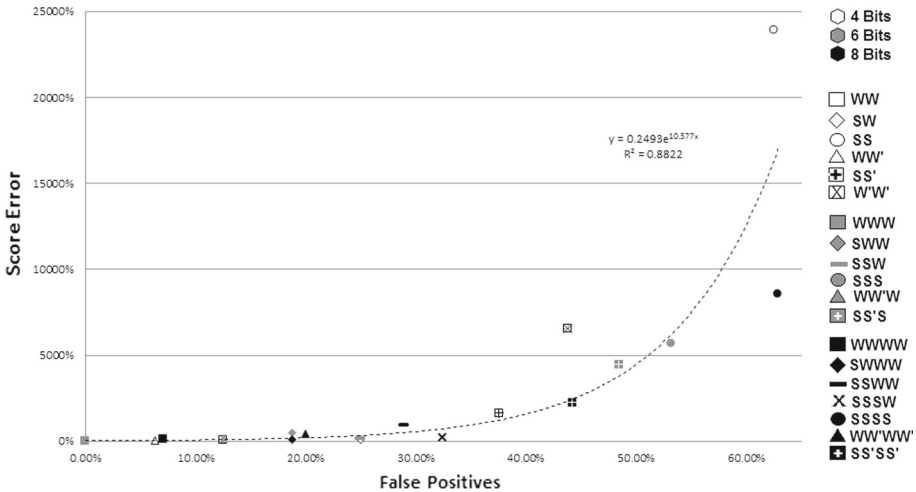


Fig. 7 Graph showing relative score error against false positive rate for different four, six and eight feature distributions

In Fig. 7 we see the relation between the relative score error, i.e., relative to the NB score, versus classification error (false positive rate). The strong relationship between them confirms the importance of the vicinity of the score threshold $S_{NB} = 0$, where the relative error would be expected to be the largest and hence the highest sensitivity to misclassification. On the contrary, the correlation between the relative score error and the distance metric is weak showing that the two metrics are sensitive to quite different characteristics of the error function.

9 Application to real world data

In this Section we will briefly show how the insights gleaned from our previous analysis can be applied to real world problems. We have seen how local error measures for the likelihoods, Equations (3.8); and for the scores, (3.12), can be used to determine which features should be combined and therefore how to construct an appropriate factorization for the GNB. However, these diagnostics are all associated with probabilities and, therefore, independent of sample size. For example, for two features, if $P(X_1X_2|C) = N_{CX_1X_2}/N_C = 0.3$, $P(X_1|C) = N_{CX_1}/N_C = 0.4$ and $P(X_2|C) = N_{CX_2}/N_C = 0.4$, we must consider the possibility that the error $\delta(X_1X_2|C) = 0.14$ is not statistically significant if $N_{X_1X_2}$, N_C , N_{X_1} or N_{X_2} are small. To determine the degree of statistical significance of the errors (3.8) we will use the following binomial tests

$$\varepsilon(\xi|C) = \frac{N_C \delta(\xi|C)}{\sqrt{N_C P_{NB}(\xi|C)(1 - P_{NB}(\xi|C))}} \tag{9.1}$$

where, as throughout the paper, $C = C$ or \bar{C} . Given that $\delta(\xi|C) = (P(\xi|C) - P_{NB}(\xi|C))$ the test is taking as null hypothesis that there are no correlations between the attributes of ξ . Thus, the test determines the degree to which the actual observation $P(\xi|C)$ is inconsistent with the null hypothesis. The likelihood error will be taken to be statistically significant if ε exceeds some hypothesis testing threshold. For instance, in the case where the binomial

distribution may be approximated by a normal distribution, $\varepsilon = 1.96$ would correspond to the 95% confidence interval that that error value did not occur by chance relative to the Naive Bayes null hypothesis. In the case where the distributions are not well approximated by a normal distribution, hypothesis testing can use a more sophisticated approximation, using for example the Wilson intervals. Another problem in the finite sample setting is the possibility of having $N_{C\xi} = 0$. This implies that $P(\xi|C) = 0$ and will lead to infinite score contributions from the corresponding schema. To avoid this a smoothing, such as the Laplace correction or m-estimates, can be used.

We use as criteria that $|\varepsilon(\xi|C)| > 2$ and $|\varepsilon(\xi|\bar{C})| > 2$ in order to determine those feature sets which should be combined together, and consider all possible combinations of feature values. We consider two different combination algorithms: one where features are combined independently in the two likelihoods—the asymmetric GNB (GNB_a)—and one where they are combined together—the symmetric GNB (GNB_s). For instance, in the former, if $\varepsilon(X_1 X_2|C) = 3.1$ and $\varepsilon(X_1 X_2|\bar{C}) = 0.9$, then the features X_1 and X_2 would be combined only in the likelihood for C and not \bar{C} . On the other hand, in the symmetric case they would be combined in both likelihoods. It may occur that a given feature qualifies to be combined as a member of more than one feature combination. For example, for three features, if $\varepsilon(X_1 X_2|C) = 3.1$, $\varepsilon(X_1 X_3|C) = 2.4$ and $\varepsilon(X_2 X_3|C) = -0.2$, then the feature X_1 qualifies to be combined with both X_2 and X_3 . If $\varepsilon(X_1 X_2 X_3|C) = 1.8$, or if we restrict to only binary combinations, then X_1 can only be combined with one other feature for a given feature vector. In this case we choose the feature combination with the highest value of ε . In the present example, this would mean that X_1 is combined with X_2 , as $\varepsilon(X_1 X_2|C) > \varepsilon(X_1 X_3|C)$. In the case where different combinations resulted in the same value of ε , then the value of δ_s for the combined features was used to break ties. For simplicity, we considered only binary schemata, i.e., we combined only up to two features. Besides simplicity, another previously discussed reason for this is that two-feature samples are inevitably larger than three-feature samples and therefore, all else being equal, lead to higher values of ε .

We considered 20 data bases from the UCI repository, as seen in Table 6. We also considered three text mining data sets as this is one area where the NBA is still considered to be competitive. These latter data sets are taken from the KEEL data set repository (<http://sci2s.ugr.es/keel/textClassification.php>) and are also shown in Table 6. For simplicity we considered each problem as a two-class problem. In multi-class problems we took the smallest class case and in the binary domains we took the class specified at the UCI repository. Numeric attribute values were discretized by dividing into a fixed number of bins and choosing the bin intervals so that each bin contained approximately the same number of elements. Ten bins were chosen as the default. However, in the case of data bases with few elements we considered a smaller number. For each data base we performed random subsampling with a 70/30 training/holdout split repeated 20 times. Note that no tuning of the GBA was considered. For example, no feature selection algorithm was used. Features in the schema ξ were combined when $|\varepsilon(\xi)| > 2$. As performance metrics we considered classifier error and AUC. We compared our two GBA approximations—symmetric and asymmetric—against the NBA implemented in WEKA, as well as three other state-of-the-art classifiers available in WEKA: AODE, WAODE and HNB. Each classifier was implemented on exactly the same set of training/holdout data for each of the 20 runs. The default Laplace smoothing in WEKA was used, where $(N_{CX}/N_C) \rightarrow (N_{CX} + 1/2)/(N_C + 1)$. The results can be seen in Tables 6 and 7 for classification error and AUC respectively.

We compared the errors and AUC of the different classifiers, where the error was averaged over the 20 different runs. We then used a binomial test to determine the statistical significance of the performance difference of our five enhanced classifiers, taking as null hypothesis

Table 6 Error for NB, GNBs, GNBa, AODE, WAODE and HNB for 20 UCI and 3 text mining data bases

Domain	Attributes	Cases	Error NB	Error GNB_S	Error GNB_A	Error AODE	Error WAODE	Error HNB
Mushroom	22	8000	4.75%	0.13%+	0.16%+	0.04%+	0.01%+	0.04%+
Pendigits	17	10,992	5.01%	1.55%+	2.05%+	0.74%+	0.44%+	1.07 %+
Segment	20	2310	12.74%	6.35%+	6.39%+	2.81%+	2.18%+	2.61 %+
Vehicle	19	846	13.34%	5.38%+	6.30%+	3.94%+	3.59%+	2.54%+
Anneal	39	898	2.39%	0.72%+	0.78%+	1.76%+	0.85%+	0.74%+
Chess (kr-kp)	37	3169	12.04%	7.48%+	7.71%+	8.84%+	6.13%+	7.41%+
Hypothyroid	26	3163	3.60%	2.22%+	2.43%+	2.40%+	1.84%+	2.01%+
Letter recognition	17	20,000	1.79%	1.36%+	1.46%+	1.05%+	1.12%+	1.09%+
Satellite	37	6435	15.12%	13.13%+	13.49%+	12.22%+	15.53%	14.38%+
Adult	15	48,842	18.17%	15.94%+	15.96%+	15.65%+	15.23%+	15.96%+
House-votes	17	435	10.08%	8.73%	8.73%+	6.17%+	5.59%+	6.44%+
Tic-tac-toe	10	958	30.17%	29.27%	30.00%	25.99%+	27.26%+	23.47%+
Ionosphere	35	351	10.65%	10.48%	10.52%	8.22%+	6.32%+	7.27%+
Credit crx	16	690	13.31%	16.71%*	16.16%*	12.44%	13.59%	13.67%
Hepatitis	20	155	19.36%	16.41%+	15.98%+	17.55%	18.08%	17.22%
Cancer (bcw)	10	699	2.84%	3.16%	2.99%	3.26%	3.79%*	4.43%*
Statlog (heart)	14	270	15.00%	17.59%*	17.41%*	15.31%	16.91%*	17.41%*
Post-operative	9	90	32.09%	28.33%	28.33%	32.27%	34.85%	36.16%*
Liver (bupa)	7	345	33.86%	37.96%*	36.75%*	34.92%	35.90%	36.14%
Wine	14	178	2.81%	1.79%	0.85%+	1.03%+	2.44%	1.13%+
BCF	100	913	1.31%	2.44%*	2.60%*	1.30%	1.41%	1.95%*
BlogGender	100	3232	32.52%	32.74%	33.51%*	32.05%+	31.88%+	32.25%+
C20	100	13,929	6.79%	5.87%+	5.84%+	5.19%+	4.97%+	5.91%+
Wilcoxon test Z				-2.35+	-2.48+	-3.43+	-2.48+	-2.37+

Table 7 AUC for NB, GNBs, GNBa, AODE, WAODE and HNB for 20 UCI and 3 text mining data bases

Domain	Attributes	Cases	AUC NB	AUC GNB_S	AUC GNB_A	AUC AODE	AUC WAODE	AUC HNB
Mushroom	22	8000	99.76%	99.87%+	99.86%+	100.0%+	100.0%+	100.0%+
Pendigits	17	10,992	97.68%	99.44%+	99.28%+	99.96%+	99.96%+	99.93%+
Segment	20	2310	94.66%	98.12%+	98.30%+	99.47%+	99.69%+	99.93%+
Vehicle	19	846	93.59%	97.60%+	97.41%+	98.86%+	99.03%+	99.31%+
Anneal	39	898	99.66%	99.48%*	99.42%*	99.78%	99.92%+	99.95%+
Chess (kr-kp)	37	3169	95.35%	98.11%+	98.03%+	97.29%+	98.50%+	98.25%+
Hypothyroid	26	3163	98.66%	98.76%	98.61%	98.69%	98.86%	98.82%
Letter recognition	17	20,000	97.95%	98.81%+	98.78%+	99.41%+	99.67%+	99.75%+
Satellite	37	6435	92.39%	92.89%+	93.05%+	93.36%+	93.73%+	94.32%+
Adult	15	48,842	90.00%	90.53%+	90.56%+	90.92%+	90.94%+	89.24%*
House-votes	17	435	97.41%	96.17%*	96.17%*	98.72%+	98.74%+	98.69%+
Tic-tac-toe	10	958	73.38%	74.35%	73.07%	82.09%+	79.54%+	84.67%+
Ionosphere	35	351	95.10%	92.55%*	93.00%*	97.98%+	98.11%+	97.86%+
Credit crx	16	690	93.25%	89.52%*	89.63%*	93.46%	92.90%	92.74%
Hepatitis	20	155	87.43%	83.67%*	84.41%*	88.00%	85.46%	86.64%
Cancer (bcw)	10	699	99.25%	97.79%*	97.78%*	99.92%	99.09%	99.11%
Statlog (heart)	14	270	90.98%	87.60%*	87.86%*	90.93%	90.04%	90.31%
Post-operative	9	90	35.50%	38.88%	39.00%	28.51%*	31.48%	28.81%*
Liver (bupa)	7	345	70.57%	62.69%*	65.94%*	69.02%	67.46%*	67.58%*
Wine	14	178	99.88%	97.26%*	97.26%*	99.90%	99.88%	99.92%
BCF	100	913	99.65%	94.07%*	94.11%*	99.67%	99.67%	99.43%*
BlogGender	100	3232	74.79%	75.01%	73.88%	75.41%	75.53%	75.24%
C20	100	13,929	93.83%	90.27%*	90.41%*	94.30%	94.56%+	93.97%
Wilcoxon Test Z			-0.28	-0.34	-2.32+	-1.65	-1.41	

the WEKA NBC. We judged the performance difference to be significant if it was at the $p < 0.05$ level. The entries denoted with a “+” are where there was a significant performance improvement for the enhanced classifier relative to the NBC and * for those cases where the NBC was significantly better. Entries without a symbol correspond to no statistically significant difference over the 20 runs considered. As multiple pair-wise comparisons can be problematic we also used the Wilcoxon rank test (Demsar 2006), comparing each enhanced classifier to the NBC. We show the corresponding z statistic defined as

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \quad (9.2)$$

where N is the number of data sets, $T = \min(R_+, R_-)$ and R_+ is the sum of ranks for the data sets on which the NBC outperforms the GBA and R_- is the sum of ranks in the contrary case. The null hypothesis that the two algorithms have equal performance can be rejected at the 95% confidence level if $z < -1.96$. In terms of classifier error, we see that the Wilcoxon rank test shows that all versions of the GNB are significantly better than the NBC. On the other hand, in terms of AUC we see that, in terms of the Wilcoxon rank test, only the AODE classifier shows a statistically significant enhanced performance relative to the NBC. The strong performance of the NBC as a ranking algorithm has been amply demonstrated previously (Zhang and Su 2004, 2008), where it has been shown that in terms of ranking the NBA is better or equivalent to C4.4.

We may, of course, analyse these results from the perspective of “designing algorithms that give better performance across a wide set of problem domains”, comparing our symmetric and asymmetric GNB to the three established classifiers AODE, WAODE and HNB. As emphasised, our goal here was not to design a new state of the art classifier. Neither was it to show that knowledge of, and diagnostics for, the errors inherent in the NBA can be used to identify which features should be combined and these combinations included into a better performing classifier, although that is, indeed, an important result of this research. Rather, our goal was to show that our diagnostics could predict a priori which classifier—NBC or GBC—would work best on which problem.

That they do predict is manifest in the relatively high degree of correlation between our chosen error metric and classifier performance. Explicitly, for the UCI and KEEL data sets we consider the relation between the average of the absolute value of the signed error ΔS_{total} , averaged over all feature vectors in the training set, versus the relative difference in classifier error between the NBC and GBC, where to calculate ΔS_{total} for a given feature vector we include only combinations with statistically significant errors. Large values of the error correspond to those problem sets that exhibit significant correlations when averaged over the full training set and therefore one would expect the NBA to be less effective. In Table 8 we give a summary of these correlations for the 20 UCI data sets and the 19 artificial distributions considered in Sect. 8.2. For the UCI data sets, we take as performance measure the relative difference in error between the GNB and the NBC, while for the artificial distributions we take error itself, as in this case the GNB is exact by construction. In Table 8 we see that the Pearson correlation coefficients for our symmetric and asymmetric GBC are smaller on the UCI data sets than was present in the artificial distributions. However, all the correlations shown are statistically significant at the 95% confidence level using a one-tailed t-test, thereby giving clear evidence that our chosen error measure is predictive of performance of the GBC versus the NBC. Interestingly, our error measure is also a good predictor of performance for the classifiers AODE, WAODE and HNB. Given that our bespoke symmetric and asymmetric GBCs were designed to account for correlation

Table 8 Pearson correlation coefficients between the average of the absolute error $|\Delta S_{total}|$, averaged over all feature vectors in the training set, versus the relative difference in classifier error between the NBC and GBC for each classifier

Classifier	Correlation coefficient	Correlation coefficient (no ionosphere)
GNB_s UCI	-0.45	-0.65
GNB_a UCI	-0.47	-0.68
AODE UCI	-0.61	-0.77
WAODE UCI	-0.53	-0.59
HNB UCI	-0.52	-0.62
Avg all	-0.54	-0.68
NBC artificial	0.78	NA

All correlation coefficients are statistically significant at the 95% confidence level

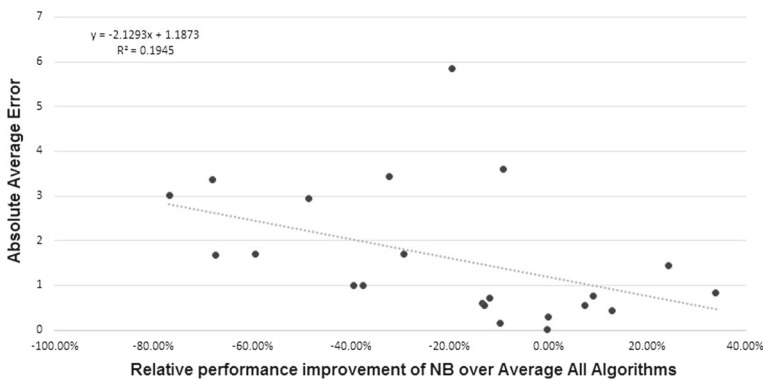


Fig. 8 Relation between the percentage relative difference in error between the NBA and GBA and the average absolute error for the 20 UCI and 3 KEEL data bases and averaged over the 5 GBA classifiers

errors by combining significantly correlated attributes in an alternative factorization of the likelihoods, it is perhaps not surprising that the GNB_s and GNB_a classifiers’ performance relative to the NBC is greater the greater the magnitude of the attribute correlations. However, it is gratifying to observe that the relative performance of the state of the art AODE, WAODE and HNB classifiers is also highly correlated with our error diagnostics. This is linked to the fact that all the classifiers we consider are trying to account for attribute dependencies by relaxing the NB maximal factorization criterion. They just do it in different ways.

In Fig. 8 we can see graphically the correlations between error and relative performance improvement for the UCI and KEEL data sets averaged over all 5 enhanced GNB classifiers. We can also notice the presence of a significant outlier—the ionosphere data set—where there is a very high correlation error but only a small enhancement of the GBA over the NBA. Indeed, if we consider the Pearson correlation coefficients for the UCI data sets in Table 8 without the ionosphere outlier we can see a significant increase in the correlation coefficients. Of course, we do not wish to data snoop in order to improve the results, but do wish to use this case to point out that, although it is remarkable the degree of correlation between our very simple error measure and relative performance of the GBC versus NBC, one would certainly expect there to be other factors, potentially many, that affect the relation

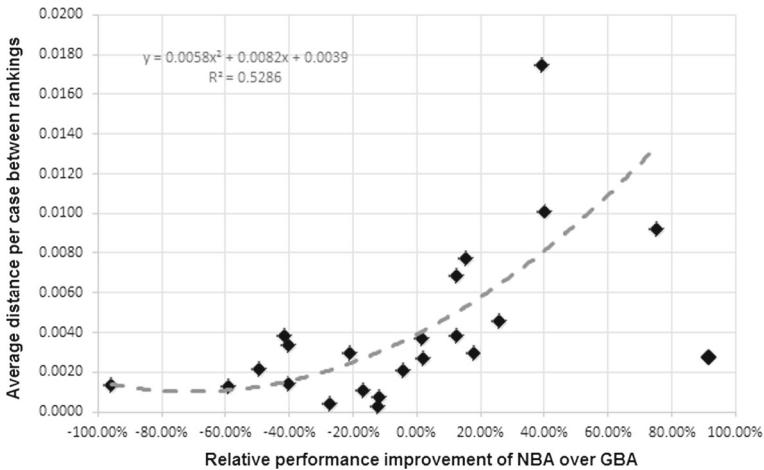


Fig. 9 Relation between the percentage relative difference in error between the NBA and GBA_s and GBA_a classifiers and the average distance in ranking per case

between them. For instance, we may note that the ionosphere data set has a substantially lower fraction of cases to attribute values than the other data sets. We may also point out that the distribution of correlation errors potentially contains a great deal of information which could be used to intuit classifier performance, much more than we have used here where only one, single overall global measure has been used. In Fig. 9 we see a graph of the average distance per case between the NBC and GBC rankings for the set of feature vectors of the training set, versus the relative difference in classifier error between the NBC and GBC. As we can see, there is a clear correlation, implying that the NBC is a better performer in conditions where the average rank distance divided by the total number of cases is larger. Given the enormous heterogeneity of the UCI and KEEL problems, it is very gratifying to see that there are diagnostic metrics that can begin to differentiate between those problems where the NBA would be adequate, versus those where a more sophisticated algorithm is required. Indeed, these results strongly indicate the potential for developing meta-prediction algorithms that could predict the performance of a given algorithm on a given problem.

The efficacy of this idea can be tested relative to the possibility of performance “prediction” via an internal cross-validation, or an internal separate hold-out set. In the latter, one considers the performance of different prediction algorithms using an internal cross-validation on a training set, then uses the ranked performance to predict how the different algorithms will perform on a separate hold-out set. In the sense of our diagnostics, we do not consider this to be a prediction but, rather, a validation, in that each and every algorithm has to be tried and tested multiple times before predicting which one will perform best. In our formalism, no algorithm is tested a priori on any data set. Rather, our statistical diagnostics are used to predict which algorithm type will give better performance before even running the algorithm. To test the two approaches we considered the combined set of UCI and KEEL data bases: in the case of a performance prediction based on an internal cross-validation, we compared the performance of each of our tested algorithms on each problem using a 15-fold internal cross validation and then tested the algorithms on a hold out set using 5 independent runs. We then determined the performance of the best algorithm on the internal cross-validation across

the hold out set, determining for each problem if the best on the internal cross validation turned out to be the best on the hold out set. As may be expected, given enough runs in order to reduce sample error, the best performer on the internal cross validation was the best performer on the hold out set with a classification accuracy of 100% and an AUC of 1. This, as we emphasise however, is not an a priori prediction, but depends on running the algorithm on each and every new problem. Such cross validation can also be used [Pazzani \(1996\)](#) to determine, in an empirical fashion, which combinations of variables lead to improved classifier performance. However, this must be done using a search algorithm to consider the different possible combinations and then run the resulting algorithm on both a training set and a cross validation set.

In contrast, our meta-algorithm approach considers only the relation between algorithm performance and our a priori diagnostic on a set of problems, which can then be used to predict the relative performance of the algorithms on a new, previously unseen problem. To illustrate the explicit implementation of this approach, we constructed the empirical relation between the percentage difference in error between the NBA and the GBA and the average absolute error, as seen in [Fig. 8](#), but keeping one problem as a hold-out for which a prediction of whether the GBA or NBA would give better performance is made. Thus, we construct a one-variable meta-prediction model which, based on the relation between our diagnostic and algorithm performance on 22 known problems, determines which algorithm type will perform better on a new unknown problem. Explicitly, we perform a linear regression on the 22 known problems which provides the required relation between the relative performance of the GBA versus NBA and absolute average error. This error is then determined for a new, unknown problem, which, using the regression model, can then be related to relative performance improvement. Thus, if this predicted performance improvement is positive/negative then we predict that the NBA will give better/worse performance than the GBA. The performance of our meta-algorithm is summarized by a 17% error rate and an AUC of 0.88. We believe that this degree of predictability given the simple nature of our diagnostic is very encouraging.

10 Conclusions

The NBA, and associated NBC, is widely used in a multitude of different contexts. It has been shown to be both remarkably robust and effective across many problem domains in spite of the strong assumption that all attributes are independent. There are instances where it does not work particularly well however, and so it has been an active area of research to both understand why it works and also to design improvements. In terms of understanding why it is so robust, there have been distinct proposals. One hypothesis is that it is an artefact of the nature of the performance measure of the majority of problems it has been applied to—binary classification. A second line of thought has been that it is possible that attribute dependence is not a sufficient condition to invalidate the NBA, rather, it is the distribution of dependencies across classes, and among the attributes themselves, that governs its validity. However, up to now, there has been no quantification of exactly how dependencies can cancel, under what circumstances, and how to measure it. This latter point is fundamental as, if it is possible to quantify deviations from the NBA, then it is possible to predict a priori when, and under what circumstances, the NBA will be inadequate and thereby know when to implement a more sophisticated, but more costly, alternative algorithm. Moreover, by better understanding the relation between problem structure and algorithm structure,

such insight should permit the design of bespoke classifiers that are better suited to a given problem. Of course, a third attribute of the NBA, that offers a relative advantage, is that it is based on a maximal factorization of the likelihoods, so each factor is associated with a larger sample, and, therefore, smaller sampling errors, than any factor containing more than one feature. This relates more to the relative advantage of the NBA in terms of model variance.

We developed a framework that can be used to determine the presence of dependencies among the attributes within an arbitrary feature combination (schema), ξ , and, more importantly, determine when, how, and to what extent, they affect the estimation of different performance metrics, such as estimating posterior probabilities and classification accuracies using the NBA. Our analysis is based on the assumption, used in generalizations of the NBA, that there exist better factorizations, for instance of the likelihood functions $P(\mathbf{X}|C)$ and $P(\mathbf{X}|\bar{C})$, than the complete factorization associated with the NBA. The question then is: what is the error associated with a given factorization, i.e., a given realization of the GBA, relative to the NBA? In analyzing this error, we made no assumption that the optimal factorization of $P(\mathbf{X}|C)$ was the same as that of $P(\mathbf{X}|\bar{C})$, presenting evidence that there are problems (probability distributions) where they manifestly were not the same. We showed that cancelations can and do occur at both the intra- and inter-schemata levels, the former showing that cancelations could occur between the likelihoods for a class and its complement, and the latter showing that cancelations could occur between the likelihoods for different feature combinations but for the same class. In fact, we showed that it was inevitable that there were such inter-schemata cancelations given that the signs of the errors for a given feature combination for different feature values could not all be the same.

In order to quantify the degree of error cancelation or error reinforcement, we introduced a set of diagnostics— $\delta(\xi|C)$, $\delta(\xi|\bar{C})$, $\delta(\xi)$, $\delta_s(\xi|C)$, $\delta_s(\xi|\bar{C})$, $\Delta_C(\xi)$, $\Delta_{\bar{C}}(\xi)$, $\Delta(\xi)$, $\Delta_s(\xi|C)$ and $\Delta(\xi)$. $\delta(\xi|C)$, and its analog for \bar{C} , measure the degree of dependency between the attributes within a given schema in the expression for the likelihood function for the class C , or its complement, \bar{C} . We showed that strong dependencies, as exhibited by $\delta(\xi|C)$ and $\delta(\xi|\bar{C})$, were not sufficient to lead to significant errors in estimating posterior probabilities or in classification. Rather, we showed how, for a given schema, that the distribution of dependencies across C and \bar{C} is what controlled the accuracy of the NBA. We showed that the errors were maximized when the errors in the likelihoods for C and \bar{C} were of large magnitude and, crucially, of opposite sign. As a corresponding diagnostic we introduced $\Delta(\xi)$.

We showed that error analysis is simplest in terms of the score function, where independent errors had an additive nature. We also saw that although errors could cancel/reinforce “locally”, i.e., between the likelihoods for C and \bar{C} within the same schema, a full error analysis for a given feature set was a “global” question, where it was not possible to say whether the total error for a given set was large or not until all contributions had been calculated. We derived explicit formulas that related local errors, in potentially distinct feature subsets for each likelihood, to global errors over the full feature set.

To relate the error analysis to model performance we considered distinct performance metrics: i) classification accuracy; ii) estimation of posterior probabilities $P(C|\mathbf{X})$; iii) the distance between the relative rankings of the NBA and GBA; and iv) AUC. Of course, the actual distribution and impact of attribute correlations on these performance metrics depends on the precise properties of the underlying correlation structure of the probability distributions we are trying to estimate. As real-world distributions are associated with finite, and very often small, samples, sampling errors play an important distinguishing role between

the NBA and GBA. As our interest here is in understanding the role of feature correlations as the cause of model bias however, we chose to restrict attention initially to a set of artificial, pre-specified probability distributions, where the degree and type of feature correlation could be chosen and tuned to illustrate the effect and impact of feature correlations. Specifically, we proposed a set of 12 probability distributions for two, binary features that illustrated different qualitative characteristics, that we believe give substantial insight into the inner workings of the NBA and its generalizations. Essentially, the distributions capture the notions of strong versus weak correlations and correlations that reinforce or cancel between the likelihoods.

We examined correlations in detail for all 12 test distributions, showing how significant error in the NBA depended crucially on the relative signs of the errors in the likelihoods for C and \bar{C} ; error reinforcement and error cancelation being associated with opposite and equal signs in $\delta(\xi|C)$ and $\delta(\xi|\bar{C})$ respectively. We examined the impact of the different correlation structures on model performance and showed that our diagnostics correlated well with model performance. In other words, the larger the error according to our diagnostics the worse the performance of the NBA. This validates the diagnostics as potential predictors of performance of the NBA. In considering a class of generalizations of the NBA, where the likelihoods are not maximally factorized, we saw the impact of choosing a factorization that did, versus did not, respect any underlying correlation structure in the context of a set of 3-feature probability distributions derived from our set of 12 two-feature distributions. We showed that a GBA factorization that captured the correlation structure of the underlying problem inevitably led to better performance. We also saw the impact of correlations that were not symmetrically distributed between the likelihoods of C and \bar{C} , i.e., that involved different feature combinations. Further, we showed that our error diagnostics correlated well with the underlying problem structure, thereby indicating which factorization was optimal.

At the level of two and three features, only intra-schemata error cancelations are visible. To investigate the role of inter-schemata cancelations we extended our analysis to four-, six-, and eight-feature distributions which were concatenations of our original two-feature distributions. We studied a set of 19 distinct, concatenated distributions with different numbers of features and correlation structures, showing how the full global error was an emergent property, resulting from a set of cancelations and reinforcements of the local errors at both the intra- and inter-schemata level. In particular, we saw the impact of the fact that the errors of any given schema had varying signs across distinct feature values, thereby guaranteeing the existence of inter-schemata error cancelations. We showed that maximal error cancelation occurred in probability distributions that exhibited both intra- and inter-schemata cancelations. We then showed that model performance was highly correlated with our global error functions, thereby validating, once again, their value as predictive diagnostics for the relative performance of the NBA. We also saw that different performance metrics were more sensitive to different characteristics of the error distribution, with classification error being particularly sensitive to errors in the likelihoods close to the NB score threshold but insensitive to those far from the threshold. On the contrary, our distance metric as a global ranking measure was equally affected by likelihood errors independent of their distance from the threshold.

We then applied our formalism to a representative set of 20 real world problems taken from the UCI repository and a further three text mining data sets. Here we arrived at two important conclusions: first that our error diagnostics allow for the identification of sets of correlated features that should be combined; and, secondly, that the combined features can be used to construct a GBC and GBA in the framework of the Semi-naive Bayes approximation.

We showed that the performance of the resulting GBC gave significant improvements over the NBC in terms of classifier error but that the differences in ranking performance, as measured by the AUC, were not statistically significant. This is in line with previous results. Importantly, we saw that our diagnostics also served to indicate a priori which problems were more likely to result in an enhanced performance using a GBC as opposed to the NBC. It should be emphasised that our GBC and GBA is not “optimized” in that we have not attempted to maximise the performance of the GBC by tuning any associated parameters. There are several areas where improvements could potentially be made: one is in the threshold used for ε for combining features.

In summary: we have proposed and tested a set of error diagnostics for detecting and quantifying the effect of feature correlations at both the local (subsets of features—schemata) and global (the full feature set) level. By interpolating between the local and global levels they allow for a full understanding of how errors cancel across different feature combinations. Thus, one can not only predict the potential performance of the NBA but can also determine which feature subsets should be combined in a generalization of the NBA. The optimal factorization for an instance of the GBA should be that which best respects the underlying correlation structure of the problem at hand. Our diagnostics are an aid in the search for that correlation structure. Obviously, in a real world problem that structure must be inferred from finite samples and therefore is subject to sampling error. Our emphasis here was on the model bias associated with the NBA versus the GBA or the exact correlation structure. However, we showed the potential for this approach by showing how a GBC could be simply generated by using statistical hypothesis testing on our error measures to determine which feature sets to combine and that the resulting classifier led to significant performance improvements on a substantive set of UCI data sets. We believe that our results are also a first step in the direction of designing prediction meta-algorithms that can provide an a priori prediction of the performance of a given algorithm on a given problem.

Acknowledgements HFH is grateful to CONACyT for financial support. AR is grateful for support from the Instituto Tecnológico de Minatitlán. This work has been supported by DGAPA, UNAM Project I113414.

Appendix A

Below, in Table 9, we show the two-feature probability distributions for the likelihoods $P(X_1X_2|C)$ and $P(X_1X_2|\bar{C})$. The first column, Dis., denotes the distribution, 0–12, with 0 denoting the parity function, second, Conf., the class and feature configuration CX_1X_2 , with $C = 1$, $\bar{C} = 0$ and $X_i = 0, 1$; the third, Lik., the corresponding likelihood for the class/feature combination, the fourth and fifth, $(\Delta_C + 1)$ and $(\Delta_{\bar{C}} + 1)$, the corresponding error functions for the likelihoods, where Δ_C is given by Eq. (5.1); the sixth column is the error function, Δ (5.3); the seventh column, Post., is the exact posterior probability and the eighth the NBA to the posterior probability, with column 9 being the percentage difference between them; column 10, SNB, is the score in the NBA and column 11, SGNB, the score in the GBA (the exact score in this case), and, finally, column 12 is the percentage difference between them. Also shown at the end of each distribution is the mean absolute error for the NBA estimates of the posterior probability and the score function.

Table 9 Characteristics of the 12 artificial probability distributions (1-12) and the parity function (0) used in the analysis

Dis	Conf	Lik	δ_C	$\delta_{\bar{C}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
0	111	0.00	0	2	-2	0	0.5	-100	0	$-\infty$	$-\infty$
0	110	0.50	2	0	2	1	0.5	100	0	∞	∞
0	101	0.50	2	0	2	1	0.5	100	0	∞	∞
0	100	0.00	0	2	-2	0	0.5	-100	0	$-\infty$	$-\infty$
0	011	0.50	2	0	2	1	0.5	100	0	∞	∞
0	010	0.00	0	2	-2	0	0.5	-100	0	$-\infty$	$-\infty$
0	001	0.00	0	2	-2	0	0.5	-100	0	$-\infty$	$-\infty$
0	000	0.50	2	0	2	1	0.5	100	0	∞	∞
1	111	0.01	0.043	1.636	-3.641	0.03	0.56	100.00	Mean abs error	∞	∞
1	110	0.456	1.957	0.222	2.175	0.93	0.61	-94.24	0.24	-3.40	2196.87
1	101	0.49	1.835	0.363	1.618	0.88	0.59	53.18	0.44	2.62	6226.83
1	100	0.044	0.164	1.777	-2.378	0.14	0.64	48.46	0.38	1.99	-5483.14
1	011	0.45	1.636	0.042	3.640	0.97	0.44	-77.88	0.58	-1.80	-1389.69
1	010	0.05	0.222	1.957	-2.175	0.07	0.39	119.77	-0.24	3.40	2196.87
1	001	0.1	0.363	1.835	-1.618	0.12	0.41	-82.61	-0.44	-2.62	6226.83
1	000	0.4	1.777	0.164	2.378	0.86	0.36	-70.58	-0.38	-1.99	-5483.14
2	111	0.01	0.042	0.363	-2.136	0.13	0.56	138.63	-0.58	1.80	-1389.69
2	110	0.456	1.957	1.777	0.096	0.63	0.61	85.67	Mean abs error	3824.13	3824.13
2	101	0.49	1.835	1.636	0.114	0.62	0.59	-76.69	0.24	-1.90	1289.34
2	100	0.044	0.164	0.222	-0.298	0.57	0.64	3.72	0.44	0.54	275.03
2	011	0.1	0.363	0.042	2.136	0.87	0.44	4.61	0.38	0.49	-388.45
2	010	0.4	1.777	1.957	-0.096	0.37	0.39	-11.14	0.58	0.28	-174.69
								97.47	-0.24	1.90	1289.34
								-5.78	-0.44	-0.54	275.03

Table 9 continued

Dis	Conf	Lik	δ_C	$\delta_{\bar{C}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
2	001	0.45	1.636	1.835	-0.114	0.38	0.41	-6.72	-0.38	-0.49	-388.45
2	000	0.05	0.222	0.164	0.298	0.43	0.36	19.83	-0.58	-0.28	-174.69
						Mean abs error		28.25	Mean abs error		531.88
3	111	0.01	0.0429	1.777	-3.723	0.04	0.61	-94.06	0.44	-3.28	-10,658.34
3	110	0.456	1.957	0.363	1.683	0.87	0.56	55.89	0.24	1.92	-1015.52
3	101	0.49	1.835	0.222	2.111	0.94	0.64	46.23	0.58	2.69	1233.57
3	100	0.044	0.164	1.636	-2.295	0.13	0.59	-78.43	0.38	-1.92	7775.56
3	011	0.4	1.777	0.042	3.723	0.96	0.39	146.10	-0.44	3.28	-10,658.34
3	010	0.1	0.363	1.957	-1.683	0.13	0.44	-71.03	-0.24	-1.92	-1015.52
3	001	0.05	0.222	1.835	-2.111	0.06	0.36	-82.29	-0.58	-2.69	1233.57
3	000	0.45	1.636	0.164	2.295	0.87	0.41	114.22	-0.38	1.92	7775.56
						Mean abs error		86.03	Mean abs error		5170.75
4	111	0.2	1.045	0.826	0.234	0.91	0.89	2.41	2.07	2.30	14.13
4	110	0.015	0.634	1.021	-0.476	0.07	0.11	-35.26	-2.11	-2.59	18.92
4	101	0.69	0.987	1.085	-0.094	0.93	0.93	-0.65	2.65	2.56	-4.20
4	100	0.095	1.100	0.989	0.106	0.19	0.18	9.01	-1.53	-1.42	-5.49
4	011	0.03	0.826	1.045	-0.234	0.09	0.11	-19.03	-2.07	-2.30	14.13
4	010	0.3	1.021	0.634	0.476	0.93	0.89	4.26	2.11	2.59	18.92
4	001	0.08	1.085	0.987	0.094	0.07	0.07	9.20	-2.65	-2.56	-4.20
4	000	0.59	0.989	1.100	-0.106	0.81	0.82	-1.96	1.53	1.42	-5.49
						Mean abs error		10.22	Mean abs error		10.68
5	111	0.53	1.092	0.595	0.607	0.89	0.81	9.34	1.47	2.07	57.27
5	110	0.24	0.842	1.127	-0.291	0.38	0.45	-15.82	-0.22	-0.51	46.72
5	101	0.1	0.690	1.944	-1.035	0.52	0.75	-31.14	1.10	0.07	-148.11

Table 9 continued

Dis	Conf	Lik	δ_C	$\delta_{\bar{c}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
5	100	0.13	1.527	0.701	0.777	0.55	0.36	53.04	-0.58	0.20	-78.93
5	011	0.1	0.595	1.092	-0.607	0.11	0.19	-40.43	-1.47	-2.07	57.27
5	010	0.6	1.127	0.842	0.291	0.63	0.55	12.71	0.22	0.51	46.72
5	001	0.14	1.944	0.690	1.035	0.48	0.25	94.01	-1.10	-0.07	-148.11
5	000	0.16	0.701	1.527	-0.777	0.45	0.64	-29.70	0.58	-0.20	-78.93
						Mean abs error		35.77	Mean abs error		82.76
6	111	0.15	0.659	1.183	-0.584	0.31	0.45	-30.54	-0.21	-0.80	94.49
6	110	0.5	1.183	0.659	0.584	0.83	0.74	13.25	1.02	1.61	94.49
6	101	0.2	1.632	0.659	0.906	0.67	0.45	49.21	-0.21	0.69	-146.47
6	100	0.15	0.659	1.632	-0.906	0.53	0.74	-28.05	1.02	0.12	-146.47
6	011	0.5	1.183	0.659	0.584	0.69	0.55	24.67	0.21	0.80	94.49
6	010	0.15	0.659	1.183	-0.584	0.17	0.26	-36.90	-1.02	-1.61	94.49
6	001	0.15	0.659	1.632	-0.906	0.33	0.55	-39.74	0.21	-0.69	-146.47
6	000	0.2	1.632	0.659	0.906	0.47	0.26	78.15	-1.02	-0.12	-146.47
						Mean abs error		37.56	Mean abs error		120.48
7	111	0.08	0.314	0.363	-0.146	0.55	0.58	-6.18	0.33	0.18	189.51
7	110	0.3	2.392	1.777	0.296	0.53	0.46	16.27	-0.18	0.12	-50.79
7	101	0.59	1.420	1.636	-0.141	0.66	0.69	-4.45	0.82	0.68	-34.33
7	100	0.03	0.146	0.222	-0.415	0.47	0.58	-17.90	0.31	-0.11	437.46
7	011	0.1	0.363	0.314	0.146	0.45	0.42	8.58	-0.33	-0.18	189.51
7	010	0.4	1.777	2.392	-0.296	0.47	0.54	-13.60	0.18	-0.12	-50.79
7	001	0.45	1.636	1.420	0.141	0.34	0.31	10.08	-0.82	-0.68	-34.33
7	000	0.05	0.222	0.146	0.415	0.53	0.42	24.42	-0.31	0.11	437.46
						Mean abs error		12.69	Mean abs error		178.02

Table 9 continued

Dis	Conf	Lik	δ_C	$\delta_{\tilde{c}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
8	111	0.1	0.363	1.069	-1.078	0.20	0.43	-52.61	-0.29	-1.37	154.87
8	110	0.4	1.777	0.887	0.695	0.67	0.50	33.47	0.00	0.69	-170.59
8	101	0.45	1.636	0.439	1.313	0.96	0.86	11.57	1.80	3.11	94.22
8	100	0.05	0.222	1.913	-2.153	0.48	0.89	-45.62	2.09	-0.06	-127.92
8	011	0.59	1.069	0.363	1.078	0.80	0.57	39.33	0.29	1.37	154.87
8	010	0.3	0.887	1.777	-0.695	0.33	0.50	-33.40	0.00	-0.69	-170.59
8	001	0.03	0.439	1.636	-1.313	0.04	0.14	-70.01	-1.80	-3.11	94.22
8	000	0.08	1.913	0.222	2.153	0.52	0.11	368.34	-2.09	0.06	-127.92
						Mean abs error		81.80	Mean abs error		136.90
9	111	0.1	0.363	0.146	0.908	0.83	0.67	24.67	0.70	1.61	307.14
9	110	0.4	1.777	1.420	0.224	0.50	0.45	12.48	-0.21	0.02	-36.61
9	101	0.45	1.636	2.392	-0.379	0.69	0.77	-9.72	1.19	0.81	-48.37
9	100	0.05	0.222	0.314	-0.346	0.48	0.57	-15.11	0.28	-0.06	280.28
9	011	0.03	0.146	0.363	-0.908	0.17	0.33	-49.73	-0.70	-1.61	307.14
9	010	0.59	1.420	1.777	-0.224	0.50	0.55	-10.14	0.21	-0.02	-36.61
9	001	0.3	2.392	1.636	0.379	0.31	0.23	31.98	-1.19	-0.81	-48.37
9	000	0.08	0.314	0.222	0.346	0.52	0.43	20.03	-0.28	0.06	280.28
						Mean abs error		21.73	Mean abs error		168.10
10	111	0.1	0.363	2.392	-1.883	0.33	0.77	-56.53	1.19	-0.69	-239.90
10	110	0.4	1.777	0.146	2.495	0.95	0.62	52.97	0.50	3.00	2625.35
10	101	0.45	1.636	0.314	1.650	0.89	0.62	44.59	0.48	2.13	2140.89
10	100	0.05	0.222	1.420	-1.854	0.11	0.45	-74.84	-0.21	-2.06	302.53
10	011	0.3	2.392	0.363	1.883	0.67	0.23	185.96	-1.19	0.69	-239.90

Table 9 continued

Dis	Conf	Lik	δ_C	$\delta_{\bar{C}}$	Δ	Post	PostNB	% diff	SNB	SGNB	% diff
10	010	0.03	0.146	1.777	-2.495	0.05	0.38	-87.38	-0.50	-3.00	2625.35
10	001	0.08	0.314	1.636	-1.650	0.11	0.38	-72.24	-0.48	-2.13	2140.89
10	000	0.59	1.420	0.222	1.854	0.89	0.55	60.81	0.21	2.06	302.53
						Mean abs error		79.42	Mean abs error		1327.17
11	111	0.02	0.078	1.816	-3.142	0.06	0.58	-90.37	0.31	-2.83	3260.35
11	110	0.48	1.959	0.115	2.828	0.96	0.59	63.71	0.35	3.18	-5021.02
11	101	0.49	1.921	0.0418	3.827	0.99	0.62	60.35	0.47	4.30	5984.44
11	100	0.01	0.040	2.038	-3.910	0.03	0.62	-94.84	0.51	-3.40	-3760.22
11	011	0.51	1.816	0.078	3.142	0.94	0.42	123.09	-0.31	2.83	3260.35
11	010	0.03	0.115	1.959	-2.828	0.04	0.41	-90.33	-0.35	-3.18	-5021.02
11	001	0.01	0.041	1.921	-3.827	0.01	0.38	-96.51	-0.47	-4.30	5984.44
11	000	0.45	2.038	0.040	3.910	0.97	0.38	157.85	-0.51	3.40	-3760.22
						Mean abs error		79.42	Mean abs error		4506.51
12	111	0.5	1.814	1.957	-0.075	0.62	0.64	-2.76	0.57	0.50	-45.14
12	110	0.02	0.081	0.164	-0.700	0.41	0.58	-29.93	0.32	-0.38	791.49
12	101	0.03	0.117	0.0429	1.010	0.82	0.62	31.78	0.49	1.50	1150.28
12	100	0.45	1.994	1.835	0.083	0.58	0.56	3.65	0.24	0.32	-49.46
12	011	0.456	1.957	1.814	0.075	0.38	0.36	4.90	-0.57	-0.50	-45.14
12	010	0.044	0.164	0.081	0.700	0.59	0.42	41.10	-0.32	0.38	791.49
12	001	0.01	0.042	0.117	-1.010	0.18	0.38	-52.04	-0.49	-1.50	1150.28
12	000	0.49	1.835	1.994	-0.083	0.42	0.44	-4.63	-0.24	-0.32	-49.46
						Mean abs error		21.35	Mean abs error		509.09

References

- Bennett, P. N. (2000). *Assessing the calibration of Naive Bayes' posterior estimates*. Technical report no. CMU-CS00-155.
- Bermejo, P., Gámez, J. A., & Puerta, J. M. (2014). Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. *Knowledge-Based Systems*, 55, 140–147.
- Broos, P. S., Getman, K. V., Povich, M. S., Townsley, L. K., Feigelson, E. D., & Garmire, G. P. (2011). A naive Bayes source classifier for X-ray sources. *The Astrophysical Journal Supplement Series*, 194(1), 4.
- Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.
- Domingos, P., & Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In *Proceedings of the thirteenth international conference on machine learning* (pp. 105–112). Morgan Kaufmann.
- Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4), 1937–1946.
- Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Naive Bayes for regression. *Machine Learning*, 41(1), 5–15.
- Friedman, J. (1997). On bias, variance, 0/1–Loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1, 55–77.
- Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29(2), 131–163.
- Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press. <http://sci2s.ugr.es/keel/textClassification.php>
- Keogh, E., & Pazzani, M. (1999). Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. In *Proceedings of the international workshop on artificial intelligence and statistics* (pp. 225–230).
- Kindermann, R., & Snell, J. L. (1980). *Markov random fields and their applications*. Providence: American Mathematical Society.
- Kohavi, R. (1996). Scaling up the accuracy of naive Bayes classifiers: A decision-tree hybrid. In *Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining (KDD-96)*, Portland, OR (pp. 202–207).
- Kononenko, I. (1991). Semi-naive Bayesian classifier. In *Proceedings of the sixth European working session on learning* (pp. 206–219). Berlin: Springer.
- Langley, P. (1993). Induction of recursive Bayesian classifiers. In *Proceedings of the 1993 European conference on machine learning* (pp. 153–164). Berlin: Springer
- Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Proceedings of the tenth conference on uncertainty in artificial intelligence* (pp. 399–406). Morgan Kaufmann.
- Liangxiao, J., Zhang, H., & Cai, Z. (2009). A novel Bayes model: Hidden naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, 21(10), 1361.
- Ling, C. X., Huang, J., & Zhang, H. (2003) AUC: A statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th international joint conference on artificial intelligence* (pp. 519–524).
- Lowd, D., & Domingos, P. (2005). Naive Bayes models for probability estimation. In *ICML '05 proceedings of the 22nd international conference on machine learning* (pp. 529–536). New York, NY: ACM.
- Mohamad, N. A., Jusoh, N. A., Htike, Z. Z., & Win, S. L. (2014). Bacteria identification from microscopic morphology using naive Bayes. *International Journal of Computer Science, Engineering and Information Technology (IJCEIT)*, 4(1).
- Monti, S., & Cooper, G. F. (1999). A Bayesian network classifier that combines a finite mixture model and a Naive Bayes model. In *Proceedings of the 15th conference on uncertainty in artificial intelligence* (pp. 447–456). Morgan Kaufmann.
- Ng, S. S. Y., Xing, Y., & Tsui, K. L. (2014). A naive Bayes model for robust remaining useful life prediction of lithium-ion battery. *Applied Energy*, 118, 114–123.
- Panda, M., & Patra, M. R. (2007). Network intrusion detection using naive Bayes. *International journal of computer science and network security*, 7(12), 258–263.
- Pazzani, M. J. (1996). Constructive induction of Cartesian product attributes. In *ISIS: information, statistics and induction in science* (pp. 66–77). Singapore: World Scientific.
- Poli, R., & Stephens, C. R. (2014). Taming the complexity of natural and artificial evolutionary dynamics. In S. Cagnoni, M. Mirolli, & M. Villani (Eds.), *Evolution, complexity and artificial life* (pp. 19–39). Berlin: Springer.

- Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41–46).
- Sahami, M. (1996). Learning limited dependence Bayesian classifiers. In *Proceedings of the second international conference on knowledge discovery and data mining* (pp. 334–338). Menlo Park, CA: AAAI Press.
- Singh, M., & Provan, G. M. (1996). Efficient learning of selective Bayesian network classifiers. In *Proceedings of the thirteenth international conference on machine learning* (pp. 453–461). San Francisco: Morgan Kaufmann.
- Stephens, C. R., Waelbroeck, H., & Talley, S. (2005, June). Predicting healthcare costs using GAs. In *Proceedings of the 2005 workshops on genetic and evolutionary computation* (pp. 159–163). ACM.
- Turhan, B., & Bener, A. (2009). Analysis of Naive Bayes' assumptions on software fault data: An empirical study. *Data & Knowledge Engineering*, 68(2), 278–290.
- Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16), 5261–5267.
- Webb, G. I. (2001). Candidate elimination criteria for lazy Bayesian rules. In *Proceedings of the fourteenth Australian joint conference on artificial intelligence* (pp. 545–556). Berlin: Springer.
- Webb, G. I., Boughton, J., & Wang, Z. (2005). Not so naive Bayes: Aggregating one-dependence estimators. *Machine Learning*, 58, 5–24.
- Webb, G. I., Boughton, J., Zheng, F., Ting, K. M., & Salem, H. (2012). Learning by extrapolation from marginal to full-multivariate probability distributions: Decreasingly naive Bayesian classification. *Machine Learning*, 86(2), 233–272.
- Webb, G. I., & Pazzani, M. J. (1998). Adjusted probability naive Bayesian induction. In *Proceedings of the eleventh Australian joint conference on artificial intelligence* (pp. 285–295). Berlin: Springer.
- Wei, W., Visweswaran, S., & Cooper, G. F. (2011). The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*, 18(4), 370–375.
- Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8, 1341–1390.
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1, 67.
- Xie, Z., Hsu, W., Liu, Z., & Lee, M. L. (2002). SNNB: A selective neighborhood based naive Bayes for lazy learning. In M.-S. Chen, P. S., Yu & B. Liu (Eds.), *Advances in knowledge discovery and data mining, proceedings PAKDD 2002* (pp. 104–114). Berlin: Springer.
- Zhang, H. (2004). The optimality of naive Bayes. In *Proceedings of the FLAIRS conference* (Vol. 1, No. 2, pp. 3–9).
- Zhang, H., & Ling, C. X. (2003). AI 2003. In Y. Xiang & B. Chaib-draa (Eds.), *LNAI* (Vol. 2671, pp. 591–595). Berlin: Springer.
- Zhang, H., & Su, J. (2004). Naive Bayesian classifiers for ranking. In J.-F. Boulicaut, et al. (Eds.), *ECML 2004, LNAI 3201* (pp. 501–512). Berlin: Springer
- Zhang, H., & Su, J. (2008). Naive Bayes for optimal ranking. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(2), 79–93.
- Zheng, Z., & Webb, G. I. (2000). Lazy learning of Bayesian rules. *Machine Learning*, 41(1), 53–84.
- Zheng, Z., Webb, G. I., & Ting, K. M. (1999). Lazy Bayesian rules: A lazy semi-naive Bayesian learning technique competitive to boosting decision trees. In *Proceedings of the sixteenth international conference on machine learning (ICML-99)* (pp. 493–502). Morgan Kaufmann.