RC 21993 (98805) 14 March 2001 **Computer Science**

IBM Research Report

An Analysis of Data Characteristics that Affect Naive Bayes Performance

Irina Rish, Joseph Hellerstein, Jayram Thathachar
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - T. J. Watson - Tokyo - Zurich

An analysis of data characteristics that affect naive Bayes performance

Irina Rish RISH@US.IBM.COM
Joseph Hellerstein HELLERS@US.IBM.COM
Jayram Thathachar JAYRAM@US.IBM.COM
IBM T.J. Watson Research Center 30 Saw Mill River Road, Hawthorne, NY 10532

Abstract

Despite its unrealistic independence assumption, the naive Bayes classifier is remarkably successful in practice. This paper identifies some data characteristics for which naive Bayes works well, such as certain deterministic and almost-deterministic dependencies (i.e., low-entropy distributions). First, we address zero-Bayes-risk problems, proving naive Bayes optimality for any two-class concept that assigns class 0 to exactly one example (i.e. $H(P(x_i|0)) = 0$). We demonstrate empirically that the entropy of $P(x_i|0)$ is a better predictor of the naive Bayes error than the class-conditional mutual information between features. Next, we consider a broader class of non-zero Bayes risk problems, further pursuing the study of low-entropy distributions. We derive error bounds for approximating the joint distribution by the product of marginals in case of nearly-deterministic class-conditional feature distributions $P(x_i|C)$, and we demonstrate how the performance of naive Bayes improves with decreasing entropy of such distributions. Finally, we consider functional dependencies between features and prove naive Bayes optimality in certain cases. Using Monte Carlo simulations, we show that naive Bayes works best in two cases: completely independent features (as expected by the assumptions made) and functionally dependent features (which is surprising). Naive Bayes has its worst performance between these extremes.

1. Introduction

Classification has been the subject of much research in the machine learning community. One well-established approach is Bayesian classification, a technique that has become increasingly popular in the recent years in part due to recent developments in learning with Bayesian belief networks (Heckerman, 1995; Friedman et al., 1997). The simplest Bayesian classifier is the widely used *naive Bayes* classifier. It greatly simplifies learning by assuming that features are independent given class, that is, $P(\mathbf{x}, c) = \prod_{i=1}^{n} P(x_i|c)$, where $\mathbf{x} = (x_1, \cdots, x_n)$ is a feature vector and c is a class. Although feature independence is generally a poor assumption, naive Bayes is surprisingly successful in practice (Langley et al., 1992; Domingos & Pazzani, 1997; Mitchell, 1997; Hellerstein et al., 2000). Naive Bayes has proven effective in text classification, medical diagnosis, and computer performance management, among many other applications.

Why does naive Bayes often work well even though its independence assumption is violated? A central observation is the following: optimality in terms of zero-one loss (classification error) is not necessarily related to the quality of the fit to a probability distribution (i.e., the appropriateness of the independence assumption). Rather, an optimal classifier is obtained as long as both the actual and estimated distributions agree on the most-probable class (Domingos & Pazzani, 1997). For example, (Domingos & Pazzani, 1997) prove naive Bayes optimality for some problems classes that have a high degree of feature dependencies, such as disjunctive and conjunctive concepts.

Herein, we probe further into the data characteristics that make naive Bayes work well. For zero-Bayes-risk problems, we prove naive Bayes optimality for any two-class concept with nominal features where only one example has class 0 (or class 1), thus generalizing the results for conjunctive and disjunctive concepts. Then, using Monte-Carlo simulation, we study the behavior of naive Bayes for increasing prior $P(0)$, observing that the entropy of class-conditional marginals, $H(P(X_i|C)$, is usually a better indicator of naive Bayes performance than the class-conditional mutual information between the features, $I(X_1, X_2|C)$.

Next, we consider a broader class of non-zero Bayes risk problems, focusing on deterministic or close-to-deterministic dependencies. We prove that naive Bayes is optimal in certain cases of functionally dependent features. Then we relax those dependencies by adding noise, demon-

strating empirically that naive Bayes reaches optimal performance in two extreme cases: completely independent features (as expected by the assumptions made) and functionally dependent features (which is surprising). Naive Bayes has its worst performance between these extremes. We also show that a joint distribution and its approximation by the product of marginals converge with decreasing entropy of the distribution (i.e., $P(a_1, ..., a_n) \approx \prod_{i=1}^{n} P(a_i)$ for low-entropy distributions), and demonstrate how decreasing entropy of class-conditional feature distributions affects the error of naive Bayes classifier.

Note, that our error analysis only focuses on the *bias* of naive Bayes classifier, not on its *variance*, i.e. we assume an infinite amount of data, or perfect knowledge of data distribution to be available and compare naive Bayes versus Bayes-optimal classifier.

Although it may seem counterintuitive, the naive Bayes error is not correlated with class-conditional mutual information between the features. This phenomenon is consistently demonstrated by our simulation for different problem classes, including both zero- and non-zero Bayes risk problems. Our results support previous observations on UCI benchmark problems that also reveal low correlation between the degree of feature dependence and the performance of naive Bayes (Domingos & Pazzani, 1997)). This motivates a search for other metrics such as entropy of class-conditional marginal feature distributions.

Our study is also motivated by significant amount of empirical evidence suggesting that approximate probabilistic inference algorithms that make independence assumptions often find accurate most-likely variable assignment on problems involving nearly-deterministic dependencies. One of the most prominent examples is successful application of Pearl's belief propagation algorithm to probabilistic decoding (Frey & MacKay, 1998): although belief propagation performs local inference ignoring long-range dependencies, its iterative variant applied to certain coding networks results into lower error rates than the state-of-the-art decoding algorithms. Another example of local inference algorithm that partially ignores dependencies is the mini-bucket approach (Dechter, 1997). When applied to finding most probable states, it performs significantly better on problems with lower "noise" (Rish, 1999).

2. Definitions and Background

Let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of observed random variables, called *features*, where each feature takes values from its *domain* D_i. A feature vector is also called an *example*, or a *state* (of nature), and the set of all possible examples, or the *state space*, is denoted $\Omega = D_1 \times ... \times D_n$. Let C be an unobserved random variable denoting *class* of an example, where C can take one of m values $c \in \{0, ..., m-1\}$. Capital letters, such as X_i, will denote variables, while lower-case letters, such as x_i, will denote their values; boldface letters will denote vectors. Also, we will sometimes use shorter notation $P(\mathbf{x})$ and $P(x_1, ..., x_n)$ instead of $P(\mathbf{X} = \mathbf{x})$ and $P(X_1 = x_1, ..., X_n = x_n)$, respectively, as well as $P(i)$ and $P(\mathbf{x}|i)$ instead of $P(C = i)$ and $P(\mathbf{x}|C = i)$, respectively. Also, we may denote $P(A|C = i)$ as $P_i(A)$.

A function $g : \Omega \rightarrow \{0, ..., m-1\}$, where $g(\mathbf{x}) = C$, denotes a *concept* to be learned. In general, $g(x)$ is a random function, and a concept is called noisy. In the absence of noise, $g(x)$ is deterministic, i.e. it always assigns same class to a given example (e.g., disjunctive and conjunctive concepts are deterministic).

A classifier is defined by a (deterministic) function $h : \Omega \rightarrow \{0, ..., m-1\}$ (a *hypothesis*) that assigns a class to any given example. A common approach is to associate each class i with a discriminant function $f_i(\mathbf{x})$, $i = 0, ..., m-1$, and let the classifier select the class with maximum discriminant function on a given example[1]:

$$h(\mathbf{x}) = \arg \max_{i \in \{0,...,m-1\}} f_i(\mathbf{x}). \quad (1)$$

The *Bayes* classifier $h^*(\mathbf{x})$ (that we also call *Bayes-optimal* classifier and denote $BO(\mathbf{x})$), uses as discriminant functions the class posterior probabilities given a feature vector, i.e. $f_i^*(\mathbf{x}) = P(C = i|\mathbf{X} = \mathbf{x})$. Applying the Bayes rule gives $P(C = i|\mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}|C=i)P(C=i)}{P(\mathbf{X}=\mathbf{x})}$, where $P(\mathbf{X} = \mathbf{x})$ is same for all classes, and therefore can be ignored. This yields Bayes discriminant functions

$$f_i^*(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|C = i)P(C = i), \quad (2)$$

where $P(\mathbf{X} = \mathbf{x}|C = i)$ is called the *class-conditional probability distribution (CPD)*. Thus, the Bayes classifier

$$h^*(\mathbf{x}) = \arg \max_i P(\mathbf{X} = \mathbf{x}|C = i)P(C = i) \quad (3)$$

finds the *maximum a posteriori probability* (MAP) hypothesis given example $\mathbf{x}$. However, direct estimation of $P(\mathbf{X} = \mathbf{x}|C = i)$ from a given set of training examples is hard when feature space is high-dimensional. Therefore, approximations are commonly used, such as using the simplifying assumption that features are independent given the class. This yields *naive Bayes* classifier $NB(\mathbf{x})$ defined by discriminant functions

$$f_i^{NB}(\mathbf{x}) = \Pi_{j=1}^{n} P(X_j = x_j|C = i)P(C = i). \quad (4)$$

[1] Clearly, discriminant functions are not unique, since classifier does not change if a monotone function (e.g., $\log$) is applied to all f_i's.

The probability of classification error, or $risk$, of a classifier h is defined as

$$R(h) = P(h(\mathbf{X}) \neq C) = \sum_{\mathbf{x}\in\Omega} P(h(\mathbf{x}) \neq C)P(\mathbf{X} = \mathbf{x}) = \\ = E_{\mathbf{x}}\{P(h(\mathbf{x}) \neq C)\},$$

where $E_{\mathbf{x}}$ is the expectation over $\mathbf{x}$. $R^* = R(h^*)$ denotes the Bayes error (Bayes risk).

We say that classifier h is *optimal* on a given problem if its risk coincides with the Bayes risk. Assuming there is no noise (i.e. zero Bayes risk), a concept is called *separable* by a set of functions $S = \{f_c(x)|c = 0, ..., m-1\}$ if every example $\mathbf{x}$ is classified correctly when using each $f_c(x)$ as discriminant functions.

As a measure of dependence between two features X_k and X_j we use the class-conditional mutual information (Cover & Thomas, 1991), which can be defined as

$$I(X_k; X_j|C) = H(X_k|C) + H(X_j|C) - H(X_k, X_j|C),$$

where $H(A|C)$ is the class-conditional entropy of A, defined as:

$$-\sum_i P(C = i) \sum_t P(A = t|C = i) \log P(A = t|C = i).$$

Mutual information $I(X_k; X_j|C)$ can be also expressed as the class-conditional *KL-divergence* between the joint distribution $P(X_k, X_j|C)$ and the product of marginals $P(X_k|C)P(X_j|C)$ (Cover & Thomas, 1991):

$$\sum_i P(C = i) \sum_{X_k=x_k, X_j=x_j} P(X_k = x_k, X_j = x_j|C = i) \\ \times \log \frac{P(X_k = x_k, X_j = x_j|C = i)}{P(X_k = x_k|C = i)P(X_j = x_j|C = i)},$$

which is zero when X_k and X_j are mutually independent given class C, and increases with increasing level of dependence, reaching the maximum when one feature is a deterministic function of the other.

3. Naive Bayes on zero-Bayes-risk problems

In this section, we will focus on concepts without noise, namely, concepts where $P(C = i|\mathbf{x}) = 0$ or 1 for any $\mathbf{x}$ and i (i.e. an example always has the same class), and thus Bayes risk is zero. We assume finite-domain, or *nominal* features, where i-th feature has k_i values. Clearly, nominal features can be transformed into numeric ones by imposing an order on their domains.) We also assume there are only two classes, $C = 0$ and $C = 1$.

It is well known that for binary features ($k_i = 2$ for all $i = 1, ..., n$) naive Bayes is a linear classifier (Duda & Hart, 1973), i.e. its discriminant functions are linear functions of features, and thus it is suboptimal for non-linearly separable concepts. For example, naive Bayes is not optimal for m-of-n concept (Kohavi, 1995; Domingos & Pazzani, 1997). However, as shown in (Domingos & Pazzani, 1997), naive Bayes is optimal disjunctive and conjunctive concepts.

When $k_i > 2$ for some features, naive Bayes yields polynomial discriminant functions (Duda & Hart, 1973). Polynomial separability of such concepts is therefore a necessary condition of naive Bayes optimality. While characterizing sufficient optimality conditions may be hard in general, we can easily identify some specific problem classes. For example, a generalization of disjunctive and conjunctive concepts to concepts with any nominal features yields the following result:

Theorem 1 *Naive Bayes classifier is optimal for any two-class concept with nominal features that assigns class 0 to exactly one example, and class 1 to the other examples, with probability 1.*

Proof. We assume that n is the number of features, k_i is the number of values of i-th feature, and $\mathbf{x}^0 = (x_1^0, ..., x_n^0)$ is the example assigned class 0. We denote by $N = \prod_{i=1}^n k_i$ the total number of examples, and by $N_i = N/k_i$ the total number of examples having a fixed value of i-th feature. Note that $P(C = 0) = 1/N$, $P(C = 1) = (N-1)/N$, $P(x_i|0) = 1$ if $x_i = x_0^*$ and 0 otherwise, $P(x_i|1) = \frac{N_i-1}{N-1}$ if $x_i = x_0^*$ and $\frac{N_i}{N-1}$ otherwise. Note that the naive Bayes rule for selecting class 0 is

$$P(0) \prod_{i=1}^n P(x_i|0) > P(1) \prod_{i=1}^n P(x_i|1). \tag{5}$$

We now prove naive Bayes optimality for the two cases: $\mathbf{x} \neq \mathbf{x}^0$ (class 1) and $\mathbf{x} = \mathbf{x}^0$ (class 0).

Class 1: $\mathbf{x} \neq \mathbf{x}^0$. Clearly, $\exists i\ x_i \neq x_i^0$ and, therefore, $P(x_i|0) = 0$. On the other hand, $\forall i\ P(x_i|1) > 0$, thus the inequality 5 does not hold, and therefore naive Bayes assigns class 1 to $\mathbf{x}$, which is a correct decision.

Class 0: $\mathbf{x} = \mathbf{x}^0$. The left-hand-side of the inequality 5 yileds

$$P(0) \prod_{i=1}^n P(x_i^0|0) = \frac{1}{N-1},$$

while the right-hand side equals

$$P(1) \prod_{i=1}^n P(x_i^0|1) = \frac{N-1}{N} \prod_{i=1}^n \frac{N_i - 1}{N-1}.$$

This yields the decision rule $(N-1)^{n-1} > \prod_{i=1}^n (N_i - 1)$

for class 0, or, equivalently,

$$N^{n-1}(1-\frac{1}{N})^{n-1} > \prod_{i=1}^{n} N_i \prod_{i=1}^{n}(1-\frac{1}{N_i}). \quad (6)$$

Since $\prod_{i=1}^{n} N_i = \prod_{i=1}^{n} \frac{N}{k_i} = \frac{N^n}{\prod_{i=1}^{n} k_i} = N^{n-1}$, we get from 6

$$(1-\frac{1}{N})^{n-1} > \prod_{i=1}^{n}(1-\frac{1}{N_i}). \quad (7)$$

Note that $N > N_i$ and $1-\frac{1}{N_i} < 1$ for any i, so that $1-\frac{1}{N} > 1-\frac{1}{N_i}$ and $(1-\frac{1}{N})^{n-1} > (1-\frac{1}{N_n}) \cdot \prod_{i=1}^{n-1}(1-\frac{1}{N_n})$. Thus, inequality 7 holds and naive Bayes assigns class 0 to $\mathbf{x}^0$ which is correct. ■

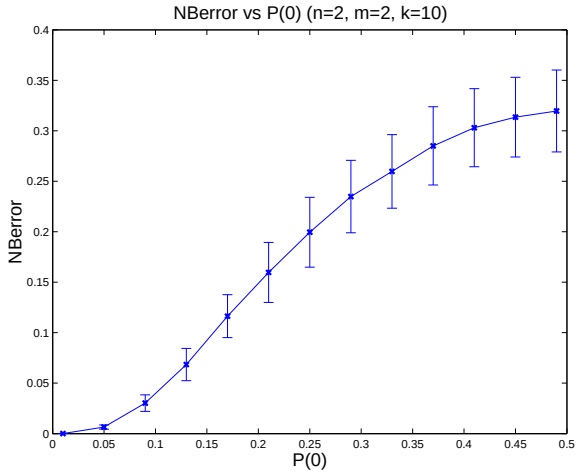
Clearly, theorem 1 also holds for concepts with only one example of class 1.

Our next questions are: how does the performance of naive Bayes change with increasing number of class-0 examples (i.e., with increasing prior $P(0)$)? Which data characteristics can be used as "error predictors"? We address these questions empirically using Monte-Carlo simulations. Assuming there are only two features each having k values, we vary the number l of class-0 examples from 1 to $k^2/2$ ($P(0)$ varies from $1/N$ to 0.5), and generate 1000 random problem instances for each value of l (it is not necessary to increase $P(0)$ beyond 0.5 due to the symmetry of the results for varying $P(0)$ and $P(1)$).

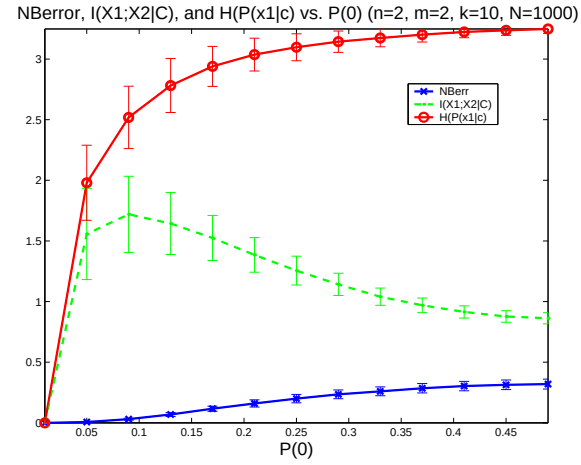
As expected, larger $P(0)$ (equivalently, larger l), yield a wider range of problems with various dependencies among features, which result into increased errors of Bayes (see Figure 1a); a closer look at the data shows no other cases of optimality besides $P(0) = 1/N$.

It is interesting to observe, however, that the strength of dependencies among features is not a good predictor of naive Bayes performance. Figure 1b plots average naive Bayes error and average class-conditional mutual information between the two features, $I(X_1; X_2|C)$, as functions of $P(0)$ (the other two plots in the figure will be discussed shortly). Note the monotone increase of the naive Bayes error and the non-monotone behavior of the mutual information. These results support some previous observations on UCI benchmarks (Domingos & Pazzani, 1997)) that also revealed low correlation between the degree of feature dependence and relative performance of naive Bayes with respect to other classifiers, such as C4.5, CN2, and PEBLS.

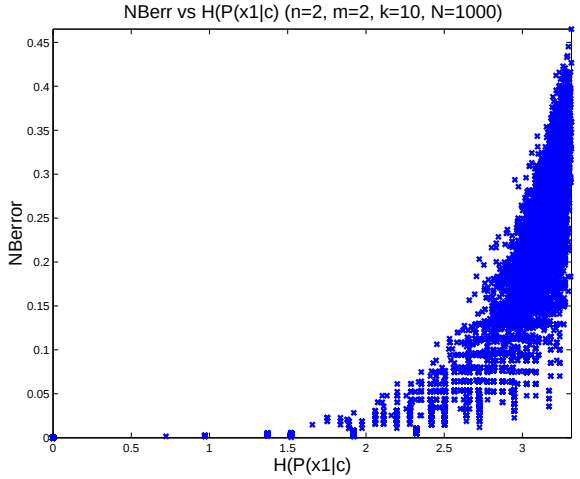
Another data characteristic related to the performance of naive Bayes is the entropy of class-conditional marginal distributions, $P(X_i|C)$. Consider again a concept defined on two features, which yields a k by k matrix filled with 0s and 1s. By analogy with the theorem 1, we would expect lower naive Bayes error when most of 0s are "con-



(a)



(b)



(c)

Figure 1. Results for the generator **ZeroBayesRisk** (k=10, 1000 problem instances per each of 13 values of $P(0)$ in $[0, 0.5]$ range): (a) Naive Bayes error vs. P(0); (b) Average values of naive Bayes error (NBerr), class-conditional mutual information between features ($I(X1; X2|C)$), and entropy of marginal distribution, $H(P(X_1|0))$; (c) Naive Bayes error vs. $H(P(X_1|0))$.

centrated" within one row or column (i.e., the entropy of $P(X_i|0)$ is low), rather than randomly distributed over the matrix (this intuition follows from the fact that the ratio $P(X_i|0)/P(X_i|1)$ is same for all examples having fixed value of X_i and is determined by majority class among those examples). Figure 1b plots average $H(P(X_1|0)$ versus $P(0)$; clearly, average naive Bayes error is a monotone function of average $H(P(X_1|0)$. Thus, an entropy of class-conditional marginals is a better indicator of naive Bayes error than the mutual information between the features. However, the variance of such prediction is quickly increasing with $P(0)$ and is quite high when $P(0)$ gets closer to 0.5 (see Figure 1c).

In the following section, we provide a more formal treatment of concepts with low-entropy marginals in the context of non-zero Bayes risk problems.

4. Naive Bayes on low-entropy distributions

We now focus on problems having non-zero Bayes risk. A natural question arises: is there an analog of theorem 1 for noisy concepts? In other words, what if $P(0|\mathbf{x}) = 1 - \delta$ rather than 1, for some small δ? We will consider next domains with almost deterministic, i.e. low-entropy, or "extreme", probability distributions, i.e. distributions having almost all the probability mass concentrated in one state. We show that approximating such distributions using independence assumption becomes more accurate with decreasing entropy, and therefore yield asymptotically optimal performance of naive Bayes.

The following lemma states that if a joint distribution over a set of variables is "extreme", than the marginal distributions of those variables are also "extreme".

Lemma 2 *Given a joint probability distribution $P(X_1, ..., X_n)$ such that $P(\mathbf{x}^*) \geq 1 - \delta$ for some state $\mathbf{x}^* = (x_1^*, ..., x_n^*)$, $\mathbf{x}^* \in \Omega$,then for each i,*

$$P(X_i = x_i^*) \geq 1 - \delta.$$

Proof. Let $S_i^* = \{x = (x_{1,...,}x_n)|x_i = x_i^*, x \in \Omega\}$. Since $P(X_i = x_i^*) = \sum_{\mathbf{x} \in S_i^*} P(\mathbf{x})$, and since S_i^* includes the point $(x_1^*, ..., x_n^*)$ that has $1 - \delta$ of all probability mass, we get $P(X_i = x_i^*) \geq 1 - \delta$. ■

The opposite result is also true: if all marginal distributions are "extreme", the joint distribution is "extreme".

Lemma 3 *Given marginal probability distributions $P(X_1), ..., P(X_n)$ such that for each i $P(X_i = x_i^*) \geq 1 - \delta$ for some x_i^*, then*

$$P(x_1^*, ..., x_n^*) \geq 1 - n\delta.$$

Proof. Since $P(X_i \neq x_i^*) \leq \delta$ for all i,

$$P(X_1 \neq x_1^* \vee \cdots \vee X_n \neq x_n^*) \leq \sum_i P(X_i \neq x_i^*) \leq n\delta,$$

using the simple union bound. The claimed bound follows by taking the complement of the event in the left hand side of the above inequality. ■

These results allow to compute a bound on approximation error when using independence assumption with "extreme" distributions, i.e. when the joint distribution is replaced by the product of marginals:

Theorem 4 *Given a joint probability distribution $P(X_1, ..., X_n)$ such that $P(x_1^*, ..., x_n^*) \geq 1 - \delta$ for some state $\mathbf{x}^* = (x_1^*, ..., x_n^*)$, then*

$$|P(x_1, ..., x_n) - \prod_{i=1}^{n} P(X_i = x_i)| \leq n\delta.$$

Proof. From lemma 2 it follows that $P(X_i = x_i^*) \geq 1 - \delta$ for any $i = 1, ..., n$. Since $1 - n\delta \leq (1-\delta)^n$ for $0 \leq \delta \leq 1$, we get

$$|P(\mathbf{x}^*) - \prod_{i=1}^{n} P(X_i = x_i^*)| \leq 1 - (1 - \delta)^n \leq n\delta.$$

On the other hand, if $\mathbf{x} = (x_1, ..., x_n)$ and $x_i \neq x_i^*$ for some i, then $P(X_i = x_i) \leq \delta$, and $P(\mathbf{x}) \leq \delta$, so that

$$|P(\mathbf{x}) - \prod_{i=1}^{n} P(X_i = x_i)| \leq \delta \leq n\delta,$$

which concludes the proof. ■

Similarly, it can be shown that

Theorem 5 *Given a set of marginal probability distributions $P(X_1), ..., P(X_n)$ such that for each i $P(X_i = x_i^*) \geq 1 - \delta$ for some x_i^*, then*

$$|P(x_1, ..., x_n) - \prod_{i=1}^{n} P(X_i = x_i)| \leq n\delta.$$

Proof. Let $\mathbf{x}^* = (x_1^*, ..., x_n^*)$. From lemma 3 it follows that $P(\mathbf{x}^*) \geq 1 - n\delta$, therefore, since $1 - n\delta \leq (1 - \delta)^n$, we get

$$|P(X = x^*) - \Pi_{i=1}^{n} P(X_i = x_i^*)| \leq n\delta.$$

On the other hand, if $\mathbf{x} = (x_1, ..., x_n)$ and $x_i \neq x_i^*$ for some i, then $P(\mathbf{x}) \leq n\delta$, and $P(X_i = x_i^*) \leq \delta$, so that

$$|P(\mathbf{x}) - \prod_{i=1}^{n} P(X_i = x_i)| \leq \delta \leq n\delta,$$

which concludes the proof. ■

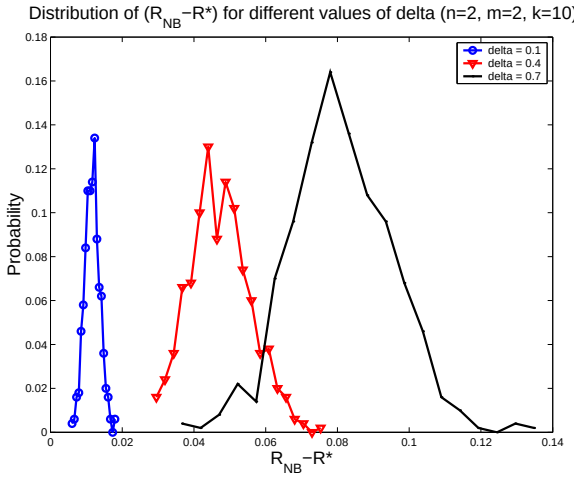
Clearly, if the difference between the joint feature distribution and its approximation by the product of marginals (all conditioned on class) vanishes with $\delta \to 0$, we would expect the naive Bayes error to vanish as well. Indeed, this is demonstrated using the following Monte-Carlo simulations.

A random problem generator, called **EXTREME**, takes the number of classes, m, number of features, n, number of values per feature, k, and the parameter δ, and creates m class-conditional feature distributions, each satisfying the condition $P(\mathbf{x}|C = c) = 1 - \delta$ if $\mathbf{x} = \mathbf{x}^c$, where $\mathbf{x}^c$ are m different states randomly selected from k^n possible states. For each class i, the remaining probability mass δ in $P(\mathbf{x}|C = i)$ is randomly distributed among the remaining $k^n - 1$ states. Class prior distributions are uniform. Once $P(\mathbf{X}|C)$ is generated, we compare naive Bayes classifier (NB) versus Bayes-optimal classifier (BO), assuming that both classifiers have perfect knowledge of data distribution (i.e., infinite amount of data).

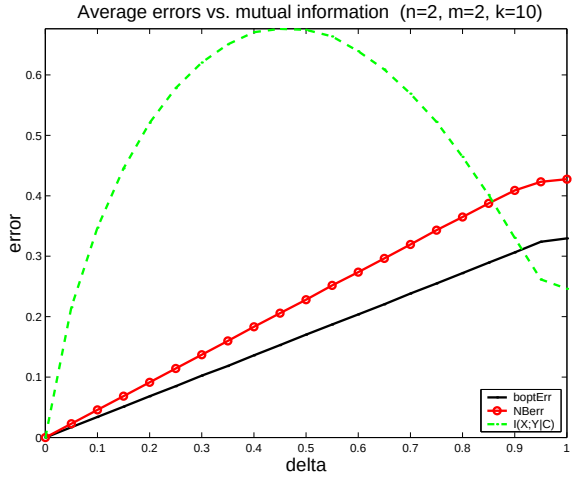
Simulation results on a set of 500 problems with $n = 2$, $m = 2$, $k = 10$, and δ varying from 0 to 1 are shown in Figure 2. Figure 2a shows the distributions of the difference between Bayes and naive Bayes errors, $R_{NB} - R^*$, as a function of δ. As expected, the distributions shift to the smaller values with decreasing δ.

Similarly to our previous results for zero Bayes risk, we observe that the strength of dependencies among features (for the sake of simplicity, we consider only two features here) is not correlated with naive Bayes error, as we see in figure 2b which plots Bayes error and naive Bayes error versus average mutual information between the two features, as functions of δ. The errors are monotone functions increasing with δ, while mutual information is a concave function reaching its maximum at intermediate value of δ (approximately between 0.45 and 0.5).

Similar results were also obtained for a different class of problems that mixes low-entropy distributions with an arbitrary ones. For one class, our generator called **MIX** creates a low-entropy class-conditional distribution as described before, and for the other one, it generates k^n random entries in the feature probability table, and then normalizes the probabilities. The naive Bayes error converges slower than in the case of two low-entropy distributions, since for same value of δ there is more noise in the problems due to random (instead of low-entropy) nature of one of the class-conditional feature distributions. The plot for average errors versus average mutual information (omitted here due to space limitations) looks very similar to the one we showed before, except that the errors are higher, and the average mutual information does not reach zero with $\delta \to 0$ since now the features do not become independent in the limit.



(a)



(b)

Figure 2. Results for the generator **EXTREME**: (a) Distribution of $R_{NB} - R^*$, the difference between naive Bayes and Bayes errors; (b) Average Bayes and naive Bayes errors versus average class-conditional mutual information.

5. Nearly-functional dependencies among features

Surprisingly, naive Bayes can be optimal in cases just opposite to the class-conditional feature independence (when mutual information is at minimum) - namely, in cases of completely deterministic dependence among the features (when mutual information achieves its maximum).

Theorem 6 *Given equal class priors, Naive Bayes is optimal if $X_i = f_i(X_1)$ for every feature X_i, $i = 2, \ldots, n$, where $f_i(\cdot)$ is a one-to-one mapping.*

Proof. Clearly, for every class c and feature i, $P_c(X_i = x_i|X_1 = x_1)$ is 1 if $x_i = f_i(x_1)$ and 0 otherwise (assuming that $f_1(x_1) = x_1$). Let $\mathbf{x}^* = (x_1, f_2(x), ..., f_n(x))$. Then this is the only non-zero-probability state. Therefore, by the theorem of total probability

$$P_c(x_1) = \sum_{\mathbf{x}:\mathbf{x}\neq\mathbf{x}^*} P_c(\mathbf{x}) + P_c(\mathbf{x}^*) = P_c(\mathbf{x}^*).$$

Then the Bayes-optimal for selecting class c can be written as

$$P_c(X_1 = x_1) > P_{c'}(X_1 = x_1) \forall c' \neq c.$$

On the other hand, the naive Bayes rule for selecting class c can be written as

$$\prod_{i=1}^{n} P_i(f_i(x_1)) > \prod_{i=1}^{n} P_{c'}(f_i(x_1)) \ \forall c' \neq c,$$

and, since $\forall c$, $P_c(f_i(x_1)) = P_c(x_1)$, we get

$$P_c(X_1 = x_1)^n > P_{c'}(X_1 = x_1)^n \ \forall c' \neq c.$$

Clearly, those two classification rules agree for every value of $\mathbf{x}$ that has nonzero probability, i.e. for every $\mathbf{x} = (x_1, f_2(x_1), ..., f_m(x_1))$. Thus naive Bayes is optimal. ■

Our next objective is to assess the naive Bayes performance for increasing noise in functional dependencies between the features. To answer this question, we used two random problem generators that relax functional dependencies using a noise parameter δ in a way similar to the low-entropy distribution generators. As before, we assume uniform class priors. Consider a simple case of two classes and two variables ($n = 2$ and $m = 2$) with k values each. Our almost-functional distribution generator called **FUNC1** selects a random permutation of k numbers, which corresponds to a one-to-one function f that binds the two features: $X_2 = f(X_1)$ $(1-\delta)$. Then it generates randomly two class-conditional (marginal) distributions for the X_1 feature, $P_0(X_1)$ and $P_1(X_1)$, for class 0 and class 1, respectively. Finally, it creates class-conditional joint feature distributions satisfying the following conditions:

$$P_c(x_1, x_2 = f(x_1)) = P_c(x_1)(1-\delta), \text{and}$$
$$P_c(x_1, x_2 \neq f(x_1)) = P_c(x_1)\frac{\delta}{k-1}, c = 0, 1.$$

This way the states satisfying functional dependence obtain $1-\delta$ probability mass, so that by controlling δ we can get as close as we want to the functional dependence described before, i.e. the generator relaxes the conditions of theorem 6. Note that, on the other hand, $\delta = \frac{k-1}{k}$ gives us uniform distributions over the second feature $P_c(x_2) = \sum_{x_1} P_c(x_1, x_2) = \frac{1}{k}$, which makes it independent of X_1 (given class c). Thus varying δ from 0 to 1 explores the whole range from deterministic dependence to complete independence between the features given class. Figure 3 summarizes the results for 500 problems with $n = 2$, $m = 2$ and $k = 10$. Figure 3a shows the average errors; as we can see, naive Bayes is optimal when $\delta = 0$ and when $\delta = 0.9$, while the maximum error is reached between the two extremes. Note again, that errors are not correlated with the average class-conditional mutual information between the features (see Figure 3b), which decreases monotonically with δ going from 0 to 0.9. So, we can conclude that naive Bayes is close to the optimal not only for close-to-independent features (low mutual information), but also for close to one-to-one functional dependence.

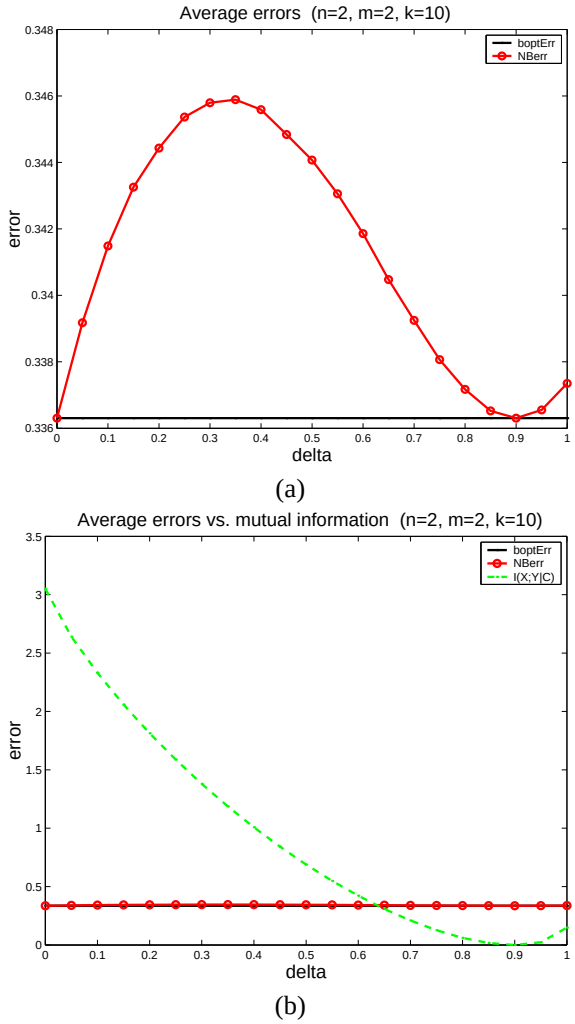


Figure 3. Results for the generator **FUNC1**: (a) Average Bayes and naive Bayes errors ; (b) Average Bayes and naive Bayes errors versus average class-conditional mutual information.

Similar results are obtained for a modification of generator **FUNC1**, called **FUNC2**, where the probability mass δ is distributed over the rest of states not uniformly but randomly. As expected, in this case we do not get complete in-

dependence, and therefore average naive Bayes error only increases with δ. Again, no (linear) correlation was observed between average error and average mutual information.

6. Conclusions

Despite its unrealistic independence assumption, the naive Bayes classifier is surprisingly effective in practice (Domingos & Pazzani, 1997; Mitchell, 1997; Friedman et al., 1997), since it often assigns maximum probability to the correct class even if its probability estimates are inaccurate. Although some optimality conditions of naive Bayes have been already identified in the past (Domingos & Pazzani, 1997), a deeper understanding of data characteristics that affect the performance of naive Bayes is still required.

In this paper, we focus on problems including deterministic and close to deterministic dependencies, often present in some practical applications. First, we address zero-Bayes-risk problems, proving naive Bayes optimality for any two-class concept that assigns class 0 to exactly one example, i.e. has zero-entropy $P(x_i|0)$ (a generalization of conjunctive and disjunctive concepts to arbitrary nominal features). Then we demonstrate empirically that the entropy of $P(x_i|0)$ is a better predictor of the naive Bayes error than the class-conditional mutual information between features. Next, we consider a broader class of non-zero Bayes risk problems, further pursuing the idea of low-entropy distributions. We derive error bounds for approximating the joint distribution by the product of marginals in case of nearly-deterministic class-conditional feature distributions $P(x_i|C)$, and we demonstrate how the performance of naive Bayes improves with decreasing entropy of such distributions. Finally, we consider functional dependencies between features and prove naive Bayes optimality in certain cases. Using Monte Carlo simulations, we show that naive Bayes works best in two cases: completely independent features (as expected by the assumptions made) and functionally dependent features (which is surprising). Naive Bayes has its worst performance between these extremes.

Directions for future work include analysis of naive Bayes on practical application that include almost-deterministic dependencies, and further investigation of data characteristics that yield a good performance of naive Bayes. An ultimate goal is to characterize probability distributions that are insensitive to the independence assumption w.r.t. to classification task. This approach aims directly at learning probabilistic models that result into better classification accuracy, even though such models may not provide good approximations with respect to other criteria, such as KL-distance between true and estimated probability distributions, or MDL criterion. Empirical results also suggest that such general model selection criteria do not necessarily lead to better classifiers (Friedman et al., 1997).

Acknowledgements

We wish to thank Mark Brodie, Vittorio Castelli, Daniel Oblinger, and Ricardo Vilalta for many insightful discussions that contributed to the ideas of this paper.

References

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York:John Wiley & Sons.

Dechter, R. (1997). Mini-buckets: A general scheme for generating approximations in automated reasoning. *Proc. Fifteenth International Joint Conference of Artificial Intelligence (IJCAI-97), Japan* (pp. 1297–1302).

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

Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis*. New York: John Wiley and Sons.

Frey, B., & MacKay, D. (1998). A revolution: Belief propagation in graphs with cycles. *Advances in Neural Information Processing Systems*, *10*.

Friedman, N., Geiger, D., & M., G. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Heckerman, D. (1995). *A tutorial on learning Bayesian networks, technical report msr-tr-95-06* (Technical Report). Microsoft Research.

Hellerstein, J., Thathachar, J., & Rish, I. (2000). Recognizing end-user transactions in performance management. *Proceedings of AAAI-2000* (pp. 596–602). Austin, Texas.

Kohavi, R. (1995). *Wrappers for performance enhancement and oblivious decision graphs* (Technical Report). PhD thesis, Department of Computer Science, Stanford, CA.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 399–406). San Jose, CA: AAAI Press.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Rish, I. (1999). *Efficient reasoning in graphical models*. PhD thesis.