Journal of Machine Learning Research 7 (2006) 2449-2480 Submitted 6/06; Revised 10/06; Published 11/06

Bounds for the Loss in Probability of Correct Classification Under Model Based Approximation

Magnus Ekdahl MAEKD@MAI.LIU.SE
Timo Koski TIKOS@MAI.LIU.SE
Matematiska Institutionen
Linköpings Universitet
581 83 Linköping, Sweden

Editor: Michael Jordan

Abstract

In many pattern recognition/classification problem the true class conditional model and class probabilities are approximated for reasons of reducing complexity and/or of statistical estimation. The approximated classifier is expected to have worse performance, here measured by the probability of correct classification. We present an analysis valid in general, and easily computable formulas for estimating the degradation in probability of correct classification when compared to the optimal classifier. An example of an approximation is the Naïve Bayes classifier. We show that the performance of the Naïve Bayes depends on the degree of functional dependence between the features and labels. We provide a sufficient condition for zero loss of performance, too.

Keywords: Bayesian networks, naïve Bayes, plug-in classifier, Kolmogorov distance of variation, variational learning

1. Introduction

Classification procedures based on probability models are widely used in data mining and machine learning (Hand et al., 2001), since such models often lead to effective algorithms and modularity in computation and have a conceptual foundation in statistical learning theory. For tractable computation and learning these models may still in many cases require steps of approximation by less complex model families (Jordan et al., 1999).

By classification we mean procedures that group items represented by a feature vector into different predefined classes. We consider classification procedures based on class conditional probabilities that belong to a model family that does not necessarily contain the true probability distribution, and analyze how the probability of correct classification is affected.

One straightforward procedure is known as a plug-in function. By this we refer to the formal operation performed by the optimal classifier based on Bayes' formula of posterior probabilities of classes, but now plugging in the modeling or approximate class conditional densities as well as approximated class probabilities. There are still a lot of unresolved issues concerning the effects of plug-in functions in the context of classification with high-dimensional feature vectors.

A well known plug-in procedure in classification is modeling by independence, which is usually called the 'Naïve Bayes' classifier. We will review, extend and sharpen the theoretical justification for this procedure while connecting it to the general approximation theory. Friedman (1997) studies also the Naïve Bayes, when the optimal classifier is estimated from training data. He shows that the bias and variance components of the estimation error affect classification error in a different way under the Gaussian approximation than the error in the estimated probabilities. This can help Naïve Bayes to perform better than expected in case the variance of the estimates of posterior probabilities is low. Our analysis in the sequel will not involve the variance $-$ bias decomposition.

Bayesian networks is a widely used class of models for probabilistic reasoning and for classification, see for example (Korb and Nicholson, 2004; Friedman et al., 1997). As the network topologies increase in size and complexity, the run-time complexity of probabilistic inference and classification procedures becomes prohibitive. In general, exact inference on Bayesian networks is known to be NP-hard (Cooper, 1990). One way of approximating or simplifying the model is to enforce additional conditional independencies or by removing edges in the graph, see van Engelen (1997) and the references therein.

Here we analyze a simplification of Bayesian networks by a strategy of approximating factors of the joint probability, and give a bound for the probability of correct classification under the ensuing plug-in function. This corresponds to some degree to the general heuristics in the work by Lewis (1959); Brown (1959); Chow and Liu (1968); Ku and Kullback (1969), who developed the idea of approximating multivariate discrete probability distributions by a product of lower order marginal distributions. The set of marginal distributions applied needs not be the full set of margins of some order, the requirements are that the product is an extension of the lower order distributions which are compatible.

2. Organization

We will start by introducing notation and basic definitions in Section 3. Section 4 provides rationales and examples of approximating models and plug-in classifiers. These will be used to illustrate the mathematical results in the following sections. Section 5 introduces results about the degradation of classifier performance with respect to the optimal probability of correct classification. The results are phrased in terms of a distance between probabilities known as the Kolmogorov variation distance. There are several well known bounds for the Kolmogorov variation distance by other distances between probability measures, quoted in Section 5, which in many examples yield explicit and computable bounds for the plug-in classifier performance. We give also a novel bound that connects the work to variational learning theory (Jordan et al., 1999). Section 6 gives a rule for potential reduction of the number of dimensions needed for evaluating the degradations, and presents more easily computable bounds. Section 7 discusses the Naïve Bayes classifier by sharpening a bound for Naïve Bayes and connecting it to one of the general approximation bounds in Section 6. Section 8 gives sufficient conditions on the margin (explained later) between two classes, which is used to generalize the possible problems where Naïve Bayes can be argued as optimal.

3. Notation, Bayes and Plug-In Classifiers

Let $(\Omega,\mathcal{F},P)$ be a probability space, such that (C,X) is a $\mathcal{F}$-measurable stochastic variable, s.v. Let $X=(X_i)_{i=1}^d$, that is X is d-dimensional. When denoting a sample (observation) of X with no missing components we use x, that is $x=(x_i)_{i=1}^d$ (x can be called a feature vector). When referring to the range of X we use $\mathcal{X}$, which for completeness of presentation is assumed to be a Borel space (Schervish, 1995). This assumption is needed to justify the use of results such as the Fubini theorem and the existence of conditional densities.

In the context of classification a sample x is assumed to have a source, one of a family of entities called classes or labels, denoted by c, which is regarded as an outcome of the random variable C. In classification C has range $\mathcal{C} = \{1,\ldots,k\}$, that is, k is the number of classes. We assume, as is common in much of classification theory, that the space of labels $\{1,\ldots,k\}$ is without relevant additional structure except whether two labels are equal or not. In order to resolve ties, it may, on the other hand, be useful to think of the labels as ordered by $1 < 2 < \ldots < k$.

Definition 1 *A classifier is a measurable function $\hat{c} : \mathcal{X} \to \mathcal{C}$ such that given x, $\hat{c}(x)$ is an estimate of c.*

In classification we do not deal directly with the whole sample (c,x), but the class c is a hidden variable. Hence we will deal with the class conditional probability. In

$$P(X \in A|C = c) = \int_A f(x|c)d\mu(x)$$

we call $f(x|c)$ the conditional density of a sample x given that the random variable C equals the label c with respect to the σ-finite measure μ. We assume in other words that μ dominates, see Schervish (1995), the probability measure $P(\cdot|C = c)$ for every c, that is, the same measure μ can be used for all $P(X\cdot|C = c)$ to define the corresponding class conditional density $f(x|c)$. The assumption of domination justifies the validity of a number of formulas of distances between probability measures. $P(c)$ is the short notation for the marginal probability $P(C = c)$. We also encounter $P(c|x)$, the probability of the class c given the sample x. $P(c|x)$ is used to define a classifier which is the cornerstone of probabilistic classification. For example, the procedure known as proportional prediction chooses the label for x by drawing c from the probability mass function $P(c|x)$ (Goodman and Kruskal, 1954). We study only deterministic classifiers $\hat{c}$.

Definition 2 *Bayes classifier for a sample x is*

$$\hat{c}_B(x) = \arg\max_{c\in\mathcal{C}} P(c|x).$$

Ties are resolved in some fixed manner, for example, by taking $\hat{c}_B(x)$ the smallest of the tied labels in (the ordered) $\mathcal{C}$.

The posterior $P(c|x)$ can be modeled directly ('the diagnostic paradigm') but this may often involve difficult computations (Ripley, 1996). Bayes' formula gives effectively

$$\hat{c}_B(x) = \arg\max_{c\in\mathcal{C}} f(x|c)P(c). \tag{1}$$

Thus we base Bayes classifier on $f(x|c)$ as well as on $P(c)$, the prior probability (or the prevalence) of class c. In essence $f(x|c)$ allows us to think of each class as generating x.

We evaluate the performance of a classifier by the probability of correct classification and assess the effect of approximating $f(x|c)$ by $\hat{f}(x|c)$ and $P(c)$ by $\hat{P}(c)$.

Definition 3 *For a classifier $\hat{c}(X)$ the probability of correct classification is $P(\hat{c}(X) = C)$.*

There is a good reason for using Bayes classifier (Definition 2), since for every $\hat{c}(X)$ it holds that

$$P(\hat{c}(X)=C) \leqslant P(\hat{c}_B(X)=C).$$

A simple way of constructing a classifier given $\hat{f}(x|c)$ and $\hat{P}(c)$ is to use these to replace the respective target probabilities in (1).

Definition 4 *$\hat{c}_{\hat{B}}(x)$ is a plug-in classifier with respect to the pair $\left(\hat{f}(x|c), \hat{P}(c)\right)$ if it is defined by*

$$\hat{c}_{\hat{B}}(x) = \arg\max_{c \in C} \hat{f}(x|c)\hat{P}(c). \tag{2}$$

Ties are resolved as in Definition 2.

The question studied here can now be stated as that of computing or bounding the difference

$$P(\hat{c}_B(X)=C) - P(\hat{c}_{\hat{B}}(X)=C).$$

It is for many P difficult or even impossible to compute explicitly $P(\hat{c}_B(X)=C)$. Hence there exists a literature for bounding the optimal probability of error, $P_e^* = 1 - P(\hat{c}_B(X)=C)$. If we set $P_e = 1 - P(\hat{c}_{\hat{B}}(X)=C)$, then

$$P(\hat{c}_B(X)=C) - P(\hat{c}_{\hat{B}}(X)=C) = P_e - P_e^*.$$

This can be bounded downwards by, for example, the upper bounds for P_e^* in Bhattacharyya and Toussaint (1982). We shall not pursue the lower bounds for $P(\hat{c}_B(X)=C) - P(\hat{c}_{\hat{B}}(X)=C)$ any further.

4. Examples of Plug-In Approximations of the Bayes Classifier

As outlined in the introduction, there are several reasons for approximating $f(x|c)$ in classification. These include the problem of digitally storing probability tables, the topic introduced in Lewis (1959), and the complexity, or even infeasibility, of computing $\hat{c}_B(x)$. Therefore we could call f the target density and $\hat{f}$ the tractable density (Wainwright and Jordan, 2003). In block transmission systems a tractable density is found for fast computation of the signal classifier (detector) (Kaleh, 1995). In this section we present some examples of plug-in classifiers motivated by these considerations in pattern recognition and detection.

Example 1 *We consider $\mathcal{X} = \{0,1\}^d$ known as the binary hypercube in d dimensions. For the binary hypercube we need in general $2^d - 1$ parameters to specify each class conditional probability mass function. Hence we may encounter a difficulty with storing of the tables of probabilities.*

There are several canonical representations of the generic probability distribution on $\mathcal{X}$ and of the $2^d - 1$ parameters. Examples of these are given in Bahadur (1961b), Devroye et al. (1996), Ott and Kronmal (1976), and Teugels (1990). We recapitulate the representation by Bahadur (1961b) in the form given by Brunk and Pierce (1974). Let $f(x|c)$ be a probability mass function on $\mathcal{X}$ such that $f(x|c) > 0$ for all $x \in \mathcal{X}$. Let

$$f_{ic} = \sum_{x \in \mathcal{X}, x_i = 1} f(x|c), \quad y_{ic} = y_{ic}(x) = \frac{x_i - f_{ic}}{\sqrt{f_{ic}(1-f_{ic})}}. \tag{3}$$

Let $\mathbf{w} = (w_1, w_2, \ldots, w_d) \in \{0,1\}^d$ *be a binary vector of zeros and ones. Then we denote by* $U_{\mathbf{w},c}(x)$ *products of a subset of* $y_{1c}, \ldots, y_{dc}$

$$U_{\mathbf{w},c}(x) = \prod_{i=1}^{d} y_{ic}(x)^{w_i}, \quad U_{\mathbf{0},c}(x) = 1.$$

We set

$$f_1(x|c) = \prod_{i=1}^{d} f(x_i|c) = \prod_{i=1}^{d} f_{ic}^{x_i}(1 - f_{ic})^{1-x_i}. \tag{4}$$

Hence $f_1(x|c)$ *is another probability mass function, which is positive on* $\{0,1\}^d$*. Its marginal distributions coincide with those of* $f(x|c)$*. With respect to* f_1 *any binary random vector* $X = (X_i)_{i=1}^d$ *consists of independent components* X_i*.*

We shall next regard the set V *of real-valued functions on* $\{0,1\}^d$ *as a vector space of dimension* 2^d*. Let us equip* V *with the scalar product defined for* $\phi \in V, \psi \in V$ *as*

$$(\phi, \psi) = \sum_{x \in \{0,1\}^d} \phi(x)\,\psi(x)\, f_1(x|c). \tag{5}$$

Next we show that the functions $\{U_{\mathbf{w},c}(x)\}_{\mathbf{w} \in \{0,1\}^d}$ *constitute an orthonormal basis with respect to this scalar product. In fact*

$$(U_{\mathbf{w},c}, U_{\mathbf{w}^*,c}) = \sum_{x \in \{0,1\}^d} U_{\mathbf{w},c}(x) U_{\mathbf{w}^*,c}(x) f_1(x|c) =$$

$$= \sum_{x \in \{0,1\}^d} \prod_{i=1}^{d} y_{ic}(x)^{w_i} y_{ic}(x)^{w_i^*} f_1(x|c).$$

The sum in the right hand side is nothing but the expectation

$$E_{f_1}\left[\prod_{i=1}^{d} y_{ic}(X)^{w_i} y_{ic}(X)^{w_i^*}\right] = \prod_{i=1}^{d} E_{f_1}\left[y_{ic}(X)^{w_i} y_{ic}(X)^{w_i^*}\right], \tag{6}$$

where we used the aforementioned independence of the components of X *under* f_1*, which yields the independence of the* $y_{ic}(X)$ *as defined by (3), too. We now show that the product in (6) equals zero, if* $\mathbf{w} \neq \mathbf{w}^*$*. In this case there is at least one i such that* $w_i \neq w_i^*$*, and for this i we get*

$$E_{f_1}\left[y_{ic}(X)^{w_i} y_{ic}(X)^{w_i^*}\right] = E_{f_1}\left[y_{ic}(X)\right] = \frac{E_{f_1}[X_i] - f_{ic}}{\sqrt{f_{ic}(1 - f_{ic})}} = 0,$$

since by the definitions above $E_{f_1}[X_i] = 1 \cdot P_1(X_i = 1) = f_{ic}$*. Hence the whole product in (6) is zero, and we have shown that* $(U_{\mathbf{w},c}, U_{\mathbf{w}^*,c}) = 0$*, if* $\mathbf{w} \neq \mathbf{w}^*$*. If* $\mathbf{w} = \mathbf{w}^*$*, then we get from (6) that*

$$(U_{\mathbf{w},c}, U_{\mathbf{w},c}) = \prod_{i=1:w_i=1} E_{f_1}\left[y_{ic}(X)^2\right].$$

Here

$$E_{f_1}\left[y_{ic}(X)^2\right] = \frac{1}{f_{ic}(1 - f_{ic})} E_{f_1}\left[(X_i - f_{ic})^2\right]$$

But since X_i is a binary random variable (or, a Bernoulli random variable) with respect to f_1, we have

$$E_{f_1}\left[\left(X_i - f_{ic}\right)^2\right] = f_{ic} - f_{ic}^2 = f_{ic}\left(1 - f_{ic}\right).$$

Hence $\left(U_{\mathbf{w},c}, U_{\mathbf{w},c}\right) = 1$, and we have shown that $\left\{U_{\mathbf{w},c}\left(x\right)\right\}_{\mathbf{w}\in\{0,1\}^d}$ is an orthonormal set in V with respect to the scalar product in (5). Since the number of functions in $\left\{U_{\mathbf{w},c}\left(x\right)\right\}_{\mathbf{w}\in\{0,1\}^d}$ equals the dimension of V (=2^d), $\left\{U_{\mathbf{w},c}\left(x\right)\right\}_{\mathbf{w}\in\{0,1\}^d}$ must be an orthonormal basis in V.

Hence every function $\phi \in V$ has a unique expansion in terms of the 2^d coordinates $\left(\phi, U_{\mathbf{w},c}\right)$ with respect to this basis written as

$$\phi\left(x\right) = \sum_{\mathbf{w}\in\{0,1\}^d} (\phi, U_{\mathbf{w},c}) U_{\mathbf{w},c}\left(x\right). \tag{7}$$

If we take $\phi\left(x\right) = f(x|c)/f_1(x|c)$ we obtain

$$\left(\frac{f}{f_1}, U_{\mathbf{w},c}\right) = \sum_{x\in\{0,1\}^d} f(x|c) U_{\mathbf{w},c}\left(x\right) = E_f\left(U_{\mathbf{w},c}(X)\right).$$

In other words the coordinate $\left(\frac{f}{f_1}, U_{\mathbf{w},c}\right)$ equals the expectation of $U_{\mathbf{w},c}(X)$ w.r.t. to the probability mass function $f(x|c)$. For this we introduce the standard notation

$$\beta_{\mathbf{w},c} = E_f\left(U_{\mathbf{w},c}(X)\right). \tag{8}$$

By substitution of (8) in (7) we obtain

$$\frac{f(x|c)}{f_1(x|c)} = \sum_{\mathbf{w}\in\{0,1\}^d} \beta_{\mathbf{w},c} U_{\mathbf{w},c}\left(x\right).$$

This gives us the the (Bahadur-Lazarsfeld) representation of any positive probability mass function $f(x|c)$ on $\{0,1\}^d$ as

$$f(x|c) = f_1(x|c) f_{c,\text{interactions}}(x), \tag{9}$$

where we have written

$$f_{c,\text{interactions}}(x) = \sum_{\mathbf{w}\in\{0,1\}^d} \beta_{\mathbf{w},c} U_{\mathbf{w},c}\left(x\right). \tag{10}$$

The rank $R(\mathbf{w})$ of the polynomial $U_{\mathbf{w},c}$ is defined as

$$R(\mathbf{w}) = \sum_{i=1}^{d} w_i.$$

Here $\beta_{\mathbf{0},c} = 1$, and if $R(\mathbf{w}) = 1$, then $\beta_{\mathbf{w},c} = 0$. The probability mass function $f_1(x|c)$ in (4) is known as the first order term. For $R(\mathbf{w}) = 2$ the coefficients $\{\beta_{\mathbf{w}}\}$ are correlations. We can think of the coefficients β as interactions of order $R(\mathbf{w})$ minus one.

One can define a family of probability mass functions called kth order Bahadur distributions as the set of all probabilities on the binary hypercube in d dimensions such that $\beta_{\mathbf{w}} = 0$ for $R(\mathbf{w}) > k$. Anoulova et al. (1996) prove, simplifying their statement, that there is an algorithm that, given enough samples, computes for any $\varepsilon > 0$ a plug-in classifier $\hat{c}_{\hat{B}}(x)$ such that $P(\hat{c}_B(X) = C) -$

$P(\hat{c}_{\hat{B}}(X) = C) \leqslant \varepsilon$, when the conditional distributions of $X|C$ are in the class of kth order Bahadur distributions.

If we expand $\log \frac{f(x|c)}{f_1(x|c)}$ with respect to the basis $\{U_{\mathbf{w},c}(x)\}_{\mathbf{w}\in\{0,1\}^d}$ we obtain the following canonical form

$$f(x|c) = f_1(x|c)e^{\sum_{\mathbf{w}\in\{0,1\}^d} \alpha_{\mathbf{w},c}U_{\mathbf{w},c}(x)}, \tag{11}$$

where it follows similarly as above that

$$\alpha_{\mathbf{w},c} = E_{f_1}\left[\log \frac{f(X|c)}{f_1(X|c)} \cdot U_{\mathbf{w},c}(X)\right]. \tag{12}$$

The two canonical forms (10) and (11) above are of interest in the sequel for defining structures of approximations and for evaluating the effect of a plug-in classifier on probability of correct classification. First, the plug-in classifier

$$\hat{c}_{\hat{B}}(x) = \arg\max_{c\in C} f_1(x|c)\hat{P}(c)$$

is an instance of the Naïve Bayes procedure to be treated in more generality in Section 7 below. In the setting of the binary hypercube the Naïve Bayes is said to take into account only the first order term. A survey of the Naïve Bayes in supervised and unsupervised learning of bacterial taxonomies using binary features is found in Gyllenberg and Koski (2001). Further structures of plug-in classifiers can be defined by adding sets of higher order interactions to the first order term. Examples of this are found in Bahadur (1961a), Chow and Liu (1968), Moore (1973), and Ott and Kronmal (1976). Here the trade-off is between the additional complexity and the more accurate statistical description, and, as it will turn out in the sequel, higher probability of correct classification with the plug-in classifier.

A successful empirical application of the Bahadur representation in classification or diagnosis of six diseases using eleven features or symptoms is reported in Scheinck (1972). The underlying requirement $f(x|c) > 0$ for all $x \in \{0,1\}^{11}$ is possibly overlooked in Scheinck (1972).

In some of the contributions referred to in the above the approximating structure is not necessarily a probability, since an arbitrary truncation of a representation of a probability mass function with respect to a basis is not always a probability mass function.

In case the support of f is a true subset of $\mathcal{X} = \{0,1\}^d$, a canonical representation (an interpolator) of f has been reported in Pistone et al. (2001). This is based on the monomials $\prod_{i=1}^d x_i^{w_i}$ and the properties of Gröbner bases.

Example 2 *One model of intersymbol interference (ISI) channels in digital communication theory, see Kaleh (1995); Barbosa (1989), can be formulated as observing a $d \times 1$ vector X with the class conditional normal distribution*

$$X|C = \mathbf{b} \sim N(H\mathbf{b}, \Sigma),$$

where $\mathbf{b}$ *is* $N \times 1$ *vector such that* $b_i \in \{-1,+1\}$*, and* Σ *is a positive definite* $d \times d$ *matrix, and* H *represents a linear, time-invariant and causal ISI channel by the* $d \times N$ *matrix*

$$H = \begin{pmatrix} h_0 & 0 & \dots & 0 \\ h_1 & h_0 & \ddots & \vdots \\ \vdots & h_1 & \ddots & 0 \\ h_{L-1} & \vdots & \ddots & h_0 \\ 0 & h_{L-1} & \ddots & h_1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & h_{L-1} \end{pmatrix}.$$

Here L *is the length of the channel memory, if* $h_0 \neq 0$ *and* $h_{L-1} \neq 0$*. Hence* $d = L+N-1$*. The set of labels* $\mathcal{C}$ *equals in this case a subset of* $\{-1,+1\}^N$*.* $\mathcal{C}$ *might be called a codebook. If all* $\mathbf{b}$ *are equally likely a priori, we have (the optimal detector)*

$$\hat{c}_B(x) = \arg\min_{\mathbf{b}\in\mathcal{C}} \|\Sigma^{-1}(x - H\mathbf{b})\|^2,$$

where $\|x\| = \sqrt{x^T x}$*.*

A suboptimal detector may be introduced, for example, for the purpose of reducing run time complexity, see Barbosa (1989), by a $d \times N$ *matrix* M *of the same structure as* H*, but with a shorter memory and the plug-in classifier*

$$\hat{c}_{\hat{B}}(x) = \arg\min_{\mathbf{b}\in\mathcal{C}} \|\Sigma^{-1}(x - M\mathbf{b})\|^2.$$

Here explicit expressions for both $P(\hat{c}_B(X) = C)$ *and* $P(\hat{c}_{\hat{B}}(X) = C)$ *are readily found, and the question of developing techniques for estimating the loss of performance incurred by the introduction of the suboptimal detector has been studied extensively for a number of designs of the matrices* M*.*

5. A Performance Bound

There are several representations of the exact difference $P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C)$.
For typographical and readability reasons we will use the notation $\hat{c}_B(x) = b$ as well as $\hat{c}_{\hat{B}}(x) = \hat{b}$.
We can write $P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C)$ as

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) = \int_{\{\hat{b}\neq b\}} \left(P(b)f(x|b) - P(\hat{b})f(x|\hat{b})\right) d\mu(x). \tag{13}$$

We may also re-write this as

$$= \int_{\{\hat{b}\neq b\}} \left(P(b)f(x|b) - \hat{P}(b)\hat{f}(x|b)\right) d\mu(x) - \int_{\{\hat{b}\neq b\}} \left(P(\hat{b})f(x|\hat{b}) - \hat{P}(\hat{b})\hat{f}(x|\hat{b})\right) d\mu(x)$$
$$- \int_{\{\hat{b}\neq b\}} \left(\hat{P}(\hat{b})\hat{f}(x|\hat{b}) - \hat{P}(b)\hat{f}(x|b)\right) d\mu(x) \tag{14}$$

since

$$= \int_{\{\hat{b}\neq b\}} P(b)f(x|b)d\mu(x) - \int_{\{\hat{b}\neq b\}} \hat{P}(b)\hat{f}(x|b)d\mu(x)$$

$$-\int_{\{\hat{b}\neq b\}} P(\hat{b})\, f(x|\hat{b}) d\mu(x) + \int_{\{\hat{b}\neq b\}} \hat{P}(\hat{b})\, \hat{f}(x|\hat{b}) d\mu(x)$$

$$-\int_{\{\hat{b}\neq b\}} \hat{P}(\hat{b})\, \hat{f}(x|\hat{b}) d\mu(x) + \int_{\{\hat{b}\neq b\}} \hat{P}(b)\, \hat{f}(x|b) d\mu(x),$$

where 4 integrals cancel each other and (13) is formed. The difficulty with these expressions is to find the set $\{x|\hat{c}_B(x) \neq \hat{c}_{\hat{B}}(x)\}$ and to compute the integrals above.

We give next a first upper bound for $P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C)$. The result for the specific case of $k = 2$ is presented in Ryzin (1966). When $k \geqslant 2$ the result in Theorem 5 can basically be found inside a proof in Glick (1972). A proof is included here for completeness and readability. For the specific approximation, where $\hat{f}$ and $\hat{P}$ are the maximum likelihood estimators, and samples are discrete, rates of convergence are provided in Glick (1973), as the sample size increases to infinity.

Theorem 5

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) \leqslant E_{\hat{f}\hat{P}} \left| \frac{f(X|C)P(C)}{\hat{f}(X|C)\hat{P}(C)} - 1 \right|. \tag{15}$$

Proof From (14) $P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) =$

$$\int_{\{x|\hat{b}\neq b\}} \left(P(b)\, f(x|b) - \hat{P}(b)\, \hat{f}(x|b)\right) d\mu(x)$$

$$-\int_{\{x|\hat{b}\neq b\}} \left(P(\hat{b})\, f(x|\hat{b}) - \hat{P}(\hat{b})\, \hat{f}(x|\hat{b})\right) d\mu(x) - \int_{\{x|\hat{b}\neq b\}} \left(\hat{f}(x|\hat{b})P(\hat{b}) - \hat{f}(x|b)P(b)\right) d\mu(x).$$

Definition 4 implies that $\hat{f}(x|\hat{b})\hat{P}(\hat{b}) \geqslant \hat{f}(x|b)\hat{P}(b)$, hence

$$\leqslant \int_{\{x|\hat{b}\neq b\}} \left(P(b)\, f(x|b) - \hat{P}(b)\, \hat{f}(x|b)\right) d\mu(x) - \int_{\{x|\hat{b}\neq b\}} \left(P(\hat{b})\, f(x|\hat{b}) - \hat{P}(\hat{b})\, \hat{f}(x|\hat{b})\right) d\mu(x).$$

To simplify further $\int a - e \leqslant |\int a - e| \leqslant |\int a| + |\int e| \leqslant \int |a| + \int |e|$ is used, resulting in

$$\leqslant \int_{\{x|\hat{b}\neq b\}} \left|P(b)\, f(x|b) - \hat{P}(b)\, \hat{f}(x|b)\right| d\mu(x) + \int_{\{x|\hat{b}\neq b\}} \left|P(\hat{b})\, f(x|\hat{b}) - \hat{P}(\hat{b})\, \hat{f}(x|\hat{b})\right| d\mu(x).$$

Then divide into cases where b as well as $\hat{b}$ are constant (they both depend on x)

$$= \sum_{c=1}^{k} \left[\int_{\left\{x|b\neq\hat{b}\bigcap b=c\right\}} \left|P(c)\, f(x|c) - \hat{P}(c)\, \hat{f}(x|c)\right| d\mu(x) \right.$$

$$\left. + \int_{\left\{x|b\neq\hat{b}\bigcap\hat{b}=c\right\}} \left|P(c)\, f(x|c) - \hat{P}(c)\, \hat{f}(x|c)\right| d\mu(x) \right].$$

Now $b \neq \hat{b} \bigcap b = c$ and $b \neq \hat{b} \bigcap \hat{b} = c$ are disjoint sets so we can write both integrals as one integral,

$$= \sum_{c=1}^{k} \int_{\left\{x|b\neq\hat{b}\bigcap\left(b=c\bigcup\hat{b}=c\right)\right\}} \left|P(c)\, f(x|c) - \hat{P}(c)\, \hat{f}(x|c)\right| d\mu(x)$$

We want an approximation that does not depend on $b, \hat{b}$, such as

$$\leqslant \sum_{c=1}^{k} \int_{X} \left| P(c) f(x|c) - \hat{P}(c) \hat{f}(x|c) \right| d\mu(x) = \sum_{c=1}^{k} \int_{X} \hat{P}(c) \hat{f}(x|c) \left| \frac{P(c) f(x|c)}{\hat{P}(c) \hat{f}(x|c)} - 1 \right| d\mu(x).$$

■

The right hand side of the inequality (15) is, when multiplied by the factor $1/2$, an instance of what is being called the Kolmogorov distance of variation, see for example Ali and Silvey (1966). We shall resort to this terminology in order to be able to refer concisely to the quantity in the right hand side of (15) (or of (16) below). The basic mathematical properties of this distance are found in Strasser (1985). Probabilistically the size of the quantity $P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C)$ in (15) is thus interpreted as the expected dispersion of $\frac{f(X|C)P(C)}{\hat{f}(X|C)\hat{P}(C)}$ around unity with respect to the approximating distribution $\hat{f}\hat{P}$.

The result above is the starting point of our development of approximations of probability to find plug-in classifiers for Bayesian networks and in particular to evaluate Naïve Bayes. We note that

$$E_{\hat{f}\hat{P}} \left| \frac{f(X|C)P(C)}{\hat{f}(X|C)\hat{P}(C)} - 1 \right| = \sum_{c=1}^{k} \int_{X} \hat{f}(x|c)\hat{P}(c) \left| \frac{f(x|c)P(c)}{\hat{f}(x|c)\hat{P}(c)} - 1 \right| d\mu(x)$$

$$= \sum_{c=1}^{k} \int_{X} \left| f(x|c)P(c) - \hat{f}(x|c)\hat{P}(c) \right| d\mu(x), \tag{16}$$

which is the bound in (15) written as in Ryzin (1966) and Glick (1972). We shall, next present examples of evaluating the bound directly.

Example 3 *Let again as in Example 1 take $\mathcal{X}$ as the binary hypercube in d dimensions. We assume that the true density (with respect to the counting measure μ) $f(x|c) > 0$ for all $x \in \mathcal{X}$ and $P(c) = \hat{P}(c)$. When we approximate this density by its first order term $f_1(x|c)$ in (4) we get from (9) and (10) that*

$$\left| f(x|c)P(c) - \hat{f}(x|c)\hat{P}(c) \right| = P(c) f_1(x|c) \left| \sum_{\mathbf{w} \neq \mathbf{0}} \beta_{\mathbf{w},c} U_{\mathbf{w},c}(x) \right|.$$

In words, here the bound in Theorem 5 expresses the deterioration of the classifier performance by means of a sum of all interactions of order higher than one. Since the measure μ is the counting measure we have the bound

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) \leqslant \sum_{c=1}^{k} P(c) \sum_{x \in X} f_1(x|c) \left| \sum_{\mathbf{w} \neq \mathbf{0}} \beta_{\mathbf{w},c} U_{\mathbf{w},c}(x) \right|.$$

Example 4 *The Kolmogorov distance of variation is very effectively evaluated and bounded for the class of two dimensional densities having an expansion with respect to an orthonormal system of polynomials. A diagonal expansion is possible, for example, for Gaussian, sinusoidal and Pearson type II distributions, see McGraw and Wagner (1968), which also contains an extensive list of references on the subject.*

We take as an illustration the two dimensional Gaussian density. Hence $X = (X_i)_{i=1}^{2}$, and $\mathcal{X} = \mathbb{R} \times \mathbb{R}$. The density $f(x)$ is determined by the respective variances σ_1^2 and σ_2^2, the respective means m_1 and m_2, and the coefficient of correlation ρ. Let $\hat{f}(x) = f_1(x_1) f_2(x_2)$ be the product of the two Gaussian marginal densities for X_1 and X_2. This corresponds again to an instance of the Naïve Bayes procedure. Then the classical Mehler expansion (Cramér, 1966) says for $|\rho| < 1$ that

$$f(x) = f_1(x_1) \cdot f_2(x_2) \cdot \sum_{n=0}^{\infty} H_n\left(\frac{x_1 - m_1}{\sigma_1}\right) \cdot H_n\left(\frac{x_2 - m_2}{\sigma_2}\right) \cdot \frac{\rho^n}{n!},$$

where $H_n(x)$ is the Hermite polynomial of order n, defined as $H_n(x) = (-1)^n e^{x^2} \frac{d^n}{x^n} e^{-x^2}$ for $n = 0, 1, \ldots,$. If we assume $P(c) = \hat{P}(c)$, then Theorem 5 entails, since $H_0(x) = 1$,

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C)$$
$$\leqslant \sum_{c=1}^{k} P(c) \int_{\mathbb{R}\times\mathbb{R}} f_1(x_1|c) f_2(x_2|c) \left| \sum_{n=1}^{\infty} H_n\left(\frac{x_1 - m_1(c)}{\sigma_1}\right) \cdot H_n\left(\frac{x_2 - m_2(c)}{\sigma_2}\right) \cdot \frac{\rho^n(c)}{n!} \right| dx_1 dx_2,$$

where the means and coefficient of correlation are chosen to depend on c. The bound on the difference between the probabilities of correct classification is seen to be a power series in the absolute value of the coefficient of correlation. There are computational routines for the Hermite polynomials, and in addition integrals of the form $\int_{\mathbb{R}} \mid x_i - m_i \mid^k f_i(x_i) dx_i$ involved here are explicitly computable. There is an extension of the Mehler expansion for n-variate densities (Slepian, 1972), which could be used in some of the examples below, but we will not expand on this due to the extensive notational machinery thereby required.

There are certain well known inequalities between the Kolmogorov distance of variation and other distances or divergences between probability measures. These distances are often readily computable in an explicit form, a compendium is recapitulated in Kailath (1967). An up-to-date discussion of the inequalities to be presented below and several others is found in Topsoe (2000). Nguyen et al. (2005) have presented techniques of replacing the Bayesian probability of error by more general risk functions and analyzing them with corresponding divergences, which are surveyed in Topsoe (2000).

For two probability densities f and $\hat{f}$ we have the inequality due to Ch. Kraft, see Hoeffding and Wolfowitz (1958); Pitman (1979),

$$\frac{1}{2} \int_{\mathcal{X}} \left| f(x) - \hat{f}(x) \right| d\mu(x) \leqslant \sqrt{1 - \left[\int_{\mathcal{X}} \sqrt{f(x) \cdot \hat{f}(x)} d\mu(x) \right]^2}. \tag{17}$$

The quantity $\int_{\mathcal{X}} \sqrt{f(x) \cdot \hat{f}(x)} d\mu(x)$ is known as the affinity or as the Bhattacharyya coefficient or as the Hellinger integral.

We note next that (17) yields in (16)

$$\sum_{c=1}^{k} \int_{\mathcal{X}} \left| f(x|c)P(c) - \hat{f}(x|c)\hat{P}(c) \right| d\mu(x)$$
$$\leqslant 2\sqrt{1 - \left[\sum_{c=1}^{k} \sqrt{P(c)\hat{P}(c)} \int_{\mathcal{X}} \sqrt{f(x|c) \cdot \hat{f}(x|c)} d\mu(x) \right]^2}. \tag{18}$$

The Kullback-Leibler divergence (in natural logarithm) (Kullback, 1997; Cover and Thomas, 1991) defined as

$$D\left(f,\hat{f}\right)=\int_{\mathcal{X}} f(x|c)\log\left(\frac{f(x|c)}{\hat{f}(x|c)}\right)d\mu(x).$$

We have

$$-\frac{1}{2}D(f,\hat{f})=\int_{\mathcal{X}} f(x|c)\log\left(\frac{\hat{f}(x|c)}{f(x|c)}\right)^{\frac{1}{2}}d\mu(x).$$

By Jensen's inequality

$$\leqslant \log\int_{\mathcal{X}} f(x|c)\left(\frac{\hat{f}(x|c)}{f(x|c)}\right)^{\frac{1}{2}}d\mu(x)=\log\int_{\mathcal{X}}\sqrt{f(x|c)\hat{f}(x|c)}d\mu(x).$$

Hence

$$\int_{\mathcal{X}}\sqrt{f(x|c)\hat{f}(x|c)}d\mu(x)\geqslant e^{-\frac{1}{2}D(f,\hat{f})}.$$

Hoeffding and Wolfowitz (1958) were probably the first to observe this inequality. Furthermore,

$$\sqrt{1-\left[\int_{\mathcal{X}}\sqrt{f(x)\cdot\hat{f}(x)}d\mu(x)\right]^2}\leqslant\sqrt{1-e^{-D(f,\hat{f})}}. \tag{19}$$

In hypothesis testing and pattern classification it is desirable that $D(f_1,f_2)$ is large, or, the affinity is small. The opposite is desirable for plug-in classifiers. By symmetry of the affinity in (19) we get ($D(f,\hat{f})$ need not be equal to $D(\hat{f},f)$) that

$$\sqrt{1-\left[\int_{\mathcal{X}}\sqrt{f(x)\cdot\hat{f}(x)}d\mu(x)\right]^2}\leqslant\sqrt{1-e^{-D(\hat{f},f)}}.$$

Brown (1959) and Ku and Kullback (1969) developed a convergent iteration that finds $\hat{f}$ minimizing $D(\hat{f},f)$ in the class of all densities on discrete $\mathcal{X}$ that have some given set of lower order marginals. The iteration may in several situations be computationally infeasible without constraining f to some suitably simplified model family.

Example 5 *In Example 3 we get by (17) the bound*

$$P(\hat{c}_B(X)=C)-P(\hat{c}_{\hat{B}}(X)=C)\leqslant 2\sum_{c=1}^{k}P(c)\sqrt{1-\left[\sum_{x\in\mathcal{X}}f_1(x|c)\sqrt{\sum_{\mathbf{w}\neq\mathbf{0}}\beta_{\mathbf{w},c}U_{\mathbf{w},c}(x)}\right]^2}.$$

Since we shall need the generic formula in the sequel, we note that the Kullback-Leibler divergence involved in this context is for discrete $\mathcal{X}$

$$D\left(f(x|c)P(c),f_1(x|c)\hat{P}(c)\right)=\sum_{c=1}^{k}\sum_{x\in\mathcal{X}}f(x|c)P(c)\log\frac{f(x|c)P(c)}{f_1(x|c)\hat{P}(c)}$$

$$=\sum_{c=1}^{k}P(c)\sum_{x\in\mathcal{X}}f(x|c)\log\frac{f(x|c)}{f_1(x|c)}+\sum_{c=1}^{k}P(c)\log\frac{P(c)}{\hat{P}(c)} \tag{20}$$

$$= \sum_{c=1}^{k} P(c) D\left(f(x|c), f_1(x|c)\right) + D\left(P(c), \hat{P}(c)\right).$$

Example 6 *We continue with Example 3 but omit the assumption that $P(c) = \hat{P}(c)$. We consider the plug-in classifier with the first order term $f_1(x|c)$ in (4). In view of (11) we get*

$$\begin{aligned} \sum_{x \in \mathcal{X}} f(x|c) \log \frac{f(x|c)}{f_1(x|c)} &= \sum_{x \in \mathcal{X}} f(x|c) \sum_{\mathbf{w} \in \{0,1\}^d} \alpha_{\mathbf{w},c} U_{\mathbf{w},c}(x) \\ = \sum_{\mathbf{w} \in \{0,1\}^d} \alpha_{\mathbf{w},c} \sum_{x \in \mathcal{X}} f(x|c) U_{\mathbf{w},c}(x) &= \sum_{\mathbf{w} \in \{0,1\}^d} \alpha_{\mathbf{w},c} E_f\left[U_{\mathbf{w},c}(X)\right] \\ &= \sum_{\mathbf{w} \in \{0,1\}^d} \alpha_{\mathbf{w},c} \beta_{\mathbf{w},c}, \end{aligned} \tag{21}$$

where we evoked (8). Hence we have by (19) and (20) obtained the following bound for the performance of the Naïve Bayes classifier for binary feature vectors

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) \leqslant 2\sqrt{1 - e^{-\left\{\sum_{c=1}^{k} P(c) \sum_{\mathbf{w} \in \{0,1\}^d} \alpha_{\mathbf{w},c}\beta_{\mathbf{w},c} + \sum_{c=1}^{k} P(c) \log \frac{P(c)}{\hat{P}(c)}\right\}}}.$$

For f in a kth order Bahadur class in (9) we can often, at least for relatively low d, readily evaluate the bounds above. An observation concerning the expression obtained in (21) is that $U_{\mathbf{0},c} = 1$ gives by (8) that $\beta_{\mathbf{0},c} = 1$, and by (12) that

$$\alpha_{\mathbf{0},c}\beta_{\mathbf{0},c} = -D\left(f_1(x|c), f(x|c)\right).$$

Example 7 *In Example 2 above the true and plug-in densities correspond to the distributions $N(H\mathbf{b}, \Sigma)$ and $N(M\mathbf{b}, \Sigma)$, respectively. Since $\mathcal{C} \subseteq \{-1, +1\}^N$ is a codebook of equally likely vectors $\{\mathbf{b}\}$, we modify the general notation for this example by denoting a label in $\mathcal{C}$ by $\mathbf{b}$. Hence $\hat{P}(\mathbf{b}) = P(\mathbf{b}) = \frac{1}{|\mathcal{C}|}$, where $|\mathcal{C}|$ is the cardinality of the codebook.*

In this example we use the bound (18). The expression for the required affinity is well known (Kailath, 1967) and equals

$$\int_{\mathbb{R}^d} \sqrt{f(x|\mathbf{b}) \cdot \hat{f}(x|\mathbf{b})} d\mu(x) = e^{-\frac{1}{4} D(N(H\mathbf{b},\Sigma), N(M\mathbf{b},\Sigma))}, \tag{22}$$

where $D(N(H\mathbf{b}, \Sigma), N(M\mathbf{b}, \Sigma))$ is in fact the Kullback-Leibler divergence

$$D(N(H\mathbf{b}, \Sigma), N(M\mathbf{b}, \Sigma)) = \frac{1}{2}\left((H - M)\mathbf{b}\right)^T \Sigma^{-1} \left((H - M)\mathbf{b}\right), \tag{23}$$

see Kullback (1997)). Therefore we obtain for the plug-in classifier (suboptimal detector) defined in Example 2 by (18) that

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) \leqslant 2\sqrt{1 - \frac{1}{|\mathcal{C}|} \sum_{\mathbf{b} \in \mathcal{C}} e^{-\frac{1}{8}((H-M)\mathbf{b})^T \Sigma^{-1}((H-M)\mathbf{b})}}. \tag{24}$$

This expression can be used to compare different designs of suboptimal detectors represented by their respective matrices M.

We want a way to calculate $P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C)$ in terms of (only) $\hat{f}(x|c)\hat{P}(c)$. We will generalize the result in Theorem 5 (15) in the sense that it can be used when only certain parts in a factorization of $\hat{f}(x|c)$ are approximations. What we mean by 'components' will be made precise later.

6. Approximating Bayesian Networks in Classification

While (14) is an exact expression of $P(\hat{c}_B(X)=C)-P(\hat{c}_{\hat{B}}(X)=C)$ it might be infeasible to calculate it in practice. Thus we introduce more easily computable bounds in Corollary 8 and Theorem 11. To avoid making these approximations, or bounds upwards too loose, we also take into account the case where we have not approximated all of $P(c|x)$ (Theorem 7). If the factor of $P(c|x)$ that is not approximated is not functionally dependent of the factor that is approximated, we can develop sharper bounds on the degradation of the approximated classifier performance. Here we consider a class of general approximations, applicable to Bayesian networks (Cowell et al., 1999). For ease of reference we recapitulate first the definition of Bayesian networks.

Definition 6 *Given a directed acyclic graph $G=(V,E)$ with $\{X_i\}_{i=1}^{d}$ designating (the random variables at) the vertices, Π_i denotes the set of parents of X_i in the graph G. We use π_i to denote the parents states. The pair (G,P), is called a Bayesian network and satisfies*

$$f(x|c,G)=\prod_{i=1}^{d} f(x_i|\pi_i,c,G). \tag{25}$$

Williamson (2005) suggests the following method of approximation. $f(x|c)$ denotes a generic target probability mass function and G is a directed acyclic graph. In principle one can compute the probabilities $f(x_i|\pi_i,c)$ on G using f. Then $\hat{f}(x|c)$ is an approximating probability obtained by taking

$$\hat{f}(x_i|\pi_i,c,G)=f(x_i|\pi_i,c), \tag{26}$$

and multiplying $\hat{f}(x_i|\pi_i,c,G)$ according to (25). The best approximating G (in some family of directed acyclic graphs) is found by maximizing

$$\sum_{x\in\mathcal{X}} f(x|c)\sum_{i=1}^{d}\log\frac{f(x_i,\pi_i|c)}{f(x_i|c)f(\pi_i|c)},$$

which is shown to minimize $D\left(f,\hat{f}\right)$. This constitutes also a method of learning network structures, in case there is an effective algorithmic implementation.

We give next a few examples of Bayesian networks, which will also be used to illustrate some of the results in the sequel.

Example 8 *As in Example 1 we take $\mathcal{X}=\{0,1\}^d$ and assume that G is a rooted tree. We order the variables so that x_1 is the state at the root. Direction is defined from parent to child. In a rooted tree any node i, except for the root, has exactly one parent node $\pi(i)<i$. We write the parent state as $\pi_i=x_{\pi(i)}$. Then the factorization in (25) becomes*

$$f(x|c,G)=f(x_1|c)\prod_{i=2}^{d} f(x_i|x_{\pi(i)},c,G). \tag{27}$$

In other words this is a joint probability factorized along a rooted tree. This is in the sense of Lewis (1959), as discussed above, a product approximation of a density with d variables with at most two components per factor. We can also talk about a tree dependence. This dependence was introduced in Chow and Liu (1968).

We will use the form in Definition 6, (25) to express partial approximations. Let $S = (S_1,\dots,S_4)$ be a partition of $\{X_i\}_{i=1}^{d}$, where $s = (s_1,\dots,s_4)$ denotes the resulting partition of $\{x_i\}_{i=1}^{d}$.

We designate by $f(s_i|G)$ the class conditional density of all s.v.'s that are in S_i given its parents, that is $f(s_i|G)$ is short notation for

$$\prod_{\{i|X_i \in S_i\}} f(x_i|\pi_i,c,G).$$

When referring to the range of S_i we use $\mathcal{S}_i$. Next we describe an efficient choice of S. We make some relevant definitions.

X_i is an proper ancestor of X_j in G and X_j is a proper descendent of X_i in G if there exist a path from X_i to X_j in G for $i \neq j$. A path is a sequence $A_0,...,A_n$ of distinct vertices such that $(A_{i-1},A_i) \in E$ for all $i = 1,...,n$. Given a Bayesian network (G,P) and $\hat{P}$, the partition S is defined for a class conditional density as follows:

- $X_i \in S_1$ if $f(x_i|\pi_i,c) \neq \hat{f}(x_i|\pi_i,c)$.
- $X_i \in S_2$ if for all x_i,π_i we have $f(x_i|\pi_i,c) = \hat{f}(x_i|\pi_i,c)$ and for all $j \neq i$ such that $X_j \in S_1$ we have $X_i \notin \pi_j$.
- $X_i \in S_3$ if for all x_i,π_i, we have $f(x_i|\pi_i,c) = \hat{f}(x_i|\pi_i,c)$, there exists $j \neq i$ such that $X_j \in S_1$ and $X_i \in \pi_j$. Furthermore no proper ancestor X_k of X_i in $G_{S \setminus S_2}$ is such that $X_k \in S_1$.
- $X_i \in S_4$ if $X_i \notin S_1$, $X_i \notin S_2$ and $X_i \notin S_3$.

Example 9 *We consider the rooted and directed tree in the Example 8. We approximate the joint density in (27) by the product of marginal densities. This corresponds to removing all the edges from the tree, the resulting set of nodes is a degenerate special case of a DAG.*

Then $S_1 = \{2,3,\dots,d\}$ and $S_3 = \{1\}$ = the root. The partitioning sets S_2 and S_4 are empty.

Example 10 *Context-Specific Independence in Bayesian networks (Boutilier et al., 1996). In this example $X_i \in \{0,1\}$ and the graph for the Bayesian network is as in Figure 1. Then X_9 is a context in the sense that*

$$f(x_1|x_5,\dots,x_9) = \begin{cases} f(x_1|x_5,x_6) & x_9 = 0 \\ f(x_1|x_7,x_8) & x_9 = 1 \end{cases}. \tag{28}$$

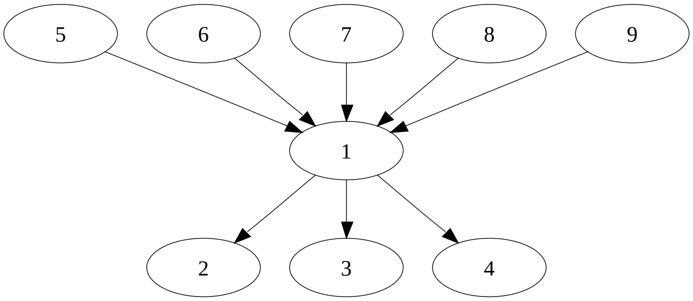
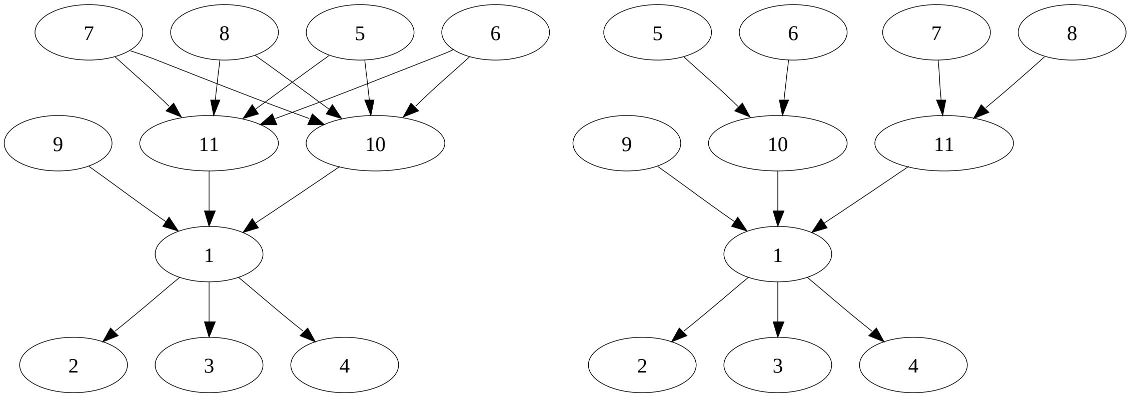


Figure 1: Original Bayesian network



(a) Transformed Bayesian network

(b) Transformed Bayesian network with the CSI assumption

Figure 2: Transformed Bayesian networks with and without the CSI assumption

To encode this in a Bayesian network we transform the original Bayesian network into the network in Figure 2(a). Figure 2(b) describes the graph, where the assumption in (28) holds, where $f(x_{10}|x_5,x_6) = f(x_1|x_5,x_6)$, $f(x_{11}|x_7,x_8) = f(x_1|x_7,x_8)$ *and*

$f(x_1\|x_9,x_{10},x_{10})$	x_9	x_{10}	x_{11}
0	*0*	*0*	*0*
0	*0*	*0*	*1*
1	*0*	*1*	*0*
1	*0*	*1*	*1*
0	*1*	*0*	*0*
1	*1*	*0*	*1*
0	*1*	*1*	*0*
1	*1*	*1*	*1* .

If the context specific assumption is introduced as an approximation this would yield

$$\begin{cases} S_1 = \{X_{10}, X_{11}\} \\ S_2 = \{X_1, X_2, X_3, X_4, X_9\} \\ S_3 = \{X_5, X_6, X_7, X_8\} \\ S_4 = \{\varnothing\} \end{cases} .$$

Example 11 *In this example we depict a graph G given a partition s, with some abuse of notation. In Figure 3 if $X_i \in S_1$ we label the vertex X_i as* 1.

Theorem 7 $\int_{\mathcal{X}} \left|P(c)f(x|c) - \hat{P}(c)\hat{f}(x|c)\right| d\mu(x) =$

$$\int_{\mathcal{S}_3} \int_{\mathcal{S}_1 \times \mathcal{S}_4} \left|P(c)f(s_1 \times s_4|G) - \hat{P}(c)\hat{f}(s_1 \times s_4|G)\right| d\mu(s_1 \times s_4) df(s_3|G).$$

Proof We use S to write $\int_{\mathcal{X}} \left|P(c)f(x|c) - \hat{P}(c)\hat{f}(x|c)\right| d\mu(x)$ as

$$= \int_{\mathcal{X}} \left|P(c)f(s_1|G) - \hat{P}(c)\hat{f}(s_1|G)\right| \left[\prod_{j=2}^{4} f(s_j|G)\right] d\mu(x). \qquad (29)$$

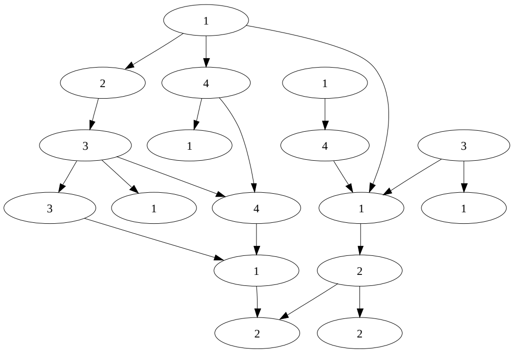


Figure 3: Bayesian network

Now we use the definition of S_2 and the Fubini theorem to write (29) as

$$\int_{\mathcal{S}_1\times\mathcal{S}_3\times\mathcal{S}_4}\left|P(c)f(s_1|G)-\hat{P}(c)\hat{f}(s_1|G)\right|\int_{\mathcal{S}_2}\left[\prod_{j=2}^{4}f(s_j|G)\right]d\mu(s_2)d\mu(s_1\times s_4)d\mu(s_3). \tag{30}$$

We can express the innermost integral as

$$\int_{\mathcal{S}_2}f(s_2\times s_3\times s_4|s_1,G)d\mu(s_2)=f(s_3\times s_4|s_1,G).$$

We continue with (30). Since for all $X_i\in S_3$ there exists no $X_j\in S_1\bigcup S_4$ such that $X_j\in\pi_i$ we can write this as

$$\int_{\mathcal{S}_3}f(s_3|G)\int_{\mathcal{S}_1\times\mathcal{S}_4}\left|P(c)f(s_1\times s_4|G)-\hat{P}(c)\hat{f}(s_1\times s_4|G)\right|d\mu(s_1\times s_4)d\mu(s_3).$$

■

Since this result is an equality, it seems to indicate that isolated approximations are stable in the sense that the classification error they introduce does only depend on a local neighborhood in the original Bayesian network.

There exist several algorithms that can be used for finding an approximation of a BN such as the ones described in Chow and Liu (1968) and Chickering (2002). If an approximation has been constructed, Theorem 7 makes it possible to evaluate its effect depending on the approximations made, rather than on the original problem.

Next we extend the result in Theorem 5, (15) by specifying the difference in probability of correct classification in terms of the partial structure specific difference through combining Theorem 7 and Theorem 5.

Corollary 8

$$P(\hat{c}_B(X)=C)-P(\hat{c}_{\hat{B}}(X)=C)\leqslant$$
$$\sum_{c=1}^{k}\int_{\mathcal{S}_3}\int_{\mathcal{S}_1\times\mathcal{S}_4}\left|P(c)f(s_1\times s_4|G)-\hat{P}(c)\hat{f}(s_1\times s_4|G)\right|d\mu(s_1\times s_4)df(s_3|G). \tag{31}$$

Of course, it might still be computationally difficult to calculate the bound in Corollary 8. When combining (31), (17) and (19) we obtain the following corollary.

Corollary 9 *If $\hat{P}(c) = P(c)$ for all $c \in \mathcal{C}$,*

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) \leqslant$$

$$2\sum_{c=1}^{k} P(c) \int_{\mathcal{S}_3} \sqrt{1 - e^{-D\left(f(s_1 \times s_4|c,G), \hat{f}(s_1 \times s_4|c,G)\right)}} d f(s_3|G).$$

The following examples demonstrate computable expressions for this bound.

Example 12 *We consider the approximation in Example 9, where $S_1 = \{2,3,\ldots,d\}$ and $S_3 = \{1\} =$ the root, and the other partitioning sets are empty. We have from (27)*

$$f(s_1|c,G) = \prod_{i=2}^{d} f(x_i|x_{\pi(i)}, c, G),$$

which is a probability mass function on S_1 and

$$\hat{f}(s_1|c,G) = \prod_{i=2}^{d} f(x_i|c,G),$$

which is a probability mass function on S_1. Then in view of (19) and (20) we compute

$$D\left(f(s_1|c,G), \hat{f}(s_1|c,G)\right) = \sum_{s_1 \in S_1} f(s_1|c,G) \log \frac{f(s_1|c,G)}{\hat{f}(s_1|c,G)}$$

$$= \sum_{i=2}^{d} \sum_{x_i, x_{\pi(i)}} f\left(x_i, x_{\pi(i)}|c,G\right) \log \frac{f\left(x_i, x_{\pi(i)}|c,G\right)}{f\left(x_i|c,G\right) \cdot f\left(x_{\pi(i)}|c,G\right)}.$$

Here we recognize, see Cover and Thomas (1991), the mutual informations $I_{c,G}\left(x_i, x_{\pi(i)}\right)$ between x_i and $x_{\pi(i)}$ so that the expression in the right hand side of the preceding equation equals

$$= \sum_{i=2}^{d} I_{c,G}\left(x_i, x_{\pi(i)}\right).$$

It should be noted that this depends on $S_3 = \{1\}$ through those nodes i that have $\pi(i) = 1$. Chow and Liu (1968) developed an algorithm for finding the tree from data that maximizes sum of the mutual informations between a variable and its parents shown above.

Example 13 *The conditionally Gaussian regressions are useful probability models for Bayesian networks with both continuous and discrete variables, see Lauritzen (1990). In order to fit the framework above to these distributions, we suppose that (X_1, X_4) is a vector of r continuous random variables, and that the variables in X_3 are decomposed into the discrete ones in $X_3(\triangle)$ and into the t continuous variables $X_3(\gamma)$. In order not to overburden the notation we omit here the dependence on c in the expressions below.*

The conditionally Gaussian regressions are defined as follows. The conditional distribution of (X_1,X_4) *given* X_3 *is a multivariate normal distribution*

$$(X_1,X_4) \mid \left(X_3(\triangle)=\pi_\triangle, X_3(\gamma)=\pi_\gamma\right) \sim N_r\left(A(\pi_\triangle)+B(\pi_\triangle)\pi_\gamma, \Sigma(\pi_\triangle)\right),$$

where $\pi_\triangle$ *the state of the discrete parents,* π_γ *is the state of the continuous parents, and* $A(\pi_\triangle)$ *is a* $r\times 1$ *vector,* $B(\pi_\triangle)$*, is an* $r\times t$ *matrix,* $\Sigma(\pi_\triangle)$ *is a positive definite symmetric matrix. The Naïve Bayes classifiers approximate* $\Sigma(\pi_\triangle)$ *by a diagonal matrix.*

We illustrate, however, the simplest of the upper bounds with the Kullback-Leibler divergence by taking the plug-in distribution

$$\hat{N}_r\left(A+B(\pi_\triangle)\pi_\gamma, \Sigma(\pi_\triangle)\right).$$

Similarly to what has been done above, or more precisely, using Theorem 7, (18), and (22) above it follows that for $X=(X_i)_{i=1}^4$ *corresponding to the decomposition* $S=(S_1,\ldots,S_4)$

$$P(\hat{c}_B(X)=C)-P(\hat{c}_{\hat{B}}(X)=C)\leqslant$$

$$2\int_{\mathcal{S}_3}\left[\sqrt{1-\left(\sum_{c=1}^{k}\sqrt{P(c)\hat{P}(c)}e^{-\frac{1}{8}\left[\left(A(\pi_\triangle)-A\right)^T\Sigma^{-1}(\pi_\triangle)\left(A(\pi_\triangle)-A\right)\right]}\right)^2}\right]df(s_3|G),$$

where, as noted above, some of the dependencies on c are not explicitly accounted for. Here $f(s_3|G)$ *is not in general a Gaussian density, as* $\mathcal{S}_3$ *may, for example, involve discrete states.*

A way of further simplification of the bound in Corollary 8 is to bound the density upwards by the following quantity.

Definition 10 *Let*

$$\varepsilon(c) := \max_{s_1,s_3,s_4}\left|P(c)f(s_1\times s_4|G)-\hat{P}(c)\hat{f}(s_1\times s_4|G)\right|.$$

With this quantity we can simplify the computation of the bound in Theorem 8.

Theorem 11 $P(\hat{c}_B(X)=C)-P(\hat{c}_{\hat{B}}(X)=C)\leqslant\sum_{c=1}^k\varepsilon(c)\mu(s_1\times s_4).$

Proof From Corollary 8 (31) we have that $P(\hat{c}_B(X)=C)-P(\hat{c}_{\hat{B}}(X)=C)\leqslant$

$$\sum_{c=1}^{k}\int_{\mathcal{S}_3}\int_{\mathcal{S}_1\times\mathcal{S}_4}\left|P(c)f(s_1\times s_4|G)-\hat{P}(c)\hat{f}(s_1\times s_4|G)\right|d\mu(s_1\times s_4)df(s_3|G).$$

$$\leqslant\sum_{c=1}^{k}\max_{s_3}\left[\int_{\mathcal{S}_1\times\mathcal{S}_4}\left|P(c)f(s_1\times s_4|G)-\hat{P}(c)\hat{f}(s_1\times s_4|G)\right|d\mu(s_1\times s_4)\right].$$

We finish by using the definition of $\varepsilon(c)$, which yields

$$\leqslant\sum_{c=1}^{k}\varepsilon(c)\int_{\mathcal{S}_1\times\mathcal{S}_4}d\mu(s_1\times s_4)=\sum_{c=1}^{k}\varepsilon(c)\mu(s_1\times s_4).$$

■

In the next section, 7, we will be able to use the bound in Theorem 11 as a way of motivating the 'Naïve Bayes' approximation.

7. Bounding the Kolmogorov Distance of Variation with Respect to Naïve Bayes

In this section we assume that the feature space is finite and discrete, that is, $\mathcal{X} = \times_{i=1}^{d} \mathcal{X}_i$ and $\mathcal{X}_i = \{1, \ldots, r_i\}$. A popular plug-in classifier is the Nav̈e Bayes, already defined for three special cases in Examples 1, 4 and 13.

Definition 12 *A Naïve Bayes plug-in classifier is a classifier that assumes that the features of X are independent given c,*

$$\hat{f}(x|c) = \prod_{i=1}^{d} f(x_i|c).$$

In order not to overburden the notation we avoid symbols like $f_{X_i}(x_i|c)$ for marginal densities. In spite of this we are not restricted to the case where all marginal densities are identical.

There are several practical reasons for the popularity of the Naïve Bayes. For example, Toussaint (1972) has shown that if $\mathcal{X}_i$ is the same for every i, then $\hat{f}(x|c)$ is a polynomial.

There exist statistical tests for whether independence holds or not. In practice, however, we often exclude independence of features by domain knowledge.

When $c \in \{1,2\}$ (that is $k = 2$) and we have n samples $(x,c)^{(n)} = \{(x,c)_l\}_{l=1}^{n}$, we define $\hat{c}_{ERM}(x|x^{(n)})$ as the classifier that minimizes the empirical error on this sample. Without the Naïve Bayes assumption we can use bounds such as the following in (Devroye et al., 1996, Page 462) (for $k = 2$)

$$E\left[P\left(\widehat{c}_{ERM}\left(X|(X,C)^{(n)}\right) \neq C|(X,C)^{(n)}\right)\right] \leqslant P(\hat{c}_B(X) \neq c) + \varepsilon_1,$$

but with Naïve Bayes we have $0 \leqslant \varepsilon_2 \leqslant \varepsilon_1$ such that for $k = 2$ see Devroye et al. (1996, Chapter 27.3)

$$E\left[P\left(\widehat{c}_{ERM}\left(X|(X,C)^{(n)}\right) \neq C|(X,C)^{(n)}\right)\right] \leqslant P(\hat{c}_B(X) \neq c) + \varepsilon_2.$$

The Naïve Bayes assumption for specific data sets can actually perform better than a plug-in classifier incorporating some dependencies as shown in Titterington et al. (1981). In Friedman et al. (1997) Naïve Bayes has been reported as performing worse than taking dependence into account (but not on all data sets), and even then the difference was in many cases not large. In Huang et al. (2003) it is found as suboptimal in most data sets. A more in-depth Meta study on the subject is Hand and Yu (2001).

Our own experience does not speak against the conjecture that the advantage of taking dependence into account may depend on the context (Ekdahl, 2006, Section 7). Instead of running yet another simulation or arguing for or against on ad hoc grounds, we expand the existing theory around Naïve Bayes. The motivation for this can be intuitively outlined as follows.

Let us assume $P(C) = \hat{P}(C)$. Then the Kolmogorov variation distance is

$$\frac{1}{2} E_{\hat{f},\hat{P}} \left| \frac{f(X|C)}{\hat{f}(X|C)} - 1 \right|,$$

where now $\hat{f}(X|C)$ is as in Definition 12. As the level of dependence between the components in X increases, the dispersion of $\frac{f(X|C)}{\hat{f}(X|C)} - 1$ increases. In the case $d = 2$, or $X = (X_i)_{i=1}^{2}$ the Kolmogorov variation distance with respect to the product of marginal densities was first studied by Hoeffding (1942), who discovered that there is an upper bound for the distance, which is assumed when one of X_1 or X_2 is a function of the other. This would seem to restrict the effectiveness of the Naïve Bayes classifier to those situations, where the degree of dependence between the components of X is moderate, as was to be expected. But to get a more diverse view we recall some of Vilmansen (1971) and Vilmansen (1973).

The Kolmogorov variation distance measuring the degree of association between X and C is in Vilmansen (1971, 1973) defined as

$$K(X,C) = \frac{1}{2}\sum_{c=1}^{k}\sum_{x\in\mathcal{X}} |f(x,c) - f(x)P(c)|.$$

The maximum of this distance is given by the next inequality, or

$$K(X,C) \leqslant 1 - \sum_{c=1}^{k} P^2(c). \tag{32}$$

This is shown in Vilmansen (1973), and is also found in Hoeffding (1942).

The maximum of $K(X,C)$ in (32) is obtained, as soon as there is a functional dependence between X and C in the sense that the supports of the class-conditional densities are disjoint subsets of $\mathcal{X}$. Hence, in the last mentioned case the probability of correct classification is one. This extreme case is approached when the dependence between X and C is “very close“ to being functional, in the sense that each $f(x|c)$ is concentrated around its mode, say m_c, so that observation of m_c is an almost noise-free message, as it were, from the source c. In such a situation it should be possible for the Naïve Bayes classifier to perform well, too. The impact of strong dependence between the labels and feature vectors for Naïve Bayes has been argued in a different manner in Rish et al. (2001).

The following theorem shows in a more precise fashion how the modes of the densities $f(x|c)$ control the difference between the pertinent probabilities of the correct decision. We do not really need unimodality for our proofs, but this is a natural assumption for model based classification and simplifies the statements. In words the result in Theorem 13 below tells that, if the class conditional probability densities are predominantly well concentrated, the Kolmogorov variation distance with respect to Naïve Bayes is small.

Theorem 13 *Assume $f(x|c)$ is unimodal for every $c \in \mathcal{C}$. Let for any $c \in \mathcal{C}$,*

$$m_c := \arg\max_{x\in\mathcal{X}} f(x|c).$$

Then we have for the Naïve Bayes plug-in that

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) \leqslant \sum_{c=1}^{k} P(c)\max\left(f(m_c|c) - f(m_c|c)^d, 1 - f(m_c|c)\right). \tag{33}$$

Proof We shall simplify notation without risk confusion by writing $f(x|c)$ as $f(x)$. We state first some elementary observations. By the chain rule

$$\begin{aligned} f(x) &= f(x_i)f(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n|x_i) \\ &\leqslant f(x_i), \text{for all } i, \end{aligned} \tag{34}$$

which implies that

$$f(x)^d = \prod_{i=1}^{d} f(x) \leqslant \prod_{i=1}^{d} f(x_i). \tag{35}$$

The claim to be established will follow if we can show that for all $x \in \mathcal{X}$

$$\left| f(x) - \prod_{i=1}^{d} f(x_i) \right| \leqslant \max\left(f(m) - f(m)^d, 1 - f(m) \right). \tag{36}$$

The proof of (36) is divided into three cases

1 Suppose that $x = m$.

1.1 If $f(m) \geqslant \prod_{i=1}^{d} f(m_i) \Rightarrow f(m) - \prod_{i=1}^{d} f(m_i) \leqslant f(m) - f(m)^d$ by (35).

1.2 If $f(m) < \prod_{i=1}^{d} f(m_i) \Rightarrow \prod_{i=1}^{d} f(m_i) - f(m) \leqslant 1 - f(m)$.

2 Consider next $x \neq m$. We have $\left| f(x) - \prod_{i=1}^{d} f(x_i) \right|$

$$= \max\left(f(x), \prod_{i=1}^{d} f(x_i) \right) - \min\left(f(x), \prod_{i=1}^{d} f(x_i) \right).$$

Since both $max(a_1, a_2)$ and $min(a_1, a_2)$ are positive for $0 \leq a_1, a_2 \leq 1$

$$\leqslant \max\left(f(x), \prod_{i=1}^{d} f(x_i) \right) \leqslant \max\left(f(x), f(x_j) \right),$$

where $f(x) \leqslant \sum_{z \neq m} f(z) = 1 - f(m)$. Here j is chosen so that $x_j \neq m_j$, which exists since $x \neq m$. By (34), $f(m_j) \geqslant f(m)$ we obtain

$$f(x_j) \leqslant \sum_{x_i \neq m_j} f(x_i) = 1 - f(m_j) \leqslant 1 - f(m).$$

The inequality (36) and Theorem 5 imply (33), and thus the assertion in Theorem 13 is established as claimed. ■

The inequality (36) is an improvement of a result in Rish et al. (2001). We have constructed a tighter upper bound for $\left| f(x) - \prod_{i=1}^{d} f(x_i) \right|$ than the one in Rish et al. (2001), which is recapitulated in the next theorem.

Theorem 14 *For all* $x \in \mathcal{X}, \left| f(x) - \prod_{i=1}^{d} f(x_i) \right| \leqslant d(1 - f(m))$.

The sharpness of Theorem 14 can be seen though a plot of $\max_{x \in \mathcal{X}} \left| f(x) - \prod_{i=1}^{d} f(x_i) \right|$ as a function of $\max_{x \in \mathcal{X}} f(x)$ in two and tree dimensions in Figure 4. The plots are for a binary hypercube ($\mathcal{X} = \{0,1\}^d$). The simulation tests several distributions and takes the worst one for each value of simulated $f(m)$ (the details of the simulation can be found in the simulation appendix).

It is possible to increase $f(m)$ dividing for example multimodal class conditional densities into many unimodal densities (Vilata and Rish, 2003), but a theoretical investigation of identification and interpretation of such a partition scheme is lacking.

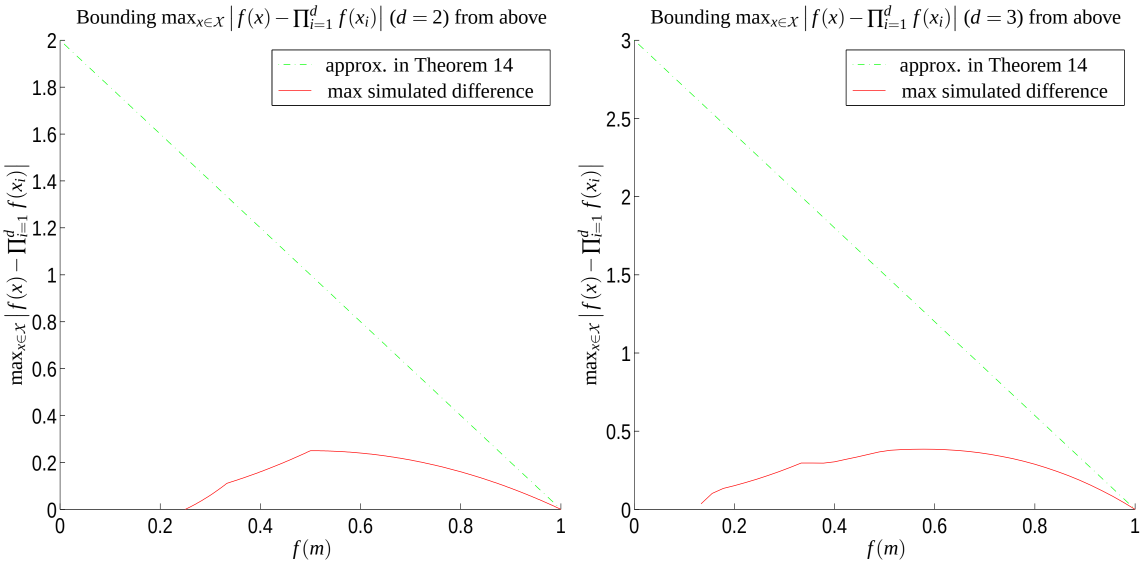


Figure 4: Illustration of the bound in Theorem 14.

Next we verify that the inequality (36) is in fact an improvement of Theorem 14, that is,

$$\max\left(f(m) - f(m)^d, 1 - f(m)\right) \leqslant d(1 - f(m)).$$

It is enough to show that $1 - f(m) \leqslant d(1 - f(m))$ **and** $f(m) - f(m)^d \leqslant d(1 - f(m))$. Here $1 - f(m) \leqslant d(1 - f(m))$ follow since $1 - f(m) \geqslant 0$ and $d \geqslant 2$. The remaining inequality can be shown using Bernoulli's inequality, so that

$$f(m) - f(m)^d = f(m) - (1 - (1 - f(m)))^d \leqslant f(m) - (1 - d(1 - f(m))) \leqslant d(1 - f(m)). \quad (37)$$

As with Theorem 14 we plot $\max_{x\in\mathcal{X}} \left| f(x) - \prod_{i=1}^{d} f(x_i) \right|$ as function of $f(m)$ in two and three dimensions and compare the maximal difference with the bounds in Theorem 14 and (36) (Figures 5 and 6).

From the three to five dimensional cases in Figures 5 and 6 we see that the inequality (36) is often sharp enough, if the probability density is concentrated, that is $f(m)$ is close to 1.

We give an additional upper bound ((39) below) readily derived from Theorem 13. We introduce the entropy (in natural logarithm)

$$H(X|C = c) = -\sum_{x\in\mathcal{X}} f(x|c) \ln f(x|c).$$

Then it is well known, see Arimoto (1971), that

$$1 - f(m_c|c) \leqslant \frac{H(X|C = c)}{\ln 2}. \quad (38)$$

Let us split $\mathcal{C}$ into two sets

$$\mathcal{C}_1 = \left\{c | f(m_c|c) - f(m_c|c)^d \geqslant 1 - f(m_c|c)\right\}$$

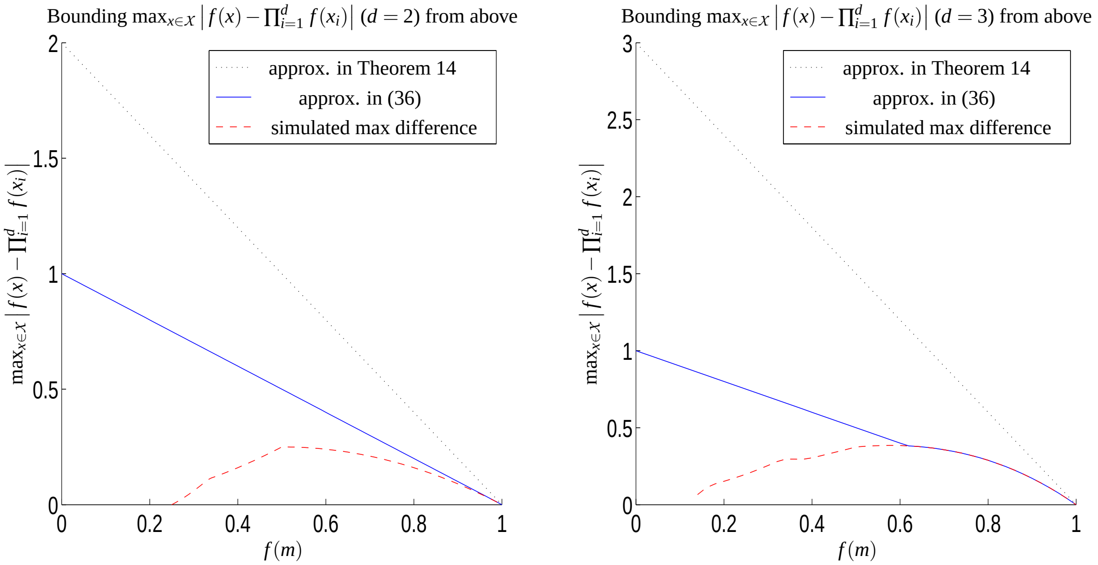


Figure 5: Illustration of the bounds in Theorem 14 and inequality (36) in two and three dimensions.

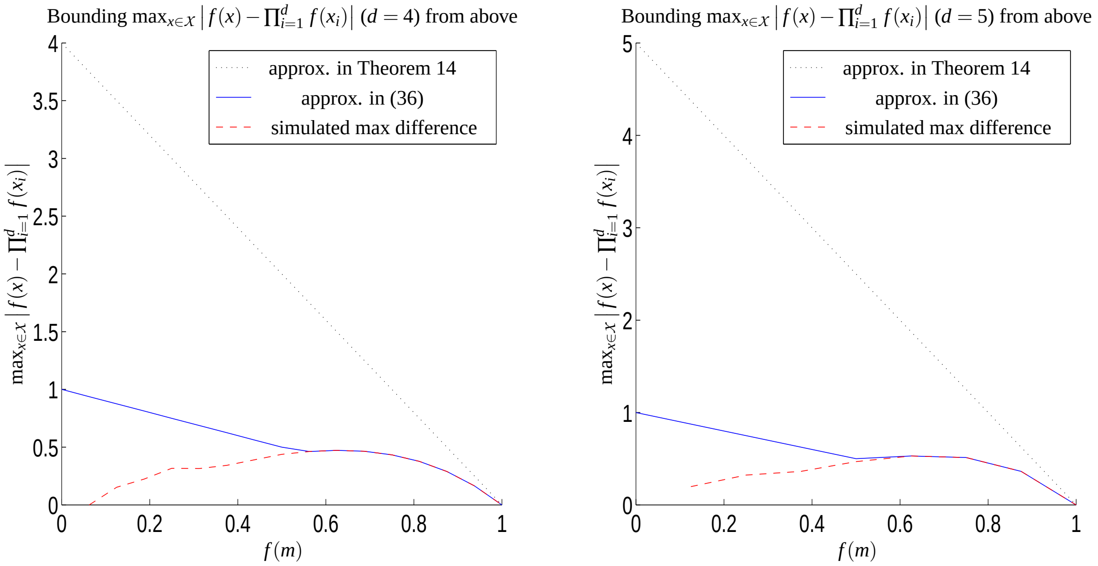


Figure 6: Illustration of the bounds in Theorem 14 and (36) in four and five dimensions.

and

$$\mathcal{C}_2 = \left\{ c | f(m_c|c) - f(m_c|c)^d < 1 - f(m_c|c) \right\}.$$

Thereby we get from (38)

$$\sum_{c=1}^{k} P(c) \max \left(f(m_c|c) - f(m_c|c)^d, 1 - f(m_c|c) \right)$$

$$\leqslant \sum_{\mathcal{C}_1} P(c) \left(f(m_c|c) - f(m_c|c)^d \right) + \sum_{\mathcal{C}_2} P(c) \frac{H(X|C=c)}{\ln 2}.$$

It was shown above, see (37), that $f(m_c|c) - f(m_c|c)^d \leqslant d\,(1 - f(m_c|c))$. We introduce the conditional entropy

$$H(X|C) = \sum_{c=1}^{k} P(c) H(X|C=c).$$

Then we obtain

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) \leqslant (d-1) \left(1 - \sum_{\mathcal{C}_1} P(c) f(m_c|c) \right) + \frac{H(X|C)}{\ln 2}. \tag{39}$$

This bound needs not to be sharp in general, but it gives some additional insight. The first term in the right hand side is of similar form as the upper bound in (32) and can therefore be thought as measuring the degree of functional dependence between X and C. The entropy $H(X|C)$ is known as equivocation and measures the uncertainty about X if C is observed (Cover and Thomas, 1991).

Corollary 8 can be combined with the inequality (36). We will do that in the same way as in Section 6, that is we will allow for a partial Naïve Bayes assumption in the following sense

$$f(x|c) = f(s_2 \times s_3 \times s_4 | c, s_1) \prod_{\{i | X_i \in S_1\}} f(x_i|c)$$

leading to the abridged notation

$$m_1 \times m_4 := \arg \max_{s_1, s_3, s_4 \in \mathcal{S}_1 \times \mathcal{S}_3 \times \mathcal{S}_4} \prod_{\{i | X_i \in S_1 \cup S_4\}} f(x_i | \pi_i, c, G).$$

Corollary 15 *Let $P(c) = \hat{P}(c)$, and take the partial Naïve Bayes on s_1 and s_4. Then*

$$P(\hat{c}_B(X) = \varsigma) - P(\hat{c}(X) = \varsigma)$$

$$\leqslant \sum_{c=1}^{k} P(c) \max \left(f(m_1 \times m_4 | G) - f(m_1 \times m_4 | G)^d, 1 - f(m_1 \times m_4 | G) \right) \prod_{\{i | X_i \in S_1 \cup S_4\}} r_i.$$

Example 14 *Let us take $k = 10$, and $d = 1000$. Most of the features are independent; however expert knowledge reveals that each class has exactly three features $(X_i, \ldots, X_{i+2})$ that depend on each other in accordance with the DAG in Figure 7. These features are very concentrated in the sense that there exists a vector a such that $P((X_i, \ldots, X_{i+2}) = a|c) > 0.995$. The expert explains that this it due to the fact that they correspond to a physical feature critical to each class. Which features have this dependence is, however, unknown and detection is complicated since*

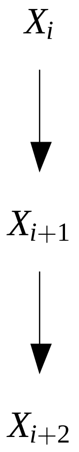


Figure 7: Graphical representation of dependence

$P(\text{there exists at least one } j \neq i \text{ such that} (X_j,\ldots,X_{j+2}|c)=a)$ *is large. Here it is possible to model with independence, since Corollary 15 then yields*

$$P(\hat{c}_B(X)=C)-P(\hat{c}_{\hat{B}}(X)=C)\leqslant\sum_{c=1}^{10}\varepsilon(c)\mu[\mathcal{S}_1\cup\mathcal{S}_4]=0.048.$$

8. Plug-In Classifiers that Make Optimal Decisions

In this section the plug-in classifier is not necessarily Naïve Bayes and $\mathcal{X}$ need not be discrete, although the conditions in the results below are more difficult to satisfy in continuous settings. As expounded above, it is easy to approximate $f(x|c)f(c)$, when the classes are well separated, in the sense there is almost a functional dependence between C and X. Sometimes it is possible to express the dependence by simply observing that

$$f(x|c)P(c)-f(x|\widetilde{c})P(\widetilde{c})$$

is large for all $x\in\mathcal{X}$ and all $c,\tilde{c}\in\mathcal{C}$ such that $c\neq\tilde{c}$. Here we present sufficient conditions for this pointwise separation between classes so that the probability of correct classification does not decrease by plugging in $\hat{f}(x|c)\hat{P}(c)$. The question is, how close must $\hat{f}(x|c)\hat{P}(c)$ be to $f(x|c)P(c)$, so that there should be no decrease in the probability of correct classification.

Definition 16 *Let $\varepsilon_2(c)$ be any bound such that for all $x\in\mathcal{X}$*

$$\left|f(x|c)P(c)-\hat{f}(x|c)\hat{P}(c)\right|\leqslant\varepsilon_2(c). \tag{40}$$

For example, let $\mathcal{X}$ be discrete and finite. If $\varepsilon_2(x,c)=f(x|c)P(c)-\hat{f}(x|c)\hat{P}(c)$, then we obviously take $\varepsilon_2(c)=\max_{x\in\mathcal{X}}\mid\varepsilon_2(x,c)\mid$. Let us suppose that the approximation is a lower bound, that is, $f(x|c)P(c)>\hat{f}(x|c)\hat{P}(c)$ for all $x\in\mathcal{X}$, which is found by variation and normalization (Jordan et al., 1999; Wainwright and Jordan, 2003). These techniques give even expressions for $\varepsilon_2(x,c)$, for example for the exponential family of densities. We continue, however, with a lemma that involves this kind of differences in general.

Lemma 17 *Assume that* $\varepsilon_2(c) > 0$ *and* $\varepsilon_2(\widetilde{c}) > 0$ *exist as defined in (40). If* $P(c|x) > P(\widetilde{c}|x)$ *and*

$$|f(x|c)P(c) - f(x|\widetilde{c})P(\widetilde{c})| \geqslant \varepsilon_2(c) + \varepsilon_2(\widetilde{c})$$

then $\hat{P}(c|x) \geqslant \hat{P}(\widetilde{c}|x)$.

Proof We prove this by contradiction. First we assume that $\hat{P}(c|x) < \hat{P}(\widetilde{c}|x)$. With the plug-in classifier and (2) we get

$$\hat{f}(x|c)\hat{P}(c) < \hat{f}(x|\widetilde{c})\hat{P}(\widetilde{c}).$$

Now we continue by applying (40), that is, increasing margin in this inequality, which gives

$$\Rightarrow f(x|c)P(c) - \varepsilon_2(c) < f(x|\widetilde{c})P(\widetilde{c}) + \varepsilon_2(\widetilde{c})$$

$$\Leftrightarrow f(x|c)P(c) - f(x|\widetilde{c})P(\widetilde{c}) < \varepsilon_2(c) + \varepsilon_2(\widetilde{c}).$$

By the assumption $P(c|x) > P(\widetilde{c}|x)$ and by (1) we get that the quantity in the left hand side of the last inequality is positive, and hence the desired contradiction follows.

■

Lemma 17 used to state sufficient conditions such that $f(x|c)$ can be approximated without affecting the probability of correct classification.

Theorem 18 *If for all* $c, \widetilde{c} \in \mathcal{C}$

$$|f(x|c)P(c) - f(x|\widetilde{c})P(\widetilde{c})| \geqslant \varepsilon_2(c) + \varepsilon_2(\widetilde{c}),$$

then $P(\hat{c}_B(X) = C) = P(\hat{c}_{\hat{B}}(X) = C)$.

Proof From (13) we have that

$$P(\hat{c}_B(X) = C) - P(\hat{c}_{\hat{B}}(X) = C) = \int_{\{x|\hat{c}_{\hat{B}}(x) \neq c\}} (P(c)f(x|c) - P(\hat{c})f(x|\hat{c}))\, d\mu(x).$$

Now the result follows since Lemma 17 implies (through (1)) that $P(c|x) = P(\hat{c}|x)$. ■

We can also combine Theorem 18 with inequality (36). This gives us next corollary.

Corollary 19 *When* $\varepsilon_2(c) = \max\left(f(m|c) - f(m|c)^d, 1 - f(m|c)\right)$ *and*
$|P(c|x)P(c) - P(\widetilde{c}|x)P(\widetilde{c})| \geqslant \varepsilon_2(c)P(c) + \varepsilon_2(\widetilde{c})P(\widetilde{c})$*, then*

$$P(\hat{c}_{\hat{B}}(X) = C) = P(\hat{c}_B(x) = C).$$

In the context of Naïve Bayes our Corollary 19 can be seen as a generalization of the results in Domingos and Pazzani (1997), the finding of which is that Naïve Bayes is optimal for learning conjunctions and disjunctions of literals, as well as an extension of the more general result in Rish et al. (2001), which says that Naïve Bayes is optimal if Bayes classifier assigns only one feature to class 1 in a two-class problem. For example Corollary 19 is more general in the sense that a classifier that assigns more than one feature to a class can be optimal if the margin is wide enough.

9. Summary

We have presented exact and easily computable bounds for the degradation of probability of correct classification when Bayes classifiers are used with respect to partial plug-in conditional densities in a Bayesian network model (Theorem 7, Corollary 8 and Theorem 11).

An example of a Bayesian network plug-in classifier is the Naïve Bayes classifier (Definition 12). In the case where a Naïve Bayesian classifier is used, we have sharpened a bound of evaluating its effect (Theorem 13).

We have presented a bound on the class conditional approximation as well as the class probabilities such that the probability of making a correct decision is not degraded when basing the decision on $\hat{c}_{\hat{B}}(x)$ instead of $\hat{c}_B(x)$ using the bound in Theorem 13, thus generalizing the theory for explaining when Naïve Bayes is optimal.

Acknowledgments

A preliminary version of the results was reported at The 23rd Annual Workshop of the Swedish Artificial Intelligence Society (Ekdahl and Koski, 2006). The research was partially supported by Swedish Research Council, grant 4042/401 and EU 6th Programme grant QLK3-CT-2002-02097 (BACDRIVERS).

Appendix A. Simulation

Algorithm 1 Simulate($granularity, d$)

```
$atom = \frac{1}{granularity}$
$state$ is a placement of atoms on $\mathcal{X}$ such that for a state $x$, $f(x) = \frac{nr\ atoms\ there}{granularity}$
while not all unique placements of atoms have been searched do
    $p = \max_{x \in \mathcal{X}} f(x)$
    $diff = \max_{x \in \mathcal{X}} \left| f(x) - \prod_{i=1}^{d} f(x_i) \right|$
    if $diff > maxdiffs[p]$ then
        $maxdiffs[p] = diff$
    end if
    $state$ = next unique placement of atoms
end while
return $maxdiffs$
```

References

S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B*, 28(1):131–142, 1966.

S. Anoulova, P. Fischer, S. Polt, and H.U. Simon. Probably almost Bayes decisions. *Information and Computation*, 129(1):63–71, 1996.

S. Arimoto. Information-theoretical considerations on estimation problems. *Information and Control*, 19:181–194, 1971.

R.R. Bahadur. On classification based on responses to *n* dichotomous items. In H. Solomon, editor, *Studies in Item Analysis and Prediction*, pages 169–177, Stanford, California, 1961a. Stanford University Press.

R.R. Bahadur. A representation of the joint distribution of responses to *n* dichotomous items. In H. Solomon, editor, *Studies in Item Analysis and Prediction*, pages 158–168, Stanford, California, 1961b. Stanford University Press.

L.C. Barbosa. Maximum likelihood sequence estimators: A geometric view. *IEEE Transactions on Information Theory*, 35(2), 1989.

B.K. Bhattacharyya and G.T. Toussaint. An upper bound on the probability of misclassification in terms of Matusita's measure of affinity. *Annals of the Institute of Statistical Mathematics*, 34: 161–165, 1982.

C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Uncertainty in Artificial Intelligence (UAI-96)*, pages 115–123, 1996.

D.T. Brown. A note on approximations to discrete probability distributions. *Information and Control*, 2:386–392, 1959.

H.D. Brunk and D.A. Pierce. Estimation of discrete multivariate densities for computer-aided differential diagnosis. *Biometrika*, 61(3):493–499, 1974.

D.M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependency trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

G. Cooper. The computational complexity of probabilistic inference using bayesian Belief networks. *Artificial Intelligence*, 42(2-3):393–405, 1990.

T.M. Cover and J.A Thomas. *Elements of Information Theory*, chapter 2.2, page 169. Wiley, 1991.

R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, 1999.

H. Cramér. *Mathematical Methods of Statistics, 11 th printing*. Princeton University Press, 1966.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

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

M. Ekdahl. *Approximations of Bayes Classifiers for Statistical Learning of Clusters*. Licentiate thesis, Linköpings Universitet, 2006.

M. Ekdahl and T. Koski. On the performance of model based approximations in classification. In *Proceedings of The 23rd Annual Workshop of the Swedish Artificial Intelligence Society*, pages 73–82. `http://sais2006.cs.umu.se/`, 2006.

J.H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29 (2):1–36, 1997.

N. Glick. Sample-based classification procedures derived from density estimators. *Journal of the American Statistical Association*, 67(337):116–122, 1972.

N. Glick. Sample based multinomial classification. *Biometrics*, 29(2):241–256, 1973.

L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–763, 1954.

M. Gyllenberg and T. Koski. Probabilistic models for bacterial taxonomy. *International Statistical Review*, 69:249–276, 2001.

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. The MIT Press, 2001.

D.J. Hand and K. Yu. Idiot's Bayes–not so stupid after all? *International Statistical Review*, 69(3): 385–398, 2001.

W. Hoeffding. Stochastische unabhängigkeit und funktionaler zusammenhang. *Skandinavisk Aktuarietidskrift*, 25:200–227, 1942.

W. Hoeffding and J. Wolfowitz. Distinguishability of sets of distributions. *The Annals of Mathematical Statistics*, 29(3):700–718, 1958.

K. Huang, I. King, and M.R. Lyu. Finite mixture model of bounded semi-naive Bayesian network classifier. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN-2003)*, volume 2714 of *Lecture Notes in Computer Science*. Springer, 2003.

M.I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications Technology*, 15(1):52–60, 1967.

G.K. Kaleh. Channel equalization for block transmission systems. *IEEE Journal on Selected Areas in Communications*, 13(1):150–121, 1995.

K.B. Korb and A.E. Nicholson. *Bayesian Artificial Intelligence*. Chapman and Hall, 2004.

H.H. Ku and S. Kullback. Approximating discrete probability distributions. *IEEE Transactions on Information Theory*, 15:444–447, 1969.

S. Kullback. *Information Theory and Statistics*, chapter 9, page 190. Dover Publications Inc., 1997.

S.L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87:1098–1108, 1990.

P.M. Lewis. Approximating probability distributions to reduce storage requirements. *Information and Control*, 2:214–225, 1959.

D.K. McGraw and J.F. Wagner. Elliptically symmetric distributions. *IEEE Transactions on Information Theory*, 14(1):110 – 120, 1968.

D.H.II. Moore. Evaluation of five discrimination procedures for binary variables. *Journal of the American Statistical Association*, 68(342):399–404, 1973.

X. Nguyen, M.J. Wainwright, and M.I. Jordan. On divergences, surrogate loss functions, and decentralized detection. Technical Report 695, University of California, Berkeley, 2005.

J. Ott and R.A. Kronmal. Some classification procedures for multivariate binary data using orthogonal functions. *Journal of the American Statistical Association*, 71(354):391–399, 1976.

G.A. Pistone, E.A. Riccomagno, and H.P.A. Wynn. Gröbner bases and factorisation in discrete probability and Bayes. *Statistics and Computing*, 11(1):37–46, 2001.

E.J.G. Pitman. *Some Basic Theory for Statistical Inference*, chapter 2. Chapman and Hall, 1979.

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

I. Rish, J. Hellerstein, and J. Thathachar. An analysis of data characteristics that affect Naive Bayes performance. Technical Report RC21993, IBM, 2001.

J.Van Ryzin. Bayes risk consistency of classification procedures using density estimation. *Sankhya Series A*, 28:261–270, 1966.

P. Scheinck. Symptom diagnosis Bayes's and Bahadur's distribution. *International Journal of Biomedical Computing*, 3:17–28, 1972.

M.J. Schervish. *Theory of Statistics*. Springer, second edition, 1995.

D. Slepian. On the symmetrized Kronecker power of a matrix and extensions of Mehler's formula for Hermite polynomials. *SIAM Journal on Mathematical Analysis*, 3:606–616, 1972.

H. Strasser. *Mathematical Theory of Statistics*, chapter 2.2. Walter de Gruyter, 1985.

J.L. Teugels. Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, 33(1):256–268, 1990.

D.M. Titterington, G.D. Murray, L.S. Murray, D.J. Spiegelhalter, A.M. Skene, J.D.F. Habbema, and G.J. Gelpke. Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society.*, 144(2):145–175, 1981.

F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

G.T. Toussaint. Polynomial representation of classifiers with independent discrete-valued features. *IEEE Transactions on Computers*, 21:205–208, 1972.

R. van Engelen. Approximating bayesian Belief networks by arc removal. *IEEE Transactions on Pattern Analysis and Machine Intellignce*, 19(8):916–920, 1997.

R. Vilata and I. Rish. A decomposition of classes via clustering to explain and improve Naive Bayes. In *Machine Learning: ECML 2003: 14th European Conference on Machine Learning*, pages 444 – 455, 2003.

T.R. Vilmansen. Feature evaluation with measures of probabilistic dependence. *IEEE Transactions on Computers*, 22(4):381–387, 1973.

T.R. Vilmansen. On dependence and discrimination in pattern recognition. *IEEE Transactions on Computers*, 21:1029–1031, 1971.

M.J Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Dept. of Statistics, 2003.

J. Williamson. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford University Press, 2005.