On the Asymptotic Optimality of Maximum Margin Bayesian Networks

Sebastian Tschiatschek
Graz University of Technology, Austria

Franz Pernkopf
Graz University of Technology, Austria

Abstract

Maximum margin Bayesian networks (MMBNs) are Bayesian networks with discriminatively optimized parameters. They have shown good classification performance in various applications. However, there has not been any theoretic analysis of their asymptotic performance, e.g. their Bayes consistency. For specific classes of MMBNs, i.e. MMBNs with fully connected graphs and discrete-valued nodes, we show Bayes consistency for binary-class problems and a sufficient condition for Bayes consistency in the multi-class case. We provide simple examples showing that MMBNs in their current formulation are not Bayes consistent in general. These examples are especially interesting, as the model used for the MMBNs can represent the assumed true distributions. This indicates that the current formulations of MMBNs may be deficient. Furthermore, experimental results on the generalization performance are presented.

1 INTRODUCTION

Maximum margin Bayesian networks (MMBNs) were first introduced in (Guo et al., 2005). The basic idea is to mimic the concept of the margin known from support vector machines (SVMs) in a probabilistic environment. SVMs are one of the best performing classifiers available. In their basic formulation, they separate samples from different classes by a linear hyperplane. While SVMs are theoretically well-understood (Vapnik, 1998; Platt, 1999; Shalev-Shwartz et al., 2007), there exist several issues that are hard to deal with. One example is the treatment of missing features in the data. SVMs usually require imputation techniques to complete the data before further processing (Little and Rubin, 2002). In contrast, Bayesian network (BN) classifiers can naturally handle missing features.

BN classifiers are composed of a directed acyclic graph and conditional probabilities associated with the nodes of this graph. The task of identifying these conditional probabilities is termed *parameter learning*. Generative parameter learning aims at finding a joint probability distribution that explains the generation of the samples, e.g. by identifying maximum likelihood (ML) parameters. ML parameters minimize the Kullback-Leibler divergence between the true joint distribution and the joint distributions that can be represented by the considered BNs (Koller and Friedman, 2009). However, in classification tasks the objective is classification rate. This is the focus of discriminative parameter learning which aims at maximizing the classification performance. Representatives of this paradigm are maximum conditional likelihood (MCL) and maximum margin (MM) parameter learning.

Previous results (Ng and Jordan, 2001) show that logistic regression, i.e. BN classifiers with naive Bayes structure and MCL parameters, exhibit lower asymptotic generalization error than classifiers with ML parameters. For BNs with MM parameters, i.e. MMBNs, no such results are available. However, comparable performance of these classifiers to SVMs has been reported (Pernkopf et al., 2012). This encourages the investigation of BN classifiers with MM parameters in more detail. Specifically, in this paper we address the issue of Bayes consistency, i.e. whether classifiers with parameters optimizing the MM objective yield asymptotically almost surely the Bayes optimal classifier. Our main results are:

1. MMBN classifiers with discrete-valued nodes are in general not Bayes consistent.
2. MMBN classifiers with discrete-valued nodes and fully connected graphs are Bayes consistent in

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

binary-class classification tasks.

3. A sufficient condition for MMBN classifiers with discrete-valued nodes and fully connected graphs to be Bayes consistent in multi-class classification tasks.

The remainder of this paper is structured as follows: In Section 2, we introduce the framework of probabilistic classifiers. In Section 3, we present definitions of MMBNs from the literature followed by our theoretical results in Section 4. In Section 5 we illustrate the theoretical insights by empirical results and subsequently discuss some implications of our results in Section 6. Finally, we conclude the paper in Section 7.

2 BACKGROUND

We consider classification problems in a probabilistic setting where $C, X_1, \ldots, X_L$ are RVs. These RVs are jointly distributed according to the distribution $\mathrm{P}^*(C, X_1, \ldots, X_L)$. We refer to this distribution as *true distribution*. The RV C corresponds to the class label and the RVs $X_1, \ldots, X_L$ are the features.

We stack the RVs $X_1, \ldots, X_L$ into the random vector $\mathbf{X} = (X_1, \ldots, X_L)$. Each X_i can take one value in the set $\mathrm{sp}(X_i)$. Similarly, C can assume values in $\mathrm{sp}(C)$. We use $\mathrm{sp}(\mathbf{X})$ to refer to the set of possible assignments of $\mathbf{X}$. Instantiations of RVs are denoted using lower case letters, i.e. $\mathbf{x}$ is an instantiation of $\mathbf{X}$ and c an instantiation of C, respectively. Whenever $\mathrm{P}(C, \mathbf{X})$ is a probability distribution over C and $\mathbf{X}$, we write $\mathrm{P}(c, \mathbf{x})$ as an abbreviation for $\mathrm{P}(C = c, \mathbf{X} = \mathbf{x})$. The expectation of a function $f(C, \mathbf{X})$ with respect to a joint distribution $\mathrm{P}(C, \mathbf{X})$ is denoted as $\mathbb{E}_{\mathrm{P}(C,\mathbf{X})}[f(C, \mathbf{X})]$.

A classifier h is a mapping $h\colon \mathrm{sp}(\mathbf{X}) \to \mathrm{sp}(C), \mathbf{x} \mapsto h(\mathbf{x})$, i.e. a classifier maps an instantiation $\mathbf{x}$ of the attributes to class c. The merit of a classifier can be quantified by its classification rate, or equivalently, by its generalization error.

Definition 1 (Generalization Error, Classification Rate)**.** *Let* $h\colon \mathrm{sp}(\mathbf{X}) \to \mathrm{sp}(C)$ *be a classifier. Its* generalization error $\mathrm{Err}(h)$ *is*

$$\mathrm{Err}(h) = \mathbb{E}_{\mathrm{P}^*(C,\mathbf{X})}\left[\mathbf{1}\{h(\mathbf{X}) \neq C\}\right], \tag{1}$$

where $\mathbf{1}\{a\}$ *is the indicator function that equals* 1 *if and only if the statement* a *is true and* 0 *otherwise. The* classification rate $\mathrm{CR}(h)$ *is* $\mathrm{CR}(h) = 1 - \mathrm{Err}(h)$.

Any probability distribution $\mathrm{P}(C, \mathbf{X})$ naturally induces a classifier $h_{\mathrm{P}(C,\mathbf{X})}$ according to

$$\begin{aligned} h_{\mathrm{P}(C,\mathbf{X})}\colon \quad & \mathrm{sp}(\mathbf{X}) \to \mathrm{sp}(C), \\ & \mathbf{x} \mapsto \arg\max_{c' \in \mathrm{sp}(C)} \mathrm{P}(c'|\mathbf{x}), \end{aligned} \tag{2}$$

i.e. an instantiation $\mathbf{x}$ of the features is classified by the maximum a posteriori estimate of class c given $\mathbf{x}$. Note that $\arg\max_{c'} \mathrm{P}(c'|\mathbf{x})$ is not necessarily unique, i.e. different classes may achieve $\max_{c'} \mathrm{P}(c'|\mathbf{x})$. These classes are collected in the set $[C|\mathbf{x}]_{\mathrm{P}(C,\mathbf{X})}$, i.e.

$$[C|\mathbf{x}]_{\mathrm{P}(C,\mathbf{X})} = \left\{ c \mid \mathrm{P}(C = c|\mathbf{X} = \mathbf{x}) = \max_{c' \in \mathrm{sp}(C)} \mathrm{P}(C = c'|\mathbf{X} = \mathbf{x}) \right\}. \tag{3}$$

Whenever $[C|\mathbf{x}]_{\mathrm{P}(C,\mathbf{X})}$ consists of more than a single element, we assume the classifiers to return one of these classes uniformly at random[1] — this is termed as *optimally* classified with respect to $\mathrm{P}(C, \mathbf{X})$.

In this paper, we consider probability distributions represented by Bayesian networks (BN). A BN $\mathcal{B} = (\mathcal{G}, \mathbf{P})$ consist of a directed acyclic graph $\mathcal{G}$ together with a set of conditional probabilities $\mathbf{P}$ (Pearl, 1988). The nodes $\mathbf{V} = \{X_0, \ldots, X_L\}$ of $\mathcal{G}$ correspond to RVs and the edges encode conditional independencies among these RVs. The conditional probabilities $\mathbf{P} = \{\mathrm{P}(X_0|Pa(X_0)), \ldots, \mathrm{P}(X_L|Pa(X_L))\}$ are associated with the nodes of the graph and $Pa(X_i)$ denotes the parents of X_i in $\mathcal{G}$. The BN defines the joint distribution

$$\mathrm{P}^{\mathcal{B}}(X_0, \ldots, X_L) = \prod_{i=0}^{L} \mathrm{P}(X_i|Pa(X_i)). \tag{4}$$

Throughout this paper, we assume that X_0 corresponds to the class RV, i.e. $X_0 = C$. According to the joint distribution, a BN $\mathcal{B}$ induces the classifier $h_{\mathcal{B}} = h_{\mathrm{P}^{\mathcal{B}}(C,\mathbf{X})}$.

In the following, we are interested in classifiers in the hypothesis class of BN classifiers with discrete RVs and fixed graph structure $\mathcal{G}$, denoted as $\mathcal{B}(\mathcal{G})$. Optimality of a classifier with respect to its hypothesis class is defined as follows:

Definition 2 (Optimal Classifier)**.** *A classifier* $h_{\mathcal{B}}$, $\mathcal{B} \in \mathcal{B}(\mathcal{G})$ *is optimal with respect to the hypothesis class* $\mathcal{B}(\mathcal{G})$ *if it satisfies*

$$\mathrm{Err}(h_{\mathcal{B}}) = \inf_{\mathcal{B}' \in \mathcal{B}(\mathcal{G})} \mathrm{Err}(h_{\mathcal{B}'}). \tag{5}$$

A classifier from any hypotheses class can not be better than the *Bayes optimal classifier* $h_{\mathrm{P}^*(C,\mathbf{X})}$ (Mitchell, 1997). The sub-optimality of a classifier $h_{\mathcal{B}} \in \mathcal{B}(\mathcal{G})$

[1] Technically, $h_{\mathrm{P}(C,\mathbf{X})}$ is not a mapping, because there is no unique assignment of $\mathbf{x} \in \mathrm{sp}(\mathbf{X})$ to some $c \in \mathrm{sp}(C)$. For ease of notation, we ignore this fact.

can be expressed as

$$\begin{aligned}&\mathrm{Err}(h_{\mathcal{B}}) - \mathrm{Err}(h_{\mathrm{P}^*(C,\mathbf{X})}) \\ &= \left(\mathrm{Err}(h_{\mathcal{B}}) - \inf_{\mathcal{B}'\in\mathcal{B}(\mathcal{G})} \mathrm{Err}(h_{\mathcal{B}'})\right) \\ &+ \left(\inf_{\mathcal{B}'\in\mathcal{B}(\mathcal{G})} \mathrm{Err}(h_{\mathcal{B}'}) - \mathrm{Err}(h_{\mathrm{P}^*(C,\mathbf{X})})\right),\end{aligned} \tag{6}$$

where the first term is referred to as *estimation error* and the second term as *approximation error*. The estimation error measures the optimality of the classifier $h_{\mathcal{B}}$ with respect to the class $\mathcal{B}(\mathcal{G})$, while the approximation error quantifies how close the best classifier in $\mathcal{B}(\mathcal{G})$ is to the Bayes optimal classifier. When considering limited graph structures $\mathcal{G}$, i.e. $\mathcal{G}$ is not fully connected, the generalization error of the Bayes optimal classifier can not be achieved in general, but there will at least be a bias corresponding to the approximation error. Throughout this paper, we consider the hypothesis classes of BNs with naive Bayes (NB) structure and fully connected graphs. In NB structures, the class node has no parents, i.e. $Pa(C) = \emptyset$, and the only parent of any feature is the class node, i.e. $Pa(X_i) = \{C\}$. In fully connected graphs, the class node has no parents, i.e. $Pa(C) = \emptyset$, and for any feature $Pa(X_i) = \{C, X_1, \ldots, X_{i-1}\}$.

The distribution $\mathrm{P}^*(C, \mathbf{X})$ is hardly ever known and can, therefore, be usually only used as a reference in synthetic experiments. In most practical situations no information about the true distribution is directly available but a training set $\mathcal{T}$ consisting of N labeled samples drawn i.i.d. from $\mathrm{P}^*(C, \mathbf{X})$ is available, i.e.

$$\mathcal{T} = \{(c^{(1)}, \mathbf{x}^{(1)}), \ldots, (c^{(N)}, \mathbf{x}^{(N)})\}. \tag{7}$$

A sequence of classifiers $h_{\mathcal{B}}^{A,N}$, where the superscript A, N denotes that the classifier is obtained from a training set of size N using the parameter learning method A (e.g. maximum likelihood or maximum margin), is *Bayes consistent* (with respect to $\mathcal{B}(\mathcal{G})$), if

$$\mathrm{Err}(h_{\mathcal{B}}^{A,N}) \to \inf_{\mathcal{B}'\in\mathcal{B}(\mathcal{G})} \mathrm{Err}(h_{\mathcal{B}'}) \text{ a.s. as } N \to \infty. \tag{8}$$

A typical approach of identifying the parameters of such classifiers is by learning the parameters in a generative sense, i.e. by maximizing the likelihood of the samples in $\mathcal{T}$. However, models with simple structure often can not represent $\mathrm{P}^*(C, \mathbf{X})$. The result is poor classification performance. To compensate for this *model mismatch*, parameters can be optimized discriminatively. For example, parameters with maximum conditional likelihood (MCL) of the samples in the training set can be identified (Roos et al., 2005). A competitive alternative to using MCL parameters, is to employ MM parameters. This type of parameter optimization is introduced in the next section.

3 MAXIMUM MARGIN BAYESIAN NETWORKS

Guo et al. (2005) introduced MMBNs as a convex optimization problem for parameter learning. Later, the maximum margin criterion was reformulated and a conjugate gradient based method for parameter learning was provided (Pernkopf et al., 2012). In experiments, both formulations have shown similar classification performance while the conjugate gradient optimization is beneficial in terms of computation cost. We shortly review both formulations and provide an example for which neither formulation retrieves a Bayes consistent classifier, although the Bayes optimal classifier is within the considered hypothesis class $\mathcal{B}(\mathcal{G})$. In the remainder of the paper we adopt the MMBN objective of Pernkopf et al. (2012).

3.1 Formulation by Pernkopf et al.

Assuming a fixed graph $\mathcal{G}$, the objective for learning the joint probability $\mathrm{P}^{\mathcal{B}}(C, \mathbf{X})$ is based on the *margins*

$$\widetilde{d}^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)}) = \frac{\mathrm{P}^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)})}{\max_{c'\neq c^{(n)}} \mathrm{P}^{\mathcal{B}}(c', \mathbf{x}^{(n)})} \tag{9}$$

of the training samples. Therefore, the n^{th} sample in the training set is classified correctly if $\widetilde{d}^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)}) > 1$. To handle non separable data, a hinge function is used such that

$$d^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)}) = \min\left(\widetilde{\gamma}, \widetilde{d}^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)})\right), \tag{10}$$

where $\widetilde{\gamma} > 1$ is a parameter that controls the influence of the margins. The objective for learning MMBNs is maximization of the product of $d^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)})$ over the samples.

Definition 3 (Maximum Margin Bayesian Network)**.** *A BN $\mathcal{B} = (\mathcal{G}, \mathbf{P})$ that achieves the optimal value of*

$$\underset{\mathcal{B}'\in\mathcal{B}(\mathcal{G})}{\text{maximize}} \quad \prod_{n=1}^{N} \min\left(\widetilde{\gamma}, \widetilde{d}^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)})\right) \tag{11}$$

is an MMBN.

This definition can be equivalently stated in the log-domain by requiring $\mathcal{B}$ to solve

$$\begin{aligned}\underset{\mathcal{B}'\in\mathcal{B}(\mathcal{G})}{\text{maximize}} \quad &\frac{1}{N}\sum_{n=1}^{N} \min\Big(\gamma, \log \mathrm{P}^{\mathcal{B}'}(c^{(n)}, \mathbf{x}^{(n)}) \\ &- \max_{c'\neq c^{(n)}} \log \mathrm{P}^{\mathcal{B}'}(c', \mathbf{x}^{(n)})\Big),\end{aligned} \tag{12}$$

where $\gamma = \log \widetilde{\gamma}$ and the objective is normalized by the number of training samples N. This allows the introduction of the empirical distribution on the training

set $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$, i.e.

$$\mathrm{P}^{\mathcal{T}}(c, \mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\{c^{(n)} = c, \mathbf{x}^{(n)} = \mathbf{x}\} \tag{13}$$

to the optimization problem. The objective (12) becomes

$$\begin{aligned} \underset{\mathcal{B}' \in \mathcal{B}(\mathcal{G})}{\text{maximize}} \quad & \sum_{c,\mathbf{x}} \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}) \min \Big(\gamma, \log \mathrm{P}^{\mathcal{B}'}(c, \mathbf{x}) \\ & - \max_{c' \neq c} \log \mathrm{P}^{\mathcal{B}'}(c', \mathbf{x}) \Big). \end{aligned} \tag{14}$$

A justification why BNs with MM parameters can be advantageous over BNs with ML or MCL parameters is given in Appendix A.

3.2 Formulation by Guo et al.

The formulation by Guo et al. is based on the representation of the probabilities $\mathrm{P}^{\mathcal{B}}(c, \mathbf{x})$ as

$$\mathrm{P}^{\mathcal{B}}(c, \mathbf{x}) = \exp(\boldsymbol{\phi}(c, \mathbf{x})^T \mathbf{w}), \tag{15}$$

where the entries of $\mathbf{w}$ correspond to the log-probabilities of the BN $\mathcal{B}$, i.e. the $(i, j, \mathbf{h})^{\text{th}}$ entry of $\mathbf{w}$ corresponds to $\mathbf{w}^i_{j|\mathbf{h}} = \log \mathrm{P}(X_i = j | Pa(X_i) = \mathbf{h})$. The vector $\boldsymbol{\phi}(c, \mathbf{x})$ is a binary vector indicating which entries of the log conditional probabilities $\log \mathrm{P}(X_i | Pa(X_i))$ are to be summed up for assignment $C = c$ and $\mathbf{X} = \mathbf{x}$. This enables to represent the logarithm of the margin (9) as

$$\log \widetilde{d}^{\mathcal{B}}(c^{(n)}, \mathbf{x}^{(n)}) = \min_{c' \neq c^{(n)}} [\boldsymbol{\phi}(c^{(n)}, \mathbf{x}^{(n)}) - \boldsymbol{\phi}(c', \mathbf{x}^{(n)})]^T \mathbf{w}. \tag{16}$$

Learning the parameters of MMBNs is then performed by solving

$$\begin{aligned} \underset{\gamma, \mathbf{w}}{\text{minimize}} \quad & \frac{1}{2\gamma^2} + \\ & B\, N \sum_{c,\mathbf{x}} \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}) \max \Big(0, \gamma - \min_{c' \neq c} [\boldsymbol{\phi}(c, \mathbf{x}) - \boldsymbol{\phi}(c', \mathbf{x})]^T \mathbf{w} \Big) \end{aligned} \tag{17}$$

$$\begin{aligned} \text{s.t.} \quad & \sum_j \exp(w^i_{j|\mathbf{h}}) = 1, \forall i, \mathbf{h} \\ & \gamma > 0, \end{aligned} \tag{18}$$

where $B \geq 0$ is a trade-off parameter between a large margin and correct classification. To end up with a convex formulation, Guo et al. replace the constraints $\sum_j \exp(w^i_{j|\mathbf{h}}) = 1$ by inequalities, i.e. $\sum_j \exp(w^i_{j|\mathbf{h}}) \leq 1$.

Due to the relaxation, the found parameters are typically not normalized. However, as pointed out in (Roos et al., 2005; Wettig et al., 2003), for certain network structures renormalization is possible without changing the classifier induced by the unnormalized parameters. The condition is for example satisfied by NB structures and fully connected graphs. The condition for renormalization is as follows:

Condition 1 (Renormalization (Wettig et al., 2003))**.** *For all feature RVs X_j with $C \in Pa(X_j)$ there exists another RV $X_i \in Pa(X_j)$ such that $Pa(X_j) \subseteq Pa(X_i) \cup \{X_i\}$.*

3.3 Inconsistent MMBNs

In this section we present an example for which both definitions of MMBNs result almost surely in inconsistent classifiers. Consider a classifier with no features, i.e. $\mathbf{X} = \emptyset$, in a three-class scenario. Let the true distribution be defined by

$$\begin{aligned} &\mathrm{P}^*(C = 1) = 0.4, \\ &\mathrm{P}^*(C = 2) = 0.3, \text{and} \\ &\mathrm{P}^*(C = 3) = 0.3. \end{aligned}$$

Hence, the Bayes optimal classifier would classify all instances as belonging to class 1. By the strong law of large numbers, the empirical distribution will satisfy asymptotically almost surely $\mathrm{P}^{\mathcal{T}}(C = 1) > \mathrm{P}^{\mathcal{T}}(C = 2)$, $\mathrm{P}^{\mathcal{T}}(C = 1) > \mathrm{P}^{\mathcal{T}}(C = 3)$ and $\mathrm{P}^{\mathcal{T}}(C = 1) < \mathrm{P}^{\mathcal{T}}(C = 2) + \mathrm{P}^{\mathcal{T}}(C = 3)$. In this case, any distribution inducing a Bayes optimal classifier has strictly smaller (larger) objective than the uniform distribution according to problem (14) (problem (17))[2]. Consequently, any MM distribution induces almost surely a Bayes inconsistent classifier.

In this example, the Bayes optimal classifier can be represented by the assumed model. Nevertheless, an inconsistent classifier is obtained. We can deduce, that in the multi-class case we must not hope for Bayes consistency of MMBN classifiers in general.

4 THEORETICAL RESULTS

4.1 Bayes Consistency of Fully Connected MMBNs

In this section, we show that fully-connected binary-class MMBN classifiers with discrete-valued nodes are Bayes consistent. Furthermore, we present a sufficient condition for Bayes consistency of multi-class MMBN classifiers. The proof consists of two parts. In the first part, we prove optimality with respect to the empirical distribution of the training set. In the second part,

[2]The necessary calculations for this result are straightforward, but provided in the supplementary material for completeness.

we conclude that MMBN classifiers are almost surely Bayes consistent.

Lemma 1 (Optimality in the binary-class case)**.** *Let C be a binary class variable, and let $\mathcal{T}$ be a training set with empirical distribution $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$ and $\mathcal{G}$ a fully connected graph. Any MMBN classifier on $\mathcal{G}$ is optimal with respect to $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$.*

The proof is provided in Appendix B.

Lemma 2 (Optimality in the multi-class case)**.** *Let C be a class variable with $|\mathrm{sp}(C)| > 2$, and let $\mathcal{T}$ be a training set with empirical distribution $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$ and $\mathcal{G}$ a fully connected graph. Any MMBN classifier on $\mathcal{G}$ is optimal with respect to $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$ if*

$$\forall \mathbf{x}\ \exists c:\ \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}) > \sum_{c' \neq c} \mathrm{P}^{\mathcal{T}}(c', \mathbf{x}). \tag{19}$$

The proof is similar to that of Lemma 1 (for reference it is provided in the supplementary material). Bluntly speaking, condition (19) requires that for every instantiation of the features $\mathbf{x}$ there is a dominant class.

Using Lemma 1 and 2 we can derive the following theorem.

Theorem 1. *Any MMBN classifier with a fully connected graph is Bayes consistent if*

(a) $|\mathrm{sp}(C)| = 2$*, i.e. the class variable is binary, or*

(b) $|\mathrm{sp}(C)| > 2$*, i.e. the multi-class case, and additionally the true distribution $\mathrm{P}^*(C, \mathbf{X})$ satisfies*

$$\forall \mathbf{x}\ \exists c:\ \mathrm{P}^{*}(c, \mathbf{x}) > \sum_{c' \neq c} \mathrm{P}^{*}(c', \mathbf{x}). \tag{20}$$

Proof. We have already established that, given the stated conditions, MMBN classifiers are optimal with respect to the empirical distribution on the training set. With growing sample size *the empirical distribution converges to the true distribution.* Therefore, the MMBN classifier converges asymptotically almost surely to the Bayes optimal classifier. □

4.2 MMBN Classifiers with not Fully Connected Graphs are not Necessarily Bayes Consistent

MMBN classifiers with not fully-connected graphs $\mathcal{G}$ are not Bayes consistent in general. This is even true in cases in which the true distribution can be represented by some BN $\mathcal{B} \in \mathcal{B}(\mathcal{G})$.

As an example consider a naive Bayes classifier with two features. Assume that the true data distribution $\mathrm{P}^*(C, \mathbf{X})$ satisfies the independence assumptions of the naive Bayes network and that the conditional probability densities are given according to Table 1(a). For $\gamma = 1$, there exist MMBN classifiers that are Bayes consistent and MMBN classifiers that are inconsistent, i.e. there is no unique optimal (and consistent) solution. The corresponding MM distributions are shown in Table 1(b), and Table 1(c), respectively. The inconsistent distribution induces a classifier which has uniform class posterior for the samples $(x_1 = 1, x_2 = 1)$ and $(x_1 = 1, x_2 = 2)$. This results in a classification rate that is 4.5 percent smaller than the maximum classification rate, i.e.

$$\begin{aligned}
&\mathrm{CR}(h_{\mathrm{P}^*(C,\mathbf{X})}) - \mathrm{CR}(h_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}) \qquad (21)\\
&= \mathrm{P}^*(c_1, \mathbf{x}_1) - \frac{1}{2}\left(\mathrm{P}^{\mathrm{MM}}(c_1, \mathbf{x}_1) + \mathrm{P}^{\mathrm{MM}}(c_2, \mathbf{x}_1)\right)\\
&\quad + \mathrm{P}^*(c_2, \mathbf{x}_2) - \frac{1}{2}\left(\mathrm{P}^{\mathrm{MM}}(c_1, \mathbf{x}_2) + \mathrm{P}^{\mathrm{MM}}(c_2, \mathbf{x}_2)\right)\\
&= 0.14 - \frac{1}{2}(0.14 + 0.12) + 0.28 - \frac{1}{2}(0.28 + 0.21)\\
&= 0.045,
\end{aligned}$$

where c_1 is a shorthand for $C = 1$, c_2 for $C = 2$, $\mathbf{x}_1$ for $X_1 = 1, X_2 = 1$, and $\mathbf{x}_2$ for $X_1 = 1, X_2 = 2$, respectively.

5 EXPERIMENTAL RESULTS

We performed two experiments supporting the theoretical results in this paper. Furthermore, an experiment demonstrating that MMBNs can perform well in the case of model mismatch is presented.

5.1 Bayes Consistent Classification Using Fully Connected Graphs

We assumed an arbitrary random distribution $\mathrm{P}^*(C, \mathbf{X})$ for a fully connected graph. These distributions were obtained by selecting each entry of the conditional probabilities associated with the nodes of the graph uniformly at random in the range $[0, 1]$. To end up with properly normalized distributions, each conditional probability distribution was normalized subsequently. From the obtained distribution we generated training sets with an increasing number of samples. On these training sets we determined BN classifiers with fully connected graphs and using ML, MCL, and MM parameters. MM parameters are determined using the linear program provided in the supplementary material and employing 5-fold cross-validation to select the value of γ. We evaluated the generalization performance of these classifiers with respect to the true distribution.

As pointed out above, classifiers with both ML and MM parameters have to converge to the optimal clas-

Table 1: Probability distribution for which MMBN classifiers can be inconsistent.

(a) True distribution $\mathrm{P}^*(C, \mathbf{X})$; Objective (14) evaluates to 0.049.

	$c=1$	$c=2$
$\mathrm{P}(C=c)$	0.5	0.5
$\mathrm{P}(X_1=1\|C=c)$	0.7	0.8
$\mathrm{P}(X_1=2\|C=c)$	0.3	0.2
$\mathrm{P}(X_2=1\|C=c)$	0.4	0.3
$\mathrm{P}(X_2=2\|C=c)$	0.6	0.7

C	X_1	X_2	$\mathrm{P}^*(C, \mathbf{X})$
1	1	1	0.14
2	1	1	0.12
1	2	1	0.06
2	2	1	0.03
1	1	2	0.21
2	1	2	0.28
1	2	2	0.09
2	2	2	0.07

(b) Inconsistent MM distribution $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$; Objective (14) evaluates to 0.05.

	$c=1$	$c=2$
$\mathrm{P}(C=c)$	0.5938	0.4062
$\mathrm{P}(X_1=1\|C=c)$	0.5	0.7311
$\mathrm{P}(X_1=2\|C=c)$	0.5	0.2689
$\mathrm{P}(X_2=1\|C=c)$	0.5	0.5
$\mathrm{P}(X_2=2\|C=c)$	0.5	0.5

C	X_1	X_2	$\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$
1	1	1	0.1485
2	1	1	0.1485
1	2	1	0.1485
2	2	1	0.0546
1	1	2	0.1485
2	1	2	0.1485
1	2	2	0.1485
2	2	2	0.0546

(c) Consistent MM distribution $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$; Objective (14) evaluates to 0.05.

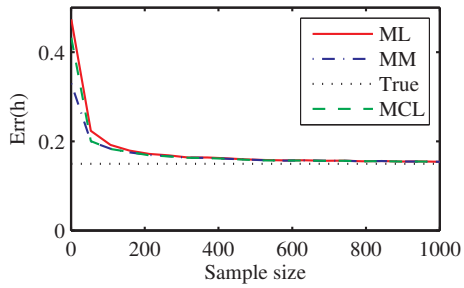
	$c=1$	$c=2$
$\mathrm{P}(C=c)$	0.5798	0.4202
$\mathrm{P}(X_1=1\|C=c)$	0.4750	0.5250
$\mathrm{P}(X_1=2\|C=c)$	0.5250	0.4750
$\mathrm{P}(X_2=1\|C=c)$	0.5	0.3100
$\mathrm{P}(X_2=2\|C=c)$	0.5	0.6900

C	X_1	X_2	$\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$
1	1	1	0.1377
2	1	1	0.0684
1	2	1	0.1522
2	2	1	0.0619
1	1	2	0.1377
2	1	2	0.1522
1	2	2	0.1522
2	2	2	0.1377

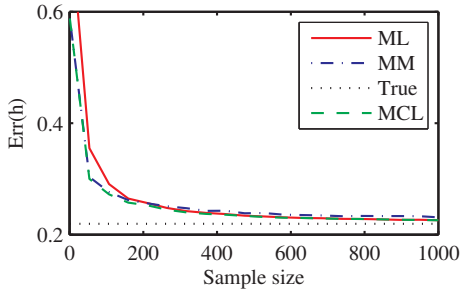
sifier as the training set size increases. Results are averaged over 100 different training sets for every sample size using 100 different random parameter sets for the true distribution (the selected true distributions satisfy the condition of Theorem 1). Results for using 5 binary features and 2 classes, as well as 5 binary features and 4 classes are shown in Figures 1(a) and 1(b), respectively. Convergence to the optimal classifier can be observed.

5.2 Convergence Experiments Assuming NB Structure

We repeated the experiment from above using true distributions satisfying the factorization properties of NB networks. BN classifiers with NB structure and ML, MCL and MM parameters are determined. In hope of obtaining unique MM parameters, we selected MM parameters with minimum ℓ_1-norm. Results for networks with 5 binary features and 2 classes, as well as 5 binary features and 4 classes are shown in Figures 2(a) and 2(b), respectively. As noticed in Section 4, the MMBN classifiers do not converge to the optimal classifier. In contrast, ML and MCL classifiers achieve the lowest possible generalization error.



(a) 2 classes



(b) 4 classes

Figure 1: Convergence of ML, MCL and MMBN classifiers to the optimal classifier assuming a fully connected graph. The generalization error of the optimal classifier is indicated by the dotted line (= True).

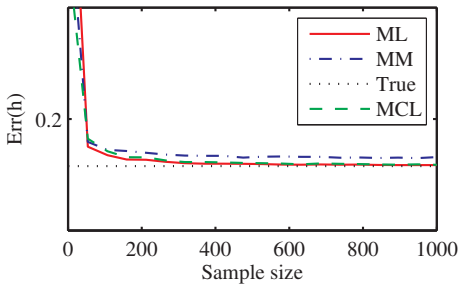
5.3 Model Mismatch

The setup for this experiments is similar to Section 5.1. For the true distribution an arbitrary distribution over $(C, \mathbf{X})$ is assumed, but the BN classifiers are determined using NB structures. The results for two-class and four-class classification are shown in Figures 3(a) and 3(b), respectively.

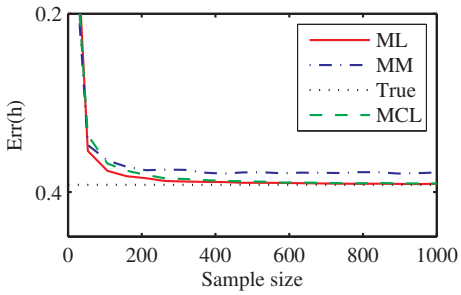
We observe that classifiers with MM parameters converge to a lower asymptotic error than classifiers with ML parameters. This is consistent with the observations in (Ng and Jordan, 2001) where generatively optimized NB classifiers are compared to logistic regression. In cases of model mismatch discriminative learning is usually beneficial.

6 DISCUSSION

We presented multi-class examples for which the Bayes optimal classifier can be represented by the considered models but is not retrieved by learning BN classifiers with MM parameters, cf. Sections 3.3 and 4.2. This suggests that the formulations of MMBNs is de-

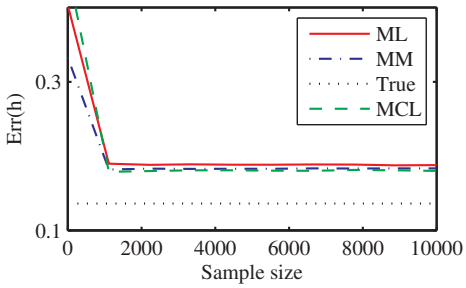


(a) 2 classes

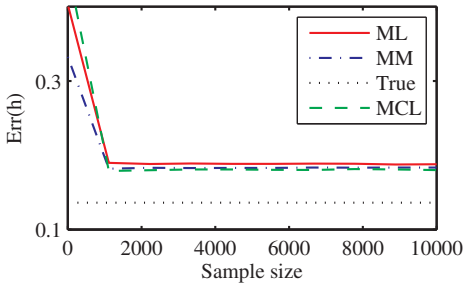


(b) 4 classes

Figure 2: Convergence of ML, MCL and MMBN classifiers assuming a NB structure. The generalization error of the optimal classifier is indicated by the dotted line (= True).



(a) 2 classes



(b) 4 classes

Figure 3: Compensation of model mismatch assuming an arbitrary true distribution and ML, MCL and MMBN classifiers with NB structure. The generalization error of the optimal classifier is indicated by the dotted line (= True).

ficient — reasonable learning algorithms for classification purposes should, asymptotically, result in a Bayes consistent classifier in this setup. This result raises the question why good classification results have been reported in the literature, e.g. in (Pernkopf et al., 2012). We attribute these results to the model mismatch and the implemented early stopping heuristic: MM parameters are obtained by starting at the maximum likelihood solution and subsequent maximization of the margin objective by gradient ascent. This maximization is not performed until a locally optimal solution is obtained, but stopped after a certain number of steps, where the stopping time is determined using 5-fold cross-validation. Consequently, the authors do not actually compute an MMBN but a blend between a BN with generatively and discriminatively optimized parameters.

Furthermore, we observed that in some binary-class examples for which the true distribution can be represented by the model, MMBNs do not necessarily induce a Bayes optimal classifiers, cf. Section 4.2. There are consistent and inconsistent parameters that achieve the same margin objective. This suggests that one should exploit the degrees of freedom still remaining after achieving a certain margin to optimize some additional criterion, e.g. maximization of the entropy or maximization of the likelihood of the training data (Peharz et al., 2013).

7 CONCLUSION

In this paper, we presented results on Bayes consistency of MMBN classifiers with fully connected graphs. We provided examples where MMBN classifiers can be inconsistent and demonstrated experimentally that these classifiers are able to efficiently compensate model mismatch.

In future work, we aim to quantify the asymptotic *suboptimality* of MMBN classifiers in terms of the true distribution. We want to establish rates of convergence to the asymptotic performance. Furthermore, we aim at extending the definition of margin objective. In particular, Bayes consistency shall be achieved whenever the true distribution can be represented by the considered BNs.

Acknowledgements

This work was supported by the Austrian Science Fund (project numbers P22488-N23 and P25244-N15).

A MMBNS MAXIMIZE A LOWER BOUND OF THE CLASSIFICATION RATE

The ideal approach for learning BN classifiers with fixed structure $\mathcal{G}$ would be solving

$$\underset{\mathcal{B}' \in \mathcal{B}(\mathcal{G})}{\text{maximize}} \quad \mathbb{E}_{\mathrm{P}^*(C,\mathbf{X})}\left[\mathbf{1}\{h_{\mathcal{B}'}(\mathbf{X}) = C\}\right], \tag{22}$$

i.e. maximization of the expected classification rate. Directly finding a solution to this problem is difficult as $\mathrm{P}^*(C, \mathbf{X})$ is unknown. Even if $\mathrm{P}^*(C, \mathbf{X})$ would be available, the maximization in general corresponds to a hard nonlinear optimization problem. Therefore, approximations are needed.

Solving (22) is equivalent to solving

$$\underset{\mathcal{B}' \in \mathcal{B}(\mathcal{G})}{\text{maximize}} \quad \sum_{c,\mathbf{x}} \mathrm{P}^*(c, \mathbf{x}) \mathbf{1}\{h_{\mathcal{B}'}(\mathbf{x}) = c\}. \tag{23}$$

The expression $\mathbf{1}\{h_{\mathcal{B}'}(\mathbf{x}) = c\}$ equals 1 if and only if $h_{\mathcal{B}'}(\mathbf{x}) = c$, or equivalently if $\mathrm{P}^{\mathcal{B}'}(c, \mathbf{x}) > \mathrm{P}^{\mathcal{B}'}(c', \mathbf{x})$ for all $c' \neq c$ (ignoring the possibility of equally large joint probabilities), otherwise it is zero. In comparison, the corresponding term in (14) with $\gamma = 1$ is at most 1 and positive if and only if $\mathrm{P}^{\mathcal{B}'}(c, \mathbf{x}) > \mathrm{P}^{\mathcal{B}'}(c', \mathbf{x})$ for all $c' \neq c$. Otherwise it is negative. Consequently,

$$\begin{aligned} \min\left(1, \log \mathrm{P}^{\mathcal{B}'}(c, \mathbf{x}) - \max_{c' \neq c} \log \mathrm{P}^{\mathcal{B}'}(c', \mathbf{x})\right) \\ \leq \mathbf{1}\{h_{\mathcal{B}'}(\mathbf{x}) = c\}. \end{aligned} \tag{24}$$

This holds for all c and $\mathbf{x}$. Therefore, the MM objective (14) lower bounds asymptotically almost surely the classification rate (as *the empirical distribution converges to the true distribution* with increasing sample size).

B PROOF OF LEMMA 1

Proof. We give a proof by contradiction. Assume that $\mathcal{B}^{\mathrm{MM}} = (\mathcal{G}, \mathrm{P}^{\mathrm{MM}}(C, \mathbf{X}))$ is an MMBN trained on the training set $\mathcal{T}$ with empirical distribution $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$. Additionally, assume that the induced classifier $h_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$ is not optimal with respect to $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$. Thus, there exists an instantiation of the features $\mathbf{x}^f$ that is not optimally classified by $h_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$, i.e. for which

$$[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})} \setminus [C|\mathbf{x}^f]_{\mathrm{P}^{\mathcal{T}}(C,\mathbf{X})} \neq \emptyset. \tag{25}$$

Because of the binary class variable, the set $[C|\mathbf{x}^f]_{\mathrm{P}^{\mathcal{T}}(C,\mathbf{X})}$ consists only of a single element (otherwise deciding for any of the two classes is optimal).

We consider the cases $|[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}| = 1$ and $|[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}| = 2$ separately. Beforehand, note that since $\mathcal{G}$ is fully connected, i.e. $\mathcal{B}(\mathcal{G})$ is the set of *all* possible distributions over $(C, \mathbf{X})$, we can arbitrarily select the probabilities $\mathrm{P}^{\mathrm{MM}}(C = c, \mathbf{X} = \mathbf{x})$, as long as a correctly normalized distribution results. Consequently, we can select $\mathrm{P}^{\mathrm{MM}}(C = c|\mathbf{X})$ without changing $\mathrm{P}^{\mathrm{MM}}(\mathbf{X})$. We use this to show that the MMBN objective (14) can be strictly increased.

Case 1. If $[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$ consists of one element $c^f \in \mathrm{sp}(C)$, then there exists a $c^* \in \mathrm{sp}(C) \setminus \{c^f\}$ such that $\mathrm{P}^{\mathrm{MM}}(c^f|\mathbf{x}^f) > \mathrm{P}^{\mathrm{MM}}(c^*|\mathbf{x}^f)$ and such that $\mathrm{P}^{\mathcal{T}}(c^*|\mathbf{x}^f) > \mathrm{P}^{\mathcal{T}}(c^f|\mathbf{x}^f)$. We generate a new probability distribution $\tilde{\mathrm{P}}^{\mathrm{MM}}(C, \mathbf{X})$ from $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$ by setting

$$\tilde{\mathrm{P}}^{\mathrm{MM}}(c, \mathbf{x}) = \mathrm{P}^{\mathrm{MM}}(c, \mathbf{x}) \quad \forall\, \mathbf{x} \neq \mathbf{x}^f\ \forall c, \tag{26}$$

$$\tilde{\mathrm{P}}^{\mathrm{MM}}(c^f, \mathbf{x}^f) = \mathrm{P}^{\mathrm{MM}}(c^*, \mathbf{x}^f), \text{ and} \tag{27}$$

$$\tilde{\mathrm{P}}^{\mathrm{MM}}(c^*, \mathbf{x}^f) = \mathrm{P}^{\mathrm{MM}}(c^f, \mathbf{x}^f). \tag{28}$$

The distribution $\tilde{\mathrm{P}}^{\mathrm{MM}}(C, \mathbf{X})$ optimally classifies $\mathbf{x}^f$. Additionally, it has higher objective (14) than $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$. Consequently, $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$ is no MMBN.

Case 2. If $[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$ consists of two elements, both classes have posterior probabilities of 0.5 according to $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$. Therefore, in the objective (14) the sum

$$\begin{aligned} \sum_c \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}^f) \min\Big(\gamma, \log \mathrm{P}^{\mathrm{MM}}(c, \mathbf{x}^f) \\ - \max_{c' \neq c} \log \mathrm{P}^{\mathrm{MM}}(c', \mathbf{x}^f)\Big) \end{aligned} \tag{29}$$

evaluates to zero.

Let $c^*, c^f \in \mathrm{sp}(C)$ satisfy $\mathrm{P}^{\mathcal{T}}(c^*|\mathbf{x}^f) > \mathrm{P}^{\mathcal{T}}(c^f|\mathbf{x}^f)$. As above, we generate a new distribution $\tilde{\mathrm{P}}^{\mathrm{MM}}(C, \mathbf{X})$ that classifies $\mathbf{x}^f$ optimally and has higher objective. The distribution $\tilde{\mathrm{P}}^{\mathrm{MM}}(C, \mathbf{X})$ is generated from $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$ by setting

$$\tilde{\mathrm{P}}^{\mathrm{MM}}(c, \mathbf{x}) = \mathrm{P}^{\mathrm{MM}}(c, \mathbf{x}) \quad \forall\, \mathbf{x} \neq \mathbf{x}^f\ \forall c, \tag{30}$$

$$\tilde{\mathrm{P}}^{\mathrm{MM}}(c^f, \mathbf{x}^f) = \frac{1}{1 + \exp(\gamma)} \cdot \mathrm{P}^{\mathrm{MM}}(\mathbf{x}^f), \text{ and} \tag{31}$$

$$\tilde{\mathrm{P}}^{\mathrm{MM}}(c^*, \mathbf{x}^f) = \frac{\exp(\gamma)}{1 + \exp(\gamma)} \cdot \mathrm{P}^{\mathrm{MM}}(\mathbf{x}^f). \tag{32}$$

The terms in the objective (14) that change their value, sum up to

$$\begin{aligned} &\sum_c \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}^f) \min\Big(\gamma, \log \tilde{\mathrm{P}}^{\mathrm{MM}}(c, \mathbf{x}^f) \\ &\qquad - \max_{c' \neq c} \log \tilde{\mathrm{P}}^{\mathrm{MM}}(c', \mathbf{x}^f)\Big) \\ &= \gamma\left(\mathrm{P}^{\mathcal{T}}(c^*, \mathbf{x}^f) - \mathrm{P}^{\mathcal{T}}(c^f, \mathbf{x}^f)\right) \\ &> 0. \end{aligned}$$

As the objective increases, $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$ is not an MMBN. □

References

Guo, Y., Wilkinson, D., and Schuurmans, D. (2005). Maximum margin Bayesian networks. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence*, pages 233–242. UAI Press.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition*. Wiley-Interscience, 2 edition.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

Ng, A. Y. and Jordan, M. I. (2001). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 14*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Peharz, R., Tschiatschek, S., and Pernkopf, F. (2013). The most generative maximum margin Bayesian networks. (submitted).

Pernkopf, F., Wohlmayr, M., and Tschiatschek, S. (2012). Maximum margin Bayesian network classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):521–531.

Platt, J. (1999). Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods-Support Vector Learning*, pages 1–21.

Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., and Tirri, H. (2005). On discriminative Bayesian network classifiers and logistic regression. *Machine Learning*, 59(3):267–296.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. (2007). Pegasos: Primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, pages 807–814.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Wettig, H., Grünwald, P., Roos, T., Myllymaki, P., and Tirri, H. (2003). When discriminative learning of Bayesian network parameters is easy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 491–496.