ASYMPTOTIC PROPERTIES OF MAXIMUM COLLECTIVE CONDITIONAL LIKELIHOOD ESTIMATORS FOR NAIVE BAYES CLASSIFIERS

PRIYANTHA WIJAYATUNGA AND SHIGERU MASE

Abstract. Bayesian networks that are probabilistic expert systems can be used as classifiers. A special type of Bayesian networks called naive Bayes classifiers are very popular in practice due to their good performance although they are relatively simple.

Enhancement of the performance of naive Bayes classifiers is often done through various parameter learning methods where the usual method is the method of maximum likelihood estimation. Nevertheless, since the true target of interest of Bayes classifiers are estimation of conditional probabilities, it is natural to learn their parameters by maximization of collective conditional likelihoods. Therefore recently there has been a growing interest in learning the parameters of naive Bayes classifiers through maximizing collective conditional likelihoods.

Strong consistency and asymptotic normality are two basic statistical properties which any decent estimators should have although they are primarily of theoretical nature. In this research, we prove the strong consistency and the asymptotic normality of the maximum collective conditional likelihood estimators for the naive Bayes classifiers. Essentially our proof follows the classical ideas well-developed for the theory of maximum likelihood estimators.

1. Introduction

A Bayesian network models probabilistic relationships among a set of random variables through a causal factorization of the joint density. There are efficient inference algorithms to calculate marginal and conditional densities of certain variable(s) given some of other variables. Therefore Bayesian networks are also referred to as probabilistic expert systems. In this section, we discuss some of the preliminaries of Bayesian networks and point out their real world applications in various fields. For extensive discussions on the topic, the reader is referred to Cowell et al. (1999) and Pearl (1988) and references therein.

Let us denote the state space of a random variable X_k by $\mathcal{X}_k$. We define a Bayesian network on a vector of n discrete finite random variables $X = (X_1, \ldots, X_n)$ as follows. From the chain rule of probability, we have

$$p(x) = p(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{v=2}^{n} p(x_v \mid x_1, \ldots, x_{v-1})$$

where $p(x_v \mid x_1, \ldots, x_{v-1})$ is the conditional density

$$\boldsymbol{P}(X_v = x_v \mid X_1 = x_1, \ldots, X_{v-1} = x_{v-1}).$$

Let X_v be conditionally independent of $\{X_1, \ldots, X_{v-1}\} \setminus X_{pa(v)}$ given $X_{pa(v)}$ is true, where $X_{pa(v)} \subseteq \{X_1, \ldots, X_{v-1}\}$, i.e., $p(x_v \mid x_1, \ldots, x_{v-1}) = p(x_v \mid x_{pa(v)})$ holds for $v = 1, \ldots, n$. Note that a conditional independence statement express an *information irrelevance*, meaning, in the above case, given $X_{pa(v)}$, the conditional

The first author would like to thank Ministry of Education, Culture, Sports, Science and Technology of Japan for supporting his research.

density of X_v can be fully defined irrespective of the realization of the variable set $\{X_1, \ldots, X_{v-1}\} \setminus X_{pa(v)}$. When *causality* is concerned, this means that, given some causes, other causes become irrelevant or neutral. Thus we have the following factorization of the joint probability density:

$$p(x) = \prod_{v=1}^{n} p(x_v \mid x_{pa(v)}).$$

This factorization can be represented by a *directed acyclic graph* where arrows are drawn so as to represent *probabilistic influences* among variables. Arrows are drawn from each variable in the set $X_{pa(v)}$ to X_v for $v = 1, \ldots, n$ and the variable set $X_{pa(v)}$ is called the *parents* of X_v. The *children* of a variable is defined by the opposite relation. And for each v there exits a family of conditional densities $\{p(x_v \mid x_{pa(v)}) : \forall x_{pa(v)}\}$ where each conditional density $p(x_v \mid x_{pa(v)})$ is assumed to be multinomial with its parameter vector, say, $\theta_{v|x_{pa(v)}} = (\theta_{x_v|x_{pa(v)}} : \forall x_v \in \mathcal{X}_v)$, which is written schematically as

$$X_v \mid x_{pa(v)}, \theta_{v|x_{pa(v)}} \sim Mn(\theta_{v|x_{pa(v)}}).$$

In the non-Bayesian setting, $\theta_{x_v|x_{pa(v)}} = \boldsymbol{P}(X_v = x_v \mid X_{pa(v)} = x_{pa(v)})$.

It is clear that, in order to use a Bayesian network model, one needs to learn the model from data on the variables of interest, perhaps along with subject domain expert knowledge. Basically, this model building process consists of two part; (1) network structure learning – extracting the conditional independence relationships among variables, (2) parameter learning given the structure –finding the factorization of the joint density numerically. In literature there are many algorithms developed for these tasks.

A simple Bayesian network is illustrated in the Fig. 1. If we have observed, say $X_1 = x_1$ and $X_5 = x_5$, then we can calculate conditional probability densities such as $p(x_3 \mid x_1, x_5)$ efficiently using inference algorithms.

Bayesian networks have vast number of applications in such fields as medicine (Cowell et al. (1993), Diez et al. (1997)), engineering (Bromley et al. (2005), Yearling and Hand (2003)), economics and finance (Cui et al. (2006), Gemela (2001), Wijayatunga et al. (2006)), law (Dawid et al. (2006)), etc. In fact, they are applicable to any field which deal with uncertainty and reasoning.

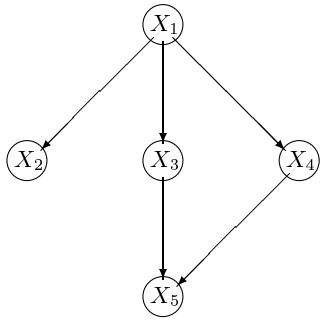

Fig.1 A simple Bayesian network with the density $p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_1)p(x_5 \mid x_3, x_4)$.

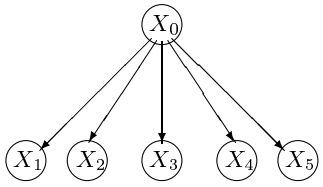

Fig.2 A naive Bayes network with the density $p(x_0, x_1, x_2, x_3, x_4, x_5) = p(x_0) \prod_{i=1}^{5} p(x_i \mid x_0)$.

It is straightforward to use Bayesian networks as classifiers (Friedman and Goldszmidt (1996), Friedman et al. (1997)) where the interest is usually on conditional probability densities of a particular random variable called the "class" given the data on some of the other variables called "attributes" that are statistically related

with it. One advantage of Bayesian networks is that classification can be done with partial observations on attributes. In fact, there is a lot of applications of Bayesian networks as classifiers in various subject domains. See, for example, Baesens et. al. (2002), Wijayatunga et al. (2006) and Porwal et al. (2006).

In Fig. 1, if the class variable is X_3 and we consider all the other variables as attributes, then we are interested in conditional densities $p(x_3 \mid x_1, x_2, x_4, x_5)$ where

$$p(x_3 \mid x_1, x_2, x_4, x_5) = \frac{p(x_3 \mid x_1)p(x_5 \mid x_3, x_4)}{\sum_{x'_3} p(x'_3 \mid x_1)p(x_5 \mid x'_3, x_4)} = p(x_3 \mid x_1, x_4, x_5).$$

Therefore, conditional probability of the class variable given all the other variables is fully determined by a subset of attributes which is called the *Markov blanket* of the class, namely, $\{X_1, X_4, X_5\}$. The Markov blanket of a variable is the set of all its parents, children and parents of those children. When we have complete observations on Markov blanket variables of the class, we need to consider only the joint density of $p(x_1, x_3, x_4, x_5)$ for predictions on the class variable but one should notice that some parameters are not necessary for the calculation of $p(x_3 \mid x_1, x_4, x_5)$, for example, $\theta_{x_4|x_1}$.

When we build a Bayesian network classifier for a chosen class variable, it seems that ideally we can relate the other variables to it as either its parents, or children or parents of those children using a Bayesian network structure learning algorithm. However, a special type of Bayesian networks called naive Bayes classifier is very popular due to its simplicity and good performance in practice.

In naive Bayes networks it is assumed that all the attributes are conditionally independent given the class. For instance, a naive Bayes network with class variable X_0 and attribute variable vector $(X_1, \ldots, X_5)$ is shown in the Fig. 2. Therefore the structure of naive Bayes classifier is predefined and what is left to do is the learning of parameters. This can be done using methods of maximum likelihood estimation, Bayesian estimation when prior knowledge is available, etc.

Those general parameter learning methods often give naive Bayes classifiers good performance in practice. Further there has been much research done to enhance their classification accuracy through various parameter learning techniques (Langley (1993), Webb and Pazzani (1998), Wijayatunga et al. (2006)) and sometimes by changing the structure of the network (Friedman and Goldszmidt (1996)).

Recently, there has been a growing interest on parameter learning of naive Bayes classifiers through maximization of collective conditional likelihood (CCL) since it is more natural in the sense that it deals directly with conditional densities of the class given the attributes, which are the object of interest of a classifier rather than the joint density as in the maximum likelihood estimation. Some of the references are Friedman et al. (1997), Rubinstein and Haste (1997), Ng and Jordon (2001), Bouchard and Trigg (2004), Chelba and Acero (2004), Grossman and Domingo (2004), Jing et al. (2005), Greiner et al. (2005), Pernkopf and Blimes (2005), Ross et al. (2005), Santafe et al. (2005) and Yakhnenko et al. (2005). Further the idea of maximization of CCL is itself interesting as a new construction principle of parameter estimators because it incorporates all the relevant conditional likelihoods simultaneously.

In the literature, parameter learning through maximizing CCL is referred as *discriminative* or *supervised* learning whereas that of through maximizing likelihood is referred as *generative* or *unsupervised* learning. It has been reported that very often the maximum CCL estimators (MCCLEs) give a classifier a better performance than maximum likelihood estimates (MLEs) (Chelba and Acero (2004); Greiner et

al. (2005); Pernkopf and Blimes (2005)), although Ng and Jordon (2001) give examples where MLEs have better performance. A demerit of MCCLE is that it has no closed form formula and one needs to calculate it numerically.

The strong consistency and asymptotic normality are two important basic properties that any decent estimator should have. But we notice that so far there are no proofs for MCCLE for naive Bayes classifiers for which it is usually applied. Although naive Bayes networks are mainly used as operational models, MCCLE should have these two asymptotic properties if they "were" true models. Our arguments follow essentially the classical ideas of proving strong consistency and asymptotic normality proof developed for asymptotic theory of MLE.

Consider a set $X = \{X_0, X_1, \ldots, X_n\}$ of $n+1$ discrete and finite random variables. X_0 is the class variable and the others are attribute variables. The state space of X_i is $\mathcal{X}_i = \{1, \ldots, a_i\}$. Assume they form a naive Bayesian network so that their joint density is

$$p(x_0, x_1, \ldots, x_n) = p(x_0) \prod_{i=1}^{n} p(x_i \mid x_0). \tag{1}$$

The conditional density of X_0 given $X_1 = x_1, \ldots, X_n = x_n$ is

$$p(x_0 \mid x_{[n]}) = \frac{p(x_0) \prod_{i=1}^{n} p(x_i \mid x_0)}{\sum_{x_0'} p(x_0') \prod_{i=1}^{n} p(x_i \mid x_0')}$$

where, and in the following, we frequently use the notation like $x_{[n]}$ meaning the ordered set $\{x_1, \ldots, x_n\}$ for convenience.

Let the parameter space Θ be $\Delta_{a_0} \times \Delta_{a_1}^{a_0} \times \cdots \times \Delta_{a_n}^{a_0}$ where $\Delta_t = \{(p_1, \ldots, p_{t-1}) : 0 \le p_i, \sum_{i=1}^{t-1} p_i \le 1\}$. The interior of Θ is denoted by Θ^o. In the following, we always assume that the true parameter is an element of Θ^o. Note that, if not, the naive Bayesian network is degenerated in the sense that some state space can be got rid of or made smaller. The naive Bayesian model is parameterized with $\theta = (\theta_{x_0}, \theta_{x_1|x_0}, \ldots, \theta_{x_n|x_0}) \in \Theta$ where $x_i \in \mathcal{X}_i$ for $i = 0, 1, \ldots, n$ so that

$$\begin{aligned}
p_\theta(x_0) &= \theta_{x_0} \text{ if } x_0 = 1, \ldots, a_0 - 1, \\
p_\theta(a_0) &= 1 - \sum_{x_0=1}^{a_0-1} \theta_{x_0}, \\
p_\theta(x_i \mid x_0) &= \theta_{x_i|x_0} \text{ if } x_i = 1, \ldots, a_i - 1, \\
p_\theta(a_i \mid x_0) &= 1 - \sum_{x_i=1}^{a_i-1} \theta_{x_i|x_0}.
\end{aligned}$$

From (**??**), its maximum likelihood estimator is easy to obtain and nothing but mere corresponding sample ratios which are sometimes meaningless because of lacks of sufficient data.

Suppose we have N complete samples on the random vector $(X_0, X_1, \ldots, X_n)$, say, $\boldsymbol{x} = \{(x_0^{(1)}, x_1^{(1)}, \ldots, x_n^{(1)}), \ldots, (x_0^{(N)}, x_1^{(N)}, \ldots, x_n^{(N)})\}$. Then the collective conditional likelihood (CCL) of θ given data $\boldsymbol{x}$ is defined as

$$CCL_N(\theta) = \prod_{j=1}^{N} p_\theta(x_0^{(j)} \mid x_{[n]}^{(j)}) = \prod_{j=1}^{N} \frac{p_\theta(x_0^{(j)}) \prod_{i=1}^{n} p_\theta(x_i^{(j)} \mid x_0^{(j)})}{\sum_{x_0'} p_\theta(x_0') \prod_{i=1}^{n} p_\theta(x_i^{(j)} \mid x_0')}. \tag{2}$$

Also we can write

$$CCL_N(\theta) = \prod_{x_0, x_{[n]}} p_\theta(x_0 \mid x_{[n]})^{N_{x_0|x_{[n]}}} = \prod_{x_{[n]}} CL_N(\theta \mid x_{[n]})$$

where $N_{x_0|x_{[n]}}$ is the number of cases in the data such that $X_0 = x_0$ given $X_{[n]} = x_{[n]}$ and

$$CL_N(\theta \mid x_{[n]}) = \prod_{x_0} p_\theta(x_0 \mid x_{[n]})^{N_{x_0|x_{[n]}}}$$

are conditional likelihoods, that is, likelihoods of conditional distributions.

The maximum collective conditional likelihood estimator (MCCLE) is defined as

$$\hat{\theta}_N = \underset{\theta \in \Theta}{\operatorname{argmax}}\, CCL_N(\theta).$$

MCCLE's have no closed form expression in general and should be solved numerically.

2. Strong consistency of MCCLE

In this section, we prove the strong consistency of MCCLE's. The proof is based on the classical strong consistency proof of MLE's due to Wald. Readers can find a good outline of Wald's idea in, e.g., Cox and Hinkley (1974) and Stuart and Ord (1991). A different approach is given in Hall and Heyde (1980). First, we need the following identifiability assumption as in the MLE case. This condition requires that θ should be uniquely determined by the corresponding density $p_\theta(. \mid .)$.

Assumption 1 (Identifiability Condition) If $p_\theta(x_0 \mid x_{[n]}) = p_{\theta'}(x_0 \mid x_{[n]})$ for all x, then $\theta = \theta'$.

Equivalently, this assumption says that, if $\theta \neq \theta'$, there exists some $x_{[n]}$ such that $p_\theta(. \mid x_{[n]}) \neq p_{\theta'}(. \mid x_{[n]})$. For such $x_{[n]}$, we have

$$\sum_{x_0} p_{\theta'}(x_0 \mid x_{[n]}) \log \frac{p_\theta(x_0 \mid x_{[n]})}{p_{\theta'}(x_0 \mid x_{[n]})} < 0,$$

which is a particular case of the so-called information inequality.

Remark 1. This assumption may not always be true. The following is a simple counter example. Let $n = 1$ and $\mathcal{X}_0 = \mathcal{X}_1 = \{1, 2\}$. Let

$$(\theta_{x_0=1}, \theta_{x_1=1|x_0=1}, \theta_{x_1=1|x_0=2}) = (1/3, 2/3, 2/3),$$
$$(\theta'_{x_0=1}, \theta'_{x_1=1|x_0=1}, \theta'_{x_1=1|x_0=2}) = (1/3, 1/4, 1/4),$$

then

$$p_\theta(x_0 = 1 \mid x_1 = 1) = p_{\theta'}(x_0 = 1 \mid x_1 = 1),$$
$$p_\theta(x_0 = 1 \mid x_1 = 2) = p_{\theta'}(x_0 = 1 \mid x_1 = 2).$$

Thus we have found parameters $\theta \neq \theta'$ for which corresponding conditional distributions are all identical.

The following lemma is essential for the consistency proof.

Lemma 1. *Let $\theta^* \in \Theta^o$ be a true parameter and $\epsilon > 0$. Let $U = \{\theta : \|\theta - \theta^*\| < \epsilon\}$ and $\Theta_1 = \Theta \setminus U$. ϵ is chosen so that $U \subset \Theta^o$. Define $CCL_N(\Theta_1) = \max_{\theta \in \Theta_1} CCL_N(\theta)$. Then, under Assumption 1,*

$$\boldsymbol{P}_{\theta^*}\left\{ \lim_{N \to \infty} \frac{CCL_N(\Theta_1)}{CCL_N(\theta^*)} = 0 \right\} = 1.$$

Proof. For suitable $d > 0$ and $\theta \in \Theta_1$, define

$$p_{\theta,d}(x_0 \mid x_{[n]}) = \max\{p_\tau(x_0 \mid x_{[n]}) : \|\tau - \theta\| \le d\},$$
$$CCL_N(\theta, d) = \prod_{i=1}^{N} p_{\theta,d}(x_0^{(i)} \mid x_{[n]}^{(i)}),$$
$$K(\theta^*, \theta, d \mid z_{[n]}) = \sum_{x_0} p_{\theta^*}(x_0 \mid z_{[n]}) \log\left\{\frac{p_{\theta,d}(x_0 \mid z_{[n]})}{p_{\theta^*}(x_0 \mid z_{[n]})}\right\}.$$

After some manipulations, we get

$$\begin{aligned}
&\frac{1}{N}\log\left\{\frac{CCL_N(\theta, d)}{CCL_N(\theta^*)}\right\} \\
&\qquad = \frac{1}{N}\log\left\{\prod_{i=1}^{N}\frac{p_{\theta,d}(x_0^{(i)} \mid x_{[n]}^{(i)})}{p_{\theta^*}(x_0^{(i)} \mid x_{[n]}^{(i)})}\right\} \\
&\qquad = \sum_{z_{[n]}}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{x_{[n]}^{(i)}=z_{[n]}}\log\left\{\frac{p_{\theta,d}(x_0^{(i)} \mid z_{[n]})}{p_{\theta^*}(x_0^{(i)} \mid z_{[n]})}\right\},
\end{aligned}$$

where $\mathbb{1}_A$ is equal to 1 (resp. 0) if the statement A is true (resp. false). Taking the expectation w.r.t. $p_{\theta^*}(x_0, x_{[n]})$ for given $z_{[n]}$,

$$\begin{aligned}
&\boldsymbol{E}_{\theta^*}\left\{\mathbb{1}_{x_{[n]}=z_{[n]}}\log\left\{\frac{p_{\theta,d}(x_0 \mid z_{[n]})}{p_{\theta^*}(x_0 \mid z_{[n]})}\right\}\right\} \\
&\qquad = \sum_{x_0, x_{[n]}} p_{\theta^*}(x_0, x_{[n]})\mathbb{1}_{x_{[n]}=z_{[n]}}\log\left\{\frac{p_{\theta,d}(x_0 \mid z_{[n]})}{p_{\theta^*}(x_0 \mid z_{[n]})}\right\} \\
&\qquad = \sum_{x_0} p_{\theta^*}(x_0, z_{[n]})\log\left\{\frac{p_{\theta,d}(x_0 \mid z_{[n]})}{p_{\theta^*}(x_0 \mid z_{[n]})}\right\} \\
&\qquad = p_{\theta^*}(z_{[n]})\sum_{x_0} p_{\theta^*}(x_0 \mid z_{[n]})\log\left\{\frac{p_{\theta,d}(x_0 \mid z_{[n]})}{p_{\theta^*}(x_0 \mid z_{[n]})}\right\} \\
&\qquad = p_{\theta^*}(z_{[n]})K(\theta^*, \theta, d \mid z_{[n]}).
\end{aligned}$$

Further we have

$$\lim_{d \downarrow 0} K(\theta^*, \theta, d \mid z_{[n]}) = K(\theta^*, \theta, 0 \mid z_{[n]})$$

and from the information inequality $K(\theta^*, \theta, 0 \mid z_{[n]}) \le 0$. By the strong law of large numbers, we get

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{x_{[n]}^{(i)}=z_{[n]}}\log\left\{\frac{p_{\theta,d}(x_0^{(i)} \mid z_{[n]})}{p_{\theta^*}(x_0^{(i)} \mid z_{[n]})}\right\} = p_{\theta^*}(z_{[n]})K(\theta^*, \theta, d \mid z_{[n]}) \le 0$$

$\boldsymbol{P}_{\theta^*}$-a.s. Therefore

$$\begin{aligned}
&\lim_{N\to\infty}\frac{1}{N}\log\left\{\frac{CCL_N(\theta,d)}{CCL_N(\theta^*)}\right\}\\
&\quad=\lim_{N\to\infty}\sum_{z_{[n]}}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{x_{[n]}^{(i)}=z_{[n]}}\log\left\{\frac{p_{\theta,d}(x_0^{(i)}\mid z_{[n]})}{p_{\theta^*}(x_0^{(i)}\mid z_{[n]})}\right\}\\
&\quad=\sum_{z_1,\ldots,z_n}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{x_{[n]}^{(i)}=z_{[n]})}\log\left\{\frac{p_{\theta,d}(x_0^{(i)}\mid z_{[n]})}{p_{\theta^*}(x_0^{(i)}\mid z_{[n]})}\right\}\\
&\quad\equiv K(\theta^*,\theta,d)\quad \boldsymbol{P}_{\theta^*}\text{-a.s.},
\end{aligned}$$

where $K(\theta^*,\theta,d)=\sum_{z_{[n]}}p_{\theta^*}(z_{[n]})K(\theta^*,\theta,d\mid z_{[n]})$. If the identifiability assumption is true, then for any $\theta\in\Theta_1$ we have $K(\theta^*,\theta,0)<0$ and, therefore, we can select a small number $d>0$ such that $K(\theta^*,\theta,d)<0$.

Since Θ_1 is a compact set, we can find finitely many points $\theta_1,\ldots,\theta_m$ in Θ_1 such that $\{U_{d(\theta_j)}(\theta_j):j=1,\ldots,m\}$ makes a covering of Θ_1 by open subset $U_{d(\theta)}(\theta)=\{\theta'\in\Theta:\|\theta'-\theta\|<d(\theta)\}$ of Θ, where $d(\theta)$ is a positive number possibly depending on θ such that $K(\theta^*,\theta,d(\theta))<0$. Therefore,

$$\Theta_1\subset\bigcup_{j=1}^{m}U_{d(\theta_j)}(\theta_j)$$

So, we have

$$\begin{aligned}
&\limsup_{N\to\infty}\frac{1}{N}\log\frac{CCL_N(\Theta_1)}{CCL_N(\theta^*)}\\
&\quad\le\lim_{N\to\infty}\max\left\{\frac{1}{N}\log\frac{CCL_N(\theta_j,d(\theta_j))}{CCL_N(\theta^*)}:j=1,\ldots,m\right\}\\
&\quad=\max\left\{K(\theta^*,\theta_j,d(\theta_j)):j=1,\ldots,m\right\}\quad \boldsymbol{P}_{\theta^*}\text{-a.s.}
\end{aligned}$$

Let $K=\max_j K(\theta^*,\theta_j,d(\theta_j))$, then $K<0$ and

$$\limsup_{N\to\infty}\frac{1}{N}\log\frac{CCL_N(\Theta_1)}{CCL_N(\theta^*)}\le K<0\quad \boldsymbol{P}_{\theta^*}\text{-a.s.}$$

This implies that

$$\lim_{N\to\infty}\frac{CCL_N(\Theta_1)}{CCL_N(\theta^*)}=0\quad \boldsymbol{P}_{\theta^*}\text{-a.s.}$$

That is, we have proved that

$$\boldsymbol{P}_{\theta^*}\left\{\lim_{N\to\infty}\frac{CCL_N(\Theta_1)}{CCL_N(\theta^*)}=0\right\}=1.$$

□

Now we can prove the strong consistency of MCCLE immediately.

Theorem 1. *Under Assumption 1, MCCLE $\hat{\theta}_N$ is strongly consistent.*

Proof. Assume the contrary, i.e., $P_{\theta^*}\{\hat{\theta}_N\nrightarrow\theta^*\}>0$. Then there exists some $d>0$ such that

$$P_{\theta^*}\left\{\hat{\theta}_N\notin U_d(\theta^*)\text{ for infinitely many }N\right\}>0.$$

Then

$$P_{\theta^*}\left\{1\le\frac{CCL_N(\hat{\theta}_N)}{CCL_N(\theta^*)}\text{ for infinitely many }N\right\}>0.$$

But by the preceding Lemma **??**, the above probability should be equal to zero. Therefore we have a contradiction that completes the proof. □

Corollary 1. *$p_{\hat{\theta}_N}(x_0 \mid x_{[n]})$ is strongly consistent estimators of $p_{\theta^*}(x_0 \mid x_{[n]})$ for each $x_{[n]}$.*

Proof. Immediate from the theorem since $p_\theta(x_0 \mid x_{[n]})$ are rational functions of the parameter which have no poles in Θ^o. □

3. Asymptotic Normality of the MCCLE

In this section, we prove the asymptotic normality of the MCCLE. We apply a general asymptotic normality theorem from van der Vaart (1998). For convenience, we cite the theorem below.

Theorem 2 (Theorem 5.41 of van der Vaart (1998))**.** *For each θ in an open subset of Euclidean space $\mathbb{R}^d$, let $\theta \mapsto \psi_\theta(x) \in \mathbb{R}^d$ be twice continuously differentiable function for every $x \in \mathbb{R}^M$. Let θ^* be the true parameter and $\boldsymbol{X} = (X_1, \ldots, X_N)$ be a corresponding iid data. Suppose that $\boldsymbol{E}_{\theta^*}\{\psi_{\theta^*}(X)\} = 0$, that $\boldsymbol{E}_{\theta^*}\{\|\psi_{\theta^*}(X)\|^2\} < \infty$ and that the matrix $\boldsymbol{E}_{\theta^*}\{\dot{\psi}_{\theta^*}(X)\}$ exists and nonsingular where $\dot{\psi}_\theta$ stands for the derivative w.r.t. θ. Assume the second-order partial derivatives of $\psi_\theta(x)$ w.r.t. θ are dominated by a fixed integrable function $\phi(x)$ uniformly for every θ in a neighborhood of θ^*. If the estimator $\hat{\theta}_N$, a zero of $\Psi_N(\theta) = \frac{1}{N}\sum_{i=1}^N \psi_\theta(X_i) = 0$, is weakly consistent, then*

$$\sqrt{N}(\hat{\theta}_N - \theta^*) = -\big(\boldsymbol{E}_{\theta^*}\dot{\psi}_{\theta^*}(X)\big)^{-1}\frac{1}{\sqrt{N}}\sum_{l=1}^N \psi_{\theta^*}(X_i) + o_P(1).$$

In particular, the sequence $\sqrt{N}(\hat{\theta}_N - \theta^*)$ is asymptotically normal with mean zero and covariance matrix

$$\big(\boldsymbol{E}_{\theta^*}\{\dot{\psi}_{\theta^*}(X)\}\big)^{-1}\boldsymbol{E}_{\theta^*}\{\psi_{\theta^*}(X)\psi_{\theta^*}(X)^T\}\big(\boldsymbol{E}_{\theta^*}\{\dot{\psi}_{\theta^*}(X)\}\big)^{-1}.$$

Before starting the proof of asymptotic normality of MCCLE, we introduce some notations. It is convenient to rewrite the parameter vector as $\theta = (\theta_1, \ldots, \theta_k)$ where $k = (a_0 - 1) + \sum_{i=1}^n (a_i - 1)a_0$ using some appropriate enumeration. Define the function $\psi_\theta(x) = \frac{\partial}{\partial\theta}\log p_\theta(x_0 \mid x_{[n]})$. A solution of the estimating equation

$$\Psi_N(\theta) \equiv \frac{1}{N}\sum_{\ell=1}^N \psi_\theta(X^{(\ell)}) = 0$$

is the MCCLE $\hat{\theta}_N$.

In order to apply Theorem **??**, we check its requirements as a series of lemmas and an assumption as follows. Let $\theta^* \in \Theta^o$ be the true parameter.

Lemma 2. *$\psi_\theta(x)$ is twice continuously differentiable in $\theta \in \Theta^o$ for every x.*

Proof. Easy since relevant conditional probabilities are rational functions in θ which have no poles in Θ^o. □

Lemma 3. *$\boldsymbol{E}_{\theta^*}\{\psi_{\theta^*}(X)\} = 0$ holds.*

Proof. First note

$$\sum_{x_0}\frac{\partial}{\partial\theta}p_\theta(x_0 \mid x_{[n]}) = \frac{\partial}{\partial\theta}\sum_{x_0}p_\theta(x_0 \mid x_{[n]}) = \frac{\partial}{\partial\theta}1 = 0$$

for every $x_{[n]}$ and θ. Hence the assertion follows from the following relation:

$$\begin{aligned}
\boldsymbol{E}_\theta\{\psi_\theta(X)\} &= \sum_{x_0,x_{[n]}} p_\theta(x_0,x_{[n]})\Big\{\frac{\partial}{\partial\theta}\log p_\theta(x_0 \mid x_{[n]})\Big\} \\
&= \sum_{x_{[n]}} p_\theta(x_{[n]}) \sum_{x_0} p_\theta(x_0 \mid x_{[n]})\Big\{\frac{\partial}{\partial\theta}\log p_\theta(x_0 \mid x_{[n]})\Big\} \\
&= \sum_{x_{[n]}} p_\theta(x_{[n]}) \sum_{x_0} \frac{\partial}{\partial\theta} p_\theta(x_0 \mid x_{[n]}) \\
&= \sum_{x_{[n]}} p_\theta(x_{[n]}) \frac{\partial}{\partial\theta} \sum_{x_0} p_\theta(x_0 \mid x_{[n]}) \\
&= \sum_{x_{[n]}} p_\theta(x_{[n]}) \frac{\partial}{\partial\theta} 1.
\end{aligned}$$

□

Lemma 4. $\boldsymbol{E}_{\theta^*}\{\|\psi_{\theta^*}(X)\|^2\} < \infty$.

Proof. This is obvious since the expectation is a finite sum and the integrand is finite. □

Assumption 2. (Non-singularity of asymptotic covariance matrix) The matrix $V_{\theta^*} = \boldsymbol{E}_{\theta^*}\{\dot{\psi}_{\theta^*}(X)\}$ is non-singular.

Remark 2. The matrix exists but may not always be non-singular. In Appendix A we give a decomposition of the matrix V_θ into those of conditional distributions and give a counter example.

Lemma 5. *There exists an integrable function $\phi(x)$ such that $\|\ddot{\psi}_\theta(x)\| \leq \phi(x)$ uniformly in x for every θ in a neighborhood of θ^*.*

Proof. This is again straightforward. Possible x are finitely many and each element of $\ddot{\psi}_\theta(x)$ is a rational function in θ which has no pole in Θ^o. Actually $\phi(x)$ can be a constant which is the maximum of maxima of absolute value of each elements of $\ddot{\psi}_\theta(x)$ w.r.t. both x and $\theta \in C$ where $C \subset \Theta^o$ is a compact neighborhood of θ^*. □

Now we can state the following theorem of asymptotic normality of MCCLE, the proof of which is immediate from Theorem 2 and previous lemmas.

Theorem 3. *Assume Assumptions 1 and 2 and let the true parameter $\theta^* \in \Theta^o$. Let $\hat{\theta}_N$ be MCCLE. Then*

$$\sqrt{N}(\hat{\theta}_N - \theta^*) = -\big(\boldsymbol{E}_{\theta^*}\{\dot{\psi}_{\theta^*}(X)\}\big)^{-1} \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \psi_{\theta^*}(X_i) + o_P(1).$$

In particular, the sequence $\sqrt{N}(\hat{\theta}_N - \theta^*)$ is asymptotically normal with mean zero and covariance matrix

$$\big(\boldsymbol{E}_{\theta^*}\{\dot{\psi}_{\theta^*}(X)\}\big)^{-1} \boldsymbol{E}_{\theta^*}\{\psi_{\theta^*}(X)\psi_{\theta^*}(X)^T\} \big(\boldsymbol{E}_{\theta^*}\{\dot{\psi}_{\theta^*}(X)\}\big)^{-1}.$$

4. Conclusion

We have proved the strong consistency and the asymptotic normality of MCCLE for naive Bayes classifier. It is interesting, and is natural in a sense from the definition of MCCLEs, that quantities of interest such as $K(\theta^*, \theta, d)$ and V_{θ^*} of the collective conditional likelihood function in our proofs are of the form of weighted averages of corresponding quantities of likelihood functions related to individual

conditional densities $p_{\theta^*}(x_0 \mid x_{[n]})$ where weights are probabilities of conditioning set $p_{\theta^*}(x_{[n]})$.

The assumption of identifiability of parameters is essential in our case of naive Bayes network. It is indispensable as Example 1 shows. We cannot give a detailed characterization of the cases that this condition are violated. But it seems such a case is rather pathological and/or has a degenerated structure, and the condition will be satisfied for almost all practically important situations. Note that, without this condition, it may be meaningless to estimate parameters. A similar comment also applies to the assumption of nonsingularity of the matrix V_{θ^*} in the asymptotic normality proof of MCCLE.

It has been shown that very often MCCLE-trained naive Bayes model has a better classification performance than that of MLE-trained. These empirical evidence are often found by performing cross-validations of each model on various data sets. But better models in terms of prediction or classification accuracy are often questioned for over-fitting, especially in the cases of these types of model training and validations. Strong consistency of MCCLE eliminates such problems against it given that the estimator is obtained from a data sample that is sufficiently large. But the problem remains how large enough should the data set be so as to have consistent classification performance for the future unseen data cases.

MLEs for Bayesian networks are mere sample conditional probabilities and may not be defined or effective because of lack of sufficient number of data. It should be stressed that MCCLEs can be useful even when MLEs fail.

Finally it remains one important question which we cannot discuss here. Are the MCCLE-estimated conditional classification probabilities $p_{\hat{\theta}_N}(x_0 \mid x_{[n]})$ asymptotically normal or not? At least, our proof shows that they are rational functions of asymptotically normal parameters.

References

[1] Baesens, B., Egmont-Petersen, M. Castelo, R. and Vanthienen, J., 2002, "Learning Bayesian Network Classifiers for Credit Scoring Using Markov Chain Monte Carlo Search", Proc. of International Congress on Pattern Recognition, IEEE Computer Society, pp. 49-52.

[2] Bouchard, G., and Triggs, B., 2004, "The Trade-off Between Generative and Discriminative Classifiers", Proc. 16^{th} Symposium of IASC (COMPSTAT'2004), Prague, Springer, pp. 721-728.

[3] Bromley, J., Jackson, N. A., Clymer, O. J., Giacomello, A. M., and Jensen, F. V., 2005, "The Use of Hugin to Develop Bayesian Networks as an Aid to Integrated Water Resource Planning", Environmental Modelling and Software, 20, pp. 231-242.

[4] Chelba, C., and Acero, A., 2004, "Conditional Maximum Likelihood Estimation of Naive Bayes Probability Models Using Rational Function Growth Transform", Microsoft Research Technical Report MSR-TR-2004-33, Microsoft Corporation, Redmond, WA. USA.

[5] Cowell, R. G., Dawid, A. P., Hutchinson, T. A., Roden, S. R., and Spiegelhalter, D. J., 1993, "Bayesian Networks for Analysis of Drug Safety", The Statistician, 42, 369-384.

[6] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J., 1999, "Probabilistic Expert Systems", Springer, NY, USA.

[7] Cox, D. R., and Hinkley, D. V., 1974, "Theoretical Statistics", Chapman and Hall, London, UK Chap. 9.

[8] Cui, G., Wong, M. L. and Lui, H., 2006, "Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming", Management Science, 52(4), pp. 597-612.

[9] Dawid, A. P., Morteria, J., and Vicard, P., 2006, "Representing and Solving Complex DNA Identification Cases Using Bayesian Networks", Progress in Forensic Genetics, 11 (Proc. of the 21^{st} International ISFG Congress), International Congress Series, 1288, Elsevier Science, Amsterdam, The Netherlands, pp. 484-491.

[10] Diez, F. J., Mira, J., Iturralde, E., and Zubillaga, S., 1997, "DIAVAL, a Bayesian Expert System for Echocardiography", Artificial Intelligence in Medicine, 10, pp. 59-73.

[11] Friedman, N., Geiger, D. and Goldszmidt, M., 1997, "Bayesian Network Classifiers", Machine Learning, 29, pp. 131-163.

[12] Friedman, N. and Goldszmidt, M., 1996, "Building Classifiers Using Bayesian Networks", Proc. National Conference on Artificial Intelligence, Menlo Park, CA, USA, AAAI Press, pp. 1277-1284.
[13] Gemela, J., "Financial Analysis Using Bayesian Networks", 2001, Applied Stochastic Models in Business and Industry, 17, pp. 57-67.
[14] Greiner, R., Su, X., Shen, B., and Zhou, W., 2005, "Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers", Machine Learning, 59, pp. 297-332.
[15] Grossman, D., and Domingo, P., 2004, "Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood," Proc. of the 21^{th} International Conference on Machine Learning, Banff, Canada, ACM Press, pp. 361-368.
[16] Hall, P. and Heyde, C. C. 1980, "Martingale Limit Theory and Its Application: Probability and Mathematical Statistics," Academic Press, NY, USA Chap. 6.
[17] Jing, Y., Pavlovic, and Rehb, J. M., 2005, "Efficient Discriminative Learning of Bayesian Network Classifiers via Boosted Augmented Naive Bayes" Proc. of the 22^{th} Intl. Conf. on Machine Learning (ICML 2005), Bonn, Germany, pp. 369-376.
[18] Langley, P., 1993, "Induction of Recursive Bayesian Network Classifiers " Proc. European Conference on Machine Learning, Lecture Notes in Artificial Intelligence, 667, Springer, Berlin, Germany, pp. 153-164.
[19] Ng, A. Y. and Jordon, M. I., 2001, "On discriminative Vs Generative Classifiers: A Comparison of Logistic Regression with Naive Bayes" Advances in Neural Information Processing Systems, 14, MIT Press, Cambridge, MA, USA, pp 605-610.
[20] Pearl, J., 1988, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, California, USA.
[21] Pernkopf, F., and Blimes, J., 2005, "Discriminative Versus Generative Parameter and Structure Learning of Bayesian Network Classifiers" Proc. of the 22^{th} International Conference on Machine Learning (ICML 2005), Bonn, Germany, pp. 657-664.
[22] Porwal, A., Carranza, E. J. M. and Hale, M., 2006,"Bayesian Network Classifiers for Mineral Potential Mapping", Computers and Geosciences, 32, pp. 1-16.
[23] Roos, T., Grunwald, P., Myllymaki, P. and Tirri, H., 2005, "On Discriminative Bayesian Network Classifiers and Logistic Regression" Machine Learning, 59, pp. 267-296.
[24] Rubinstein, Y. D., and Hastie, T., 1997, "Discriminative Vs. Informative Learning" Proc. of the 3^{th} International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp. 49-53.
[25] Santafe, G., Lozano, J. A., and Larranaga, P., 2005, "Discriminative Learning of Bayesian Network Classifiers Via TM Algorithm", Lecture Notes in Artificial Intelligence, 3571, Springer, Germany, pp. 148-160.
[26] Stuart, A and Ord, J. K., 1991, "Kendall's Advanced Theory of Statistics Volume 2", Fifth Edition, Edward Arnorld, London, UK Chap. 17-18.
[27] van der Vaart, A.W., 1998, "Asymptotic Statistics", Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK Chap. 5.
[28] Webb, G. I. and Pazzani, M. J., 1998, "Adjusted Probability Naive Bayes Induction", Proc. 11^{th} Australian Joint Conference on Artificial Intelligence. Lecture Notes in Computer Science, 1502, Springer-Verlag, Heidelberg, Germany, pp. 285-295.
[29] Wijayatunga, P., Mase, S. and Nakamura, M., 2006, "Appraisal of Companies with Bayesian Networks", International Journal of Business Intelligence and Data Mining, 1(3), pp. 329-346.
[30] Yakhnenko, O., Silvescu, A. and Honavar, V., 2005, "Discriminatively Trained Markov Model for Sequence Classification", Proc. 5^{th} IEEE Conference on Data Mining (ICDM'05), IEEE Computer Society, Digital Object Identifier 10.1109/ICDM.2005.52.
[31] Yearling, D. and Hand, D. J., 2003, "A Bayesian Network Datamining Approach for Modelling the Physical Condition of Copper Access Networks", BT Technology Journal, 21(2), pp. 90-100.

Appendix A

Here we give a counter example which violates Assumption 2. First we show a decomposition of the matrix V_{θ^*}.

Lemma 6. *The matrix V_{θ^*} is a weighted average of matrices $V_{\theta^*|x_{[n]}}$, the negative of Fisher information matrix of the conditional density $p_{\theta^*}(x_0 \mid x_{[n]})$.*

Proof.

$$\begin{aligned} V_{\theta^*} &= \sum_{x_0, x_{[n]}} p_{\theta^*}(x_0, x_{[n]}) \dot{\psi}_{\theta^*}(x) \\ &= \sum_{x_{[n]}} p_{\theta^*}(x_{[n]}) \sum_{x_0} p_{\theta^*}(x_0 \mid x_{[n]}) \dot{\psi}_{\theta^*}(x_0, x_{[n]}) \\ &= \sum_{x_{[n]}} p_{\theta^*}(x_{[n]}) V_{\theta^* \mid x_{[n]}}. \end{aligned}$$

□

We can see that matrices $-V_{\theta^*|x_{[n]}}$ for each $x_{[n]}$ are nonnegative definite since we have for $z = (z_1, \ldots, z_k)^T$

$$\begin{aligned} z^T(-V_{\theta^*|x_{[n]}})z &= \sum_{i,j} z_i z_j E_{\theta^*|x_{[n]}} \Big\{ \frac{-\partial^2}{\partial\theta_j \partial\theta_i} \log p_\theta(x_0 \mid x_{[n]}) \Big|_{\theta=\theta^*} \Big\} \\ &= E_{\theta^*|x_{[n]}} \Big\{ \Big\{ \sum_i z_i \frac{\partial}{\partial\theta_i} \log p_\theta(x_0 \mid x_{[n]}) \Big|_{\theta=\theta^*} \Big\}^2 \Big\} \\ &\geq 0, \end{aligned}$$

where $E_{\theta^*|x_{[n]}}$ stands for the expectation w.r.t. $p_{\theta^*}(x_0 \mid x_{[n]})$. It follows that $-V_{\theta^*}$ is also a nonnegative definite matrix:

$$z^T(-V_{\theta^*})z = \sum_{x_{[n]}} p_{\theta^*}(x_{[n]}) z^T(-V_{\theta^*|x_{[n]}})z \geq 0.$$

The matrix $-V_{\theta^*|x_{[n]}}$ is not positive definite if there exits $z \neq 0$ such that

$$\sum_i z_i \frac{\partial}{\partial\theta_i} \log p_\theta(x_0 \mid x_{[n]}) \Big|_{\theta=\theta^*} = 0 \tag{3}$$

for $\boldsymbol{P}_{\theta^*}$-a.s x_0, hence for all x_0, that is, if $\frac{\partial}{\partial\theta_i} \log p_\theta(x_0 \mid x_{[n]})\big|_{\theta=\theta^*}$ is linearly dependent in i. The matrix $-V_{\theta^*}$ is not positive definite if, for some $z \neq 0$, eq. (**??**) holds for all $x_{[n]}$.

Let $n = 1$ and $\mathcal{X}_0 = \mathcal{X}_1 = \{1, 2\}$. Then $\theta = (\theta_{x_0=1}, \theta_{x_1=1|x_0=1}, \theta_{x_1=1|x_0=2}) = (\theta_1, \theta_2, \theta_3)$. Let us find some $z \neq 0$ such that $z^T(-V_{\theta^*|x_1})z = 0$. That is, we have to find $z \neq 0$ so that eq. (**??**) holds for both $x_1 = 1, 2$:

$$\begin{aligned} z_1\theta_2^*\theta_3^* + z_2(1-\theta_1^*)\theta_1^*\theta_3^* + z_3(-\theta_1^*)(1-\theta_1^*)\theta_2^* = 0, \\ z_1(1-\theta_2^*)(1-\theta_3^*) + z_2(1-\theta_1^*)(-\theta_1^*)(1-\theta_3^*) + z_3\theta_1^*(1-\theta_1^*)(1-\theta_2^*) = 0. \end{aligned}$$

If we let $\theta_2^* = \theta_3^*$, then $z = (0, 1, 1)$ is a solution to the above two equations. Therefore V_{θ^*} is not positive definite.

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama 2-12-1 W8-28, Meguro-ku, Tokyo, 152-8552, Japan

E-mail address: spwijay@is.titech.ac.jp

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama 2-12-1 W8-28, Meguro-ku, Tokyo, 152-8552, Japan

E-mail address: mase@is.titech.ac.jp