Conditional Probability Estimation

Marco E. G. V. Cattaneo   M.CATTANEO@HULL.AC.UK
University of Hull
Kingston upon Hull (UK)

Abstract

This paper studies in particular an aspect of the estimation of conditional probability distributions by maximum likelihood that seems to have been overlooked in the literature on Bayesian networks: The information conveyed by the conditioning event should be included in the likelihood function as well.

Keywords: Bayesian networks; maximum likelihood; conditional probabilities.

1. Introduction

The estimation of conditional probability distributions is of central importance for the theory of Bayesian networks. The present paper studies a particular theoretical aspect of this estimation problem that seems to have been overlooked in the literature on Bayesian networks. In fact, the local probability models of Bayesian networks are often estimated from data, and the resulting global model then used to calculate conditional probabilities of future events. As we will see, this way of using Bayesian networks is in agreement with Bayesian estimation, but not with maximum likelihood estimation.

The problem is that when calculating the conditional probability of a future event by maximum likelihood, the information conveyed by the conditioning event should be included in the likelihood function as well. The omission of this information in part explains the unsatisfactory performance of maximum likelihood estimation for Bayesian networks, in particular in the presence of zero counts, which is well known from the literature (see for example Cowell et al., 1999, Jensen and Nielsen, 2007, or Koller and Friedman, 2009).

The effect of including also the information conveyed by the conditioning event in the likelihood function is illustrated by the next example. The following two Sections 2 and 3 will discuss the Bayesian approach and the maximum likelihood method, respectively, as regards the estimation of conditional probability distributions. In Section 4 we will then see that the usual way of applying the maximum likelihood method for estimating conditional probabilities turns out to be correct in many situations. However, in other situations the usual maximum likelihood estimates are incorrect: their repeated sampling performances in such a situation will be compared with the ones of the correct maximum likelihood estimates in Section 5. Finally, the last section concludes the paper and outlines future work.

Example 1 *Let X_1, X_2, Y be three binary random variables, with the two possible values of Y denoted by y and $\neg y$, respectively, and analogously for X_1 and X_2. Assume that X_1 and X_2 are conditionally independent given Y: that is, we have the independence structure of the naive Bayes classifier with class label Y and two features X_1, X_2 (the naive Bayes classifier goes back at least to Warner et al., 1961, and was first studied in a general setting by Duda and Hart, 1973).*

$n(y)$	50	50
$n(x_1, y)$	50	50
$n(x_2, y)$	0	0
$n(x_1, x_2, y)$	0	0
$n(\neg y)$	50	50
$n(x_1, \neg y)$	1	1
$n(x_2, \neg y)$	1	0
$n(x_1, x_2, \neg y)$	0	0

Table 1: Counts from the training datasets of Examples 1 (central column) and 2 (last column).

Assume that we have a complete training dataset consisting of $n = 100$ labeled instances, and we want to classify a new (unlabeled) instance with features x_1 and x_2. The counts from the training dataset are given in the central column of Table 1, with for instance $n(x_1, \neg y)$ denoting the number of instances with class label $\neg y$ and first feature x_1 (regardless of the second feature). In order to classify the new instance, we can estimate

$$p(y \mid x_1, x_2) = \frac{p(y)\, p(x_1 \mid y)\, p(x_2 \mid y)}{p(y)\, p(x_1 \mid y)\, p(x_2 \mid y) + p(\neg y)\, p(x_1 \mid \neg y)\, p(x_2 \mid \neg y)} \tag{1}$$

and check if it is greater than 0.5 (in which case the new instance is classified as y) or less than 0.5 (in which case the new instance is classified as $\neg y$).

If we use a Bayesian approach with uniform independent priors (i.e. we use Laplace's rule of succession to estimate the local probability models of the Bayesian network), we obtain the estimate

$$
\begin{aligned}
\hat{p}(y \mid x_1, x_2) &= \frac{\frac{n(y)+1}{n+2}\frac{n(x_1,y)+1}{n(y)+2}\frac{n(x_2,y)+1}{n(y)+2}}{\frac{n(y)+1}{n+2}\frac{n(x_1,y)+1}{n(y)+2}\frac{n(x_2,y)+1}{n(y)+2} + \frac{n(\neg y)+1}{n+2}\frac{n(x_1,\neg y)+1}{n(\neg y)+2}\frac{n(x_2,\neg y)+1}{n(\neg y)+2}} \\
&= \frac{\frac{51}{102}\frac{51}{52}\frac{1}{52}}{\frac{51}{102}\frac{51}{52}\frac{1}{52} + \frac{51}{102}\frac{2}{52}\frac{2}{52}} = \frac{51}{55} \approx 0.927.
\end{aligned}
\tag{2}
$$

Alternatively, if we apply the maximum likelihood method as described for example by Cowell et al. (1999), Jensen and Nielsen (2007), or Koller and Friedman (2009) (i.e. we estimate the local probability models of the Bayesian network on the basis of the training dataset only), we obtain the estimate

$$
\begin{aligned}
\hat{p}(y \mid x_1, x_2) &= \frac{\frac{n(y)}{n}\frac{n(x_1,y)}{n(y)}\frac{n(x_2,y)}{n(y)}}{\frac{n(y)}{n}\frac{n(x_1,y)}{n(y)}\frac{n(x_2,y)}{n(y)} + \frac{n(\neg y)}{n}\frac{n(x_1,\neg y)}{n(\neg y)}\frac{n(x_2,\neg y)}{n(\neg y)}} \\
&= \frac{\frac{50}{100}\frac{50}{50}\frac{0}{50}}{\frac{50}{100}\frac{50}{50}\frac{0}{50} + \frac{50}{100}\frac{1}{50}\frac{1}{50}} = 0.
\end{aligned}
\tag{3}
$$

Finally, if we apply the maximum likelihood method on the basis of all available data, including the new, unlabeled instance, we obtain the estimate

$$\hat{p}(y \mid x_1, x_2) = \frac{\frac{n(y)+\hat{p}}{n+1} \frac{n(x_1,y)+\hat{p}}{n(y)+\hat{p}} \frac{n(x_2,y)+\hat{p}}{n(y)+\hat{p}}}{\frac{n(y)+\hat{p}}{n+1} \frac{n(x_1,y)+\hat{p}}{n(y)+\hat{p}} \frac{n(x_2,y)+\hat{p}}{n(y)+\hat{p}} + \frac{n(\neg y)+1-\hat{p}}{n+1} \frac{n(x_1,\neg y)+1-\hat{p}}{n(\neg y)+1-\hat{p}} \frac{n(x_2,\neg y)+1-\hat{p}}{n(\neg y)+1-\hat{p}}}$$

$$= \frac{\frac{50+\hat{p}}{101} \frac{50+\hat{p}}{50+\hat{p}} \frac{\hat{p}}{50+\hat{p}}}{\frac{50+\hat{p}}{101} \frac{50+\hat{p}}{50+\hat{p}} \frac{\hat{p}}{50+\hat{p}} + \frac{51-\hat{p}}{101} \frac{2-\hat{p}}{51-\hat{p}} \frac{2-\hat{p}}{51-\hat{p}}} = \frac{\hat{p}\,(51-\hat{p})}{4 + 47\,\hat{p}} = \frac{47}{48} \approx 0.979,$$

(4)

where $\hat{p}(y \mid x_1, x_2) = \hat{p}$ is the only attractive fixed point of the EM algorithm described by (4): that is, the unique solution of $51 - \hat{p} = 4 + 47\,\hat{p}$ (the EM algorithm was first studied in a general setting by Dempster et al., 1977, and is described in the case of Bayesian networks for instance by Cowell et al., 1999, Jensen and Nielsen, 2007, or Koller and Friedman, 2009).

Practically, all three estimates (2), (3), and (4) are obtained by replacing the local probabilities on the right-hand side of (1) by the corresponding estimates. In particular, since the training dataset is complete, a likelihood function based only on this dataset factorizes as the product of five binomial likelihood functions (one for each local probability distribution). Hence, the local probabilities can be independently estimated by maximum likelihood: the estimates are the local relative frequencies, and we obtain expression (3). If we assume uniform independent priors for the local probabilities, and update them using the complete training dataset, then the posteriors have independent beta distributions, and the local probability estimates are their expectations, corresponding to Laplace's rule of succession. That is, we modify the local relative frequencies in (3) by adding 1 to the numerators and 2 to the denominators, obtaining the Bayesian estimate (2).

By contrast, since the new instance is unlabeled, the likelihood function based on all available data does not factorize (it is the sum of two products of five binomial likelihood functions), and maximum likelihood estimates cannot be so easily calculated. However, the EM algorithm theory implies that the maximum likelihood estimates of the local probabilities are the local relative frequencies obtained from the expected counts according to the maximum likelihood model, and thus we obtain the first equality of expression (4). The EM algorithm consists in iteratively using this equality to calculate a better approximation of $\hat{p}$ by using the old approximation in the right-hand side. Anyway, in this particular case we do not need to approximate the solution, since the equality can be solved explicitly: we obtain the maximum likelihood estimate (4).

The main topic of this paper is the question wether the correct way of applying the maximum likelihood method for estimating conditional probabilities is the one employed in (4), or the usual one employed in (3). In Section 3 we will see arguments in favour of (4) and against (3). The goal of the present example is to show that the way in which the maximum likelihood method is applied indeed makes a difference.

In fact, using (4) we would clearly classify the new instance as y, while using (3) we would clearly classify it as $\neg y$. One of the reasons for the huge difference between the conditional probability estimates (4) and (3) is that the training dataset presents zero counts. Looking at the central column of Table 1, we see that exactly half of the 100 labeled instances in the training dataset are classified as y, and in 99 cases we have an agreement between x_1 and y (in the sense that we have either x_1 and y, or $\neg x_1$ and $\neg y$), while x_2 and y agree in 49 cases. That is, X_1 seems to be a very good predictor of Y, while X_2 appears as almost uninformative. Hence, it seems natural to classify the new instance as y, as would be the case when using (2) or (4). By contrast, using (3) means

judging the class label y incompatible with the feature x_2, because no instance with x_2 and y has been observed (the difficulties of Bayesian network classifiers with zero counts are discussed for example by Domingos and Pazzani, 1997, Friedman et al., 1997, or Madden, 2009).

In the context of classifiers, the dispute about the correct way of applying the maximum likelihood method for estimating conditional probabilities, which is the main topic of the present paper, appears as a competition between supervised learning and a very special case of semi-supervised learning, in which a single unlabeled instance is available (see for example Nigam et al., 2000, Chapelle et al., 2006, or Zhu and Goldberg, 2009).

2. Bayesian Approach

Let $p_\theta(x)$ denote the probability function (or density function, in the continuous case) of a random variable X, where "random variable" is not interpreted as implying that X is necessarily real-valued (X could also be categorical, vector-valued, set-valued, and so on). The parameter θ represents what is not known about the probability distribution: it is typically a vector, but more generally it could also simply be an element of a given parameter space.

In the Bayesian approach to statistics (see for example Bernardo and Smith, 1994, Schervish, 1995, or Robert, 2001) the uncertainty about θ is described by a probability (or density) function π. The resulting (marginal, or predictive) probability distribution of X is then given by

$$p(x) = E_\pi[p_\Theta(x)] = \int p_\theta(x)\,\pi(\theta)\,d\theta, \tag{5}$$

where the capital letter Θ indicates that the expectation is taken with respect to θ (the integral is a sum when Θ is discrete).

Let $p_\theta(x, y)$ denote the joint probability (or density) function of two random variables X, Y, and let $p_\theta(y \mid x)$ denote the conditional probability (or density) function of Y given the value of X. In the Bayesian approach, a direct analogy with (5) suggests that the conditional probability distribution of Y given the value of X is

$$p(y \mid x) = E_\pi[p_\Theta(y \mid x)] = E_\pi\left[\frac{p_\Theta(x, y)}{p_\Theta(x)}\right], \tag{6}$$

where the second equality is based on the definition of conditional probability distribution. If this definition is used before applying (5),

$$p(y \mid x) = \frac{p(x, y)}{p(x)} = \frac{E_\pi[p_\Theta(x, y)]}{E_\pi[p_\Theta(x)]} \tag{7}$$

is obtained. However, (6) and (7) cannot be both correct, because expectation and division do not commute.

In fact, (7) is correct, since it only uses the definition of conditional probability distribution and two times the result (5), applied two the random variables (X, Y) and X, respectively. As a consequence, the analogy on which (6) is based must be wrong: the error is that the expectation with respect to θ should not be calculated using the prior π, but using the posterior $\pi \mid x$, where

$$\pi \mid x\,(\theta) = \frac{p_\theta(x)\,\pi(\theta)}{\int p_{\theta'}(x)\,\pi(\theta')\,d\theta'}. \tag{8}$$

This result can be easily proved using (7), (5), and (8):

$$p(y \mid x) = \frac{\int p_\theta(x, y)\,\pi(\theta)\,d\theta}{\int p_\theta(x)\,\pi(\theta)\,d\theta} = \int p_\theta(y \mid x)\,\frac{p_\theta(x)\,\pi(\theta)}{\int p_{\theta'}(x)\,\pi(\theta')\,d\theta'}\,d\theta = E_{\pi \mid x}[p_\Theta(y \mid x)]. \quad (9)$$

Although π is the prior that is updated in (8), it can at the same time be the posterior resulting from some previous update. In particular, in the case of Bayesian networks π could be the posterior resulting from a complete training dataset and a conjugate prior. In this case, the calculation of the probability distributions (5) is relatively simple, and therefore conditional probability distributions are usually calculated using expression (7), as we did in Example 1 for the estimate (2). By contrast, first updating π according to (8) and then calculating the (conditional) probability distribution $E_{\pi \mid x}[p_\Theta(y \mid x)]$ would be computationally much more complicated, but (9) guarantees that we would obtain the same result. Hence, in particular, supervised learning corresponds in the Bayesian approach to the very special case of semi-supervised learning in which the only available unlabeled instance is the one we want to classify.

3. Likelihood Approach

In the likelihood (or classical) approach to statistics (see for instance Schervish, 1995, Pawitan, 2001, or Casella and Berger, 2002) the uncertainty about θ is described by a likelihood function λ. For example, the maximum likelihood estimate of the probability distribution of X is then given by $p_{\hat\theta}(x)$, when $\lambda(\theta)$ has a unique global maximum at $\hat\theta$. As in the Bayesian case of Section 2, a direct analogy suggests that the maximum likelihood estimate of the conditional probability distribution of Y given the value of X is

$$p_{\hat\theta}(y \mid x) = \frac{p_{\hat\theta}(x, y)}{p_{\hat\theta}(x)}. \quad (10)$$

However, as in the Bayesian case of Section 2, the analogy on which (10) is based is wrong. In fact, if the value of X has been observed, basing the estimate on the likelihood function λ contradicts a fundamental principle of statistics, which says that if possible, all data should be used. That is, the (prior) likelihood function λ should first be updated to the (posterior) likelihood function $\lambda \mid x$, where

$$\lambda \mid x\,(\theta) \propto p_\theta(x)\,\lambda(\theta). \quad (11)$$

Assuming that this function has a unique global maximum at $\hat\theta \mid x$, the resulting maximum likelihood estimate of the conditional probability distribution of Y given the value of X is

$$p_{\hat\theta \mid x}(y \mid x) = \frac{p_{\hat\theta \mid x}(x, y)}{p_{\hat\theta \mid x}(x)}. \quad (12)$$

An additional argument in favour of (12) and against (10) is given by the Bayesian approach to statistics. In fact, in the likelihood approach the combination of two different descriptions of uncertainty (probability and likelihood) may give rise to some confusion, while in the Bayesian approach all uncertainties are described probabilistically, and therefore definitive answers are delivered by probability theory, as seen in Section 2. But the two approaches are otherwise very strictly related, the main difference in the case of probability estimation being the use of the expectation with respect to θ based on π in the Bayesian case versus the value at the maximum $\hat\theta$ of λ in the likelihood

case. That is, (10) corresponds to $E_\pi[p_\Theta(y \mid x)]$, while (12) corresponds to $E_{\pi \mid x}[p_\Theta(y \mid x)]$ and is thus the correct maximum likelihood estimate.

The correspondence between the likelihood and Bayesian approaches to statistics, and in particular between (12) and $E_{\pi \mid x}[p_\Theta(y \mid x)]$, answers also the following two possible concerns. First, it could appear that the maximum likelihood estimate (12) uses the observation of the value of X twice: once to update the likelihood, and once to condition the probability distribution. However, if this were true, then $E_{\pi \mid x}[p_\Theta(y \mid x)]$ would also use the same data twice, but (9) shows that this is not the case. Second, the updating of λ according to (11) could appear as unjustified when the value of X has not really been observed (i.e. when the conditioning is only hypothetical). However, since in the Bayesian approach (i.e. in probability theory) there is no difference between conditioning on real or hypothetical observations, the same must be true in the likelihood approach, and therefore (12) is the correct maximum likelihood estimate even when the conditioning is only hypothetical.

In the case of Bayesian networks, if λ is the likelihood function resulting from a complete training dataset, then for probability distributions the calculation of the maximum likelihood estimates $p_{\hat{\theta}}(x)$ is relatively simple, and for conditional probability distributions the maximum likelihood estimates are usually calculated using expression (10), corresponding to the estimate (3) in Example 1. But the correct expression for the maximum likelihood estimates of conditional probability distributions is (12), which is in general more difficult to calculate. However, Example 1 shows that using the correct expression, corresponding to the estimate (4), can make a huge difference, and is relatively simple thanks to the EM algorithm (this remains true also when the estimate cannot be expressed in closed form: see for instance Antonucci et al., 2011, 2012). Hence, in particular, also in the likelihood approach supervised learning should correspond to the very special case of semi-supervised learning in which the only available unlabeled instance is the one we want to classify (though this is not the case when the maximum likelihood method is applied in the usual, incorrect way).

4. When Does This Matter?

Conditional probability estimation is of central importance in statistics and machine learning, in particular in the fields of regression and classification, where one often tries to estimate the conditional probability distribution of the response or class label Y given the value of the vector X of predictors or features. In Section 3 we have seen that the usual way (10) of applying the maximum likelihood method for estimating conditional probability distributions is wrong and should be replaced by (12). Since the maximum likelihood method is by far the most popular estimation technique, one could have the impression that the arguments of Section 3 would have a huge impact on statistics and machine learning, at least from the theoretical standpoint. However, the impact of these new arguments is limited, because (10) and (12) turn out to be identical in many situations.

In particular, assume that the joint probability (or density) function of X and Y can be written as

$$p_\theta(x, y) = p_{\theta_1}(x) \, p_{\theta_2}(y \mid x), \tag{13}$$

where $\theta = (\theta_1, \theta_2)$, and the parameter space for θ is the Cartesian product of the ones for θ_1 and θ_2. Furthermore, assume that we have a complete training dataset, in the sense that we have observed some instances (x, y) from the distribution (13). In this case, the resulting likelihood function factorizes as $\lambda(\theta) = \lambda_1(\theta_1) \, \lambda_2(\theta_2)$, and the observation of one or more additional unlabeled instances x would affect only λ_1. As a consequence, the maximum likelihood estimate of θ_2 is not affected

by such additional observations, and since the conditional probability distribution of Y given the value of X depends only on θ_2, (10) and (12) are identical in this case.

Assumption (13) holds in a large part of the models used in statistics, in particular in the fields of regression and classification, including linear regression models, logistic regression models, analysis of variance models, general linear models, additive models, generalized linear models, generalized additive models, and so on (see for example McCullagh and Nelder, 1989, and Hastie and Tibshirani, 1990). Hence, in all these cases the arguments of Section 3 have no impact on the estimation of the conditional probability distribution of the response given the predictors, since the usual way in which the maximum likelihood method is applied turns out to be correct.

This is not always true in the case of Bayesian networks, as shown by Example 1. However, also in this case the following, more specific version of assumption (13) is often satisfied:

$$p_\theta(w, x, y, z) = p_{\theta_1}(w, x) \, p_{\theta_2}(y \mid x) \, p_{\theta_1}(z \mid w, x, y), \qquad (14)$$

where $p_\theta(w, x, y, z)$ denotes the joint probability (or density) function of four random variables W, X, Y, Z. As above, if we have a complete training dataset, in the sense that we have observed some instances (w, x, y, z) from the distribution (14), then the observation of one or more additional unlabeled instances x would not affect the maximum likelihood estimate of θ_2. Hence, the maximum likelihood estimate of the conditional probability distribution of Y given the value of X is not affected by such additional observations, and therefore (10) and (12) are identical in this case.

Assumption (14) corresponds to the assumptions that W and Y are conditionally independent given X, and that the conditional probability distribution of Y given the value of X is parameterized separately from the rest of the model. Given a Bayesian network, let Y denote any variable, and let X, Z, W denote the parents of Y, the descendants of Y, and all remaining variables, respectively. Then the local Markov property implies that assumption (14) is satisfied, when the the local probability model of Y given its parents is estimated separately from the rest of the model. That is, the usual way of applying the maximum likelihood method for estimating the local probability models of a Bayesian network is correct (notice also that this is an example of a situation in which the conditioning is only hypothetical, since we are not considering any new data).

The last result can be misleading, because the way in which Bayesian networks are usually employed consists of two separate steps: first the model is estimated, and then the estimated model is used for inference and decision making (see for example Cowell et al., 1999, Jensen and Nielsen, 2007, or Koller and Friedman, 2009). As seen in Section 2, these two steps are in agreement with the Bayesian approach, since the conditional probability distributions (7) obtained from the estimated model are correct. By contrast, as seen in Section 3, the two steps are not in agreement with the likelihood approach, because the conditional probability distributions (10) obtained from the estimated model are wrong according to the maximum likelihood method.

A manifest consequence, illustrated by the next example, of the usual, wrong way of employing Bayesian networks in the likelihood approach is that the estimate (10) can be undefined, when the denominator of the fraction is 0. Now, if we have just observed the value of X, to estimate the probability of this value as 0 is nonsense (this is true also when the conditioning is hypothetical, and thus the observation of the value of X is only imagined). In fact, the correct maximum likelihood estimate (12) of the conditional probability distribution of Y given the value of X is always well defined, when the updated likelihood function $\lambda \mid x$ has a unique global maximum at $\hat{\theta} \mid x$, since (11) implies

$$p_{\hat{\theta} \mid x}(x) \, \lambda(\hat{\theta} \mid x) \propto \lambda \mid x \, (\hat{\theta} \mid x) > 0, \qquad (15)$$

and therefore the denominator of the fraction in (12) is certainly positive. The same is true for the Bayesian conditional probability distribution (7), when the prior is not too extreme.

Example 2 *Assume that for the Bayesian network of Example 1 the counts from the training dataset are given in the last column of Table 1. That is, the only difference from Example 1 is that instead of an instance with class label $\neg y$ and features $\neg x_1, x_2$ we have observed an instance with class label $\neg y$ and features $\neg x_1, \neg x_2$. Hence, the estimates (2) and (4) of the (conditional) probability of class label y for a new (unlabeled) instance with features x_1, x_2 are modified as follows. The Bayesian estimate, corresponding to (7), becomes*

$$\hat{p}(y \mid x_1, x_2) = \frac{\frac{51}{102} \frac{51}{52} \frac{1}{52}}{\frac{51}{102} \frac{51}{52} \frac{1}{52} + \frac{51}{102} \frac{2}{52} \frac{1}{52}} = \frac{51}{53} \approx 0.962, \tag{16}$$

while the maximum likelihood estimate, corresponding to (12), becomes

$$\hat{p}(y \mid x_1, x_2) = \frac{\frac{50+\hat{p}}{101} \frac{50+\hat{p}}{50+\hat{p}} \frac{\hat{p}}{50+\hat{p}}}{\frac{50+\hat{p}}{101} \frac{50+\hat{p}}{50+\hat{p}} \frac{\hat{p}}{50+\hat{p}} + \frac{51-\hat{p}}{101} \frac{2-\hat{p}}{51-\hat{p}} \frac{1-\hat{p}}{51-\hat{p}}} = \frac{\hat{p}\,(51-\hat{p})}{2 + 48\,\hat{p}} = 1, \tag{17}$$

where $\hat{p}(y \mid x_1, x_2) = \hat{p}$ is the only attractive fixed point of the EM algorithm described by (17): that is, the unique solution of $51 - \hat{p} = 2 + 48\,\hat{p}$. In both cases the conditional probability estimate becomes even larger than in Example 1, and thus we would still classify the new instance as y. By contrast, if we try to modify the estimate (3), corresponding to (10), we obtain

$$\hat{p}(y \mid x_1, x_2) = \frac{\frac{50}{100} \frac{50}{50} \frac{0}{50}}{\frac{50}{100} \frac{50}{50} \frac{0}{50} + \frac{50}{100} \frac{1}{50} \frac{0}{50}} = \frac{0}{0}. \tag{18}$$

5. Performance Comparison

In Section 3 we have seen that the usual way (10) of applying the maximum likelihood method for estimating conditional probability distributions is wrong and should be replaced by (12). In Section 4 we have seen that the two estimates (10) and (12) turn out to be identical when some assumptions are satisfied, but when these are not satisfied the difference between the estimates can be important, as illustrated by Examples 1 and 2. In this section we compare the repeated sampling performances of the two estimates (10) and (12) and of the Bayesian estimate (7) with uniform independent priors, in the case of the simple Bayesian network of Examples 1 and 2.

Assume that for the Bayesian network of Examples 1 and 2 we have a complete training dataset consisting of $n = 100$ labeled instances, and we want to estimate the (conditional) probability $p(y \mid x_1, x_2)$ of class label y for a new (unlabeled) instance with features x_1 and x_2. We consider the training dataset to be a random sample from the probability distribution described by the Bayesian network, and therefore each estimate $\hat{p}(y \mid x_1, x_2)$ based on the training dataset will be random as well. As usual in statistics, we can evaluate the repeated sampling performance of an estimate by its mean squared error

$$E\left[\left(\hat{p}(y \mid x_1, x_2) - p(y \mid x_1, x_2)\right)^2\right], \tag{19}$$

where the expectation is taken with respect to $\hat{p}(y \mid x_1, x_2)$, while y, x_1, x_2, and $p(y \mid x_1, x_2)$ are constant. The relative efficiency of two estimates is then the inverse ratio of their mean squared errors, so that to a smaller mean squared error corresponds a higher (relative) efficiency.

The relative efficiency of the estimates depends on the probability distribution of which the training dataset is a random sample. For instance, both training datasets summarized in Table 1 are perfectly in agreement with the probability distribution determined by the Bayesian network of Examples 1 and 2, by $p(y) = 50\%$, and by the conditional probabilities

$$p(x_1 \,|\, y) = p(\neg x_1 \,|\, \neg y) = 99\%,$$
$$p(\neg x_2 \,|\, y) = p(\neg x_2 \,|\, \neg y) = 99\%. \tag{20}$$

However, in this case the maximum likelihood estimate (10) based on the incomplete likelihood function has a high probability of being undefined: it exists with 49% probability only (all expectations and probabilities in this section were obtained by simulations with ten million repetitions, and 100 iterations of the EM algorithm for each repetition, when needed). Hence, the mean squared error of this estimate is also undefined: to avoid this problem, we can calculate mean squared errors and relative efficiencies conditional on the existence of the incomplete maximum likelihood estimate (10). That is, we can replace the expectation in (19) by the conditional expectation given that the incomplete maximum likelihood estimate (i.e. the one based on the training dataset only) of $p(x_1, x_2)$ is positive. The resulting conditional relative efficiency of the complete and incomplete maximum likelihood estimates (12) and (10) (i.e. the ones based on the complete and incomplete likelihood functions, respectively) is 641. That is, even when ignoring the fact that the estimate does not always exist, the usual way of applying the maximum likelihood method is much less efficient that the correct one in this case. The correct maximum likelihood estimate is also 5.95 times more efficient than the Bayesian estimate (7) with uniform independent priors, corresponding to Laplace's rule of succession (in this case, the unconditional relative efficiency has been used, since both estimates always exist).

Figure 1 shows what happens in this example when the conditional probabilities (20) are kept constant, while $p(y)$ can vary over the interval $[0, 1]$. In general, the incomplete maximum likelihood estimate (10) has a high probability of being undefined, and this is its most important deficiency. Interestingly, in this case the (conditional or unconditional) mean squared errors of the complete maximum likelihood and Bayesian estimates (12) and (7) are rather similar, while the conditional mean squared error of the incomplete maximum likelihood estimate (10) is completely different: this could also be interpreted as an indication that something is wrong with this estimate.

However, for other probability distributions the complete and incomplete maximum likelihood estimates (12) and (10) can be very similar (when the latter is well defined), as shown by Figure 2, which is based on the conditional probabilities

$$p(x_1 \,|\, y) = p(\neg x_1 \,|\, \neg y) = 99\%,$$
$$p(\neg x_2 \,|\, y) = \quad p(x_2 \,|\, \neg y) = 90\%. \tag{21}$$

That is, in this case also the second feature X_2 is informative (though pointing in the opposite direction than X_1) and much less likely to produce zero counts than in the case of Figure 1. For example, when $p(y) = 50\%$, the complete maximum likelihood estimate is 1.02 times more efficient than the incomplete one (conditionally on the latter being well defined, which happens with almost 100% probability) and 1.36 times more efficient than the Bayesian estimate (unconditionally).
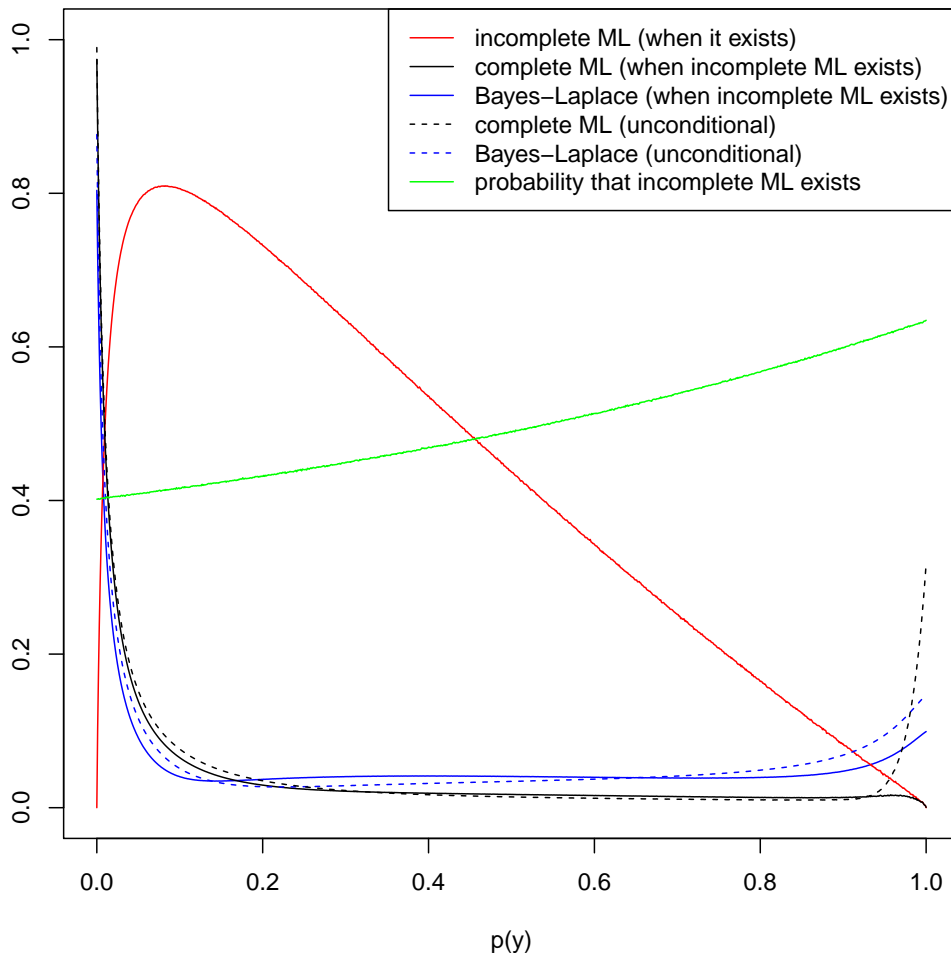
Figure 1: Square root of the (conditional or unconditional) mean squared errors as functions of $p(y)$ in the case with conditional probabilities (20).

6. Conclusion

In this paper we have seen that the maximum likelihood method is often used incorrectly when estimating conditional probabilities, because the information conveyed by the conditioning event should be included in the likelihood function as well. A particularly disturbing consequence of the omission of this information is that the usual, wrong maximum likelihood estimates do not always exist, while the correct ones do. But using an incorrect likelihood function has consequences also on other kinds of likelihood-based inference, such as likelihood-ratio tests or likelihood-based confidence regions, which will be the object of future work (see also Cattaneo, 2010). Furthermore, the effect of correcting the likelihood function on the performance of Bayesian network classifiers will be studied empirically, also in comparison with alternative methods for coping with zero counts.
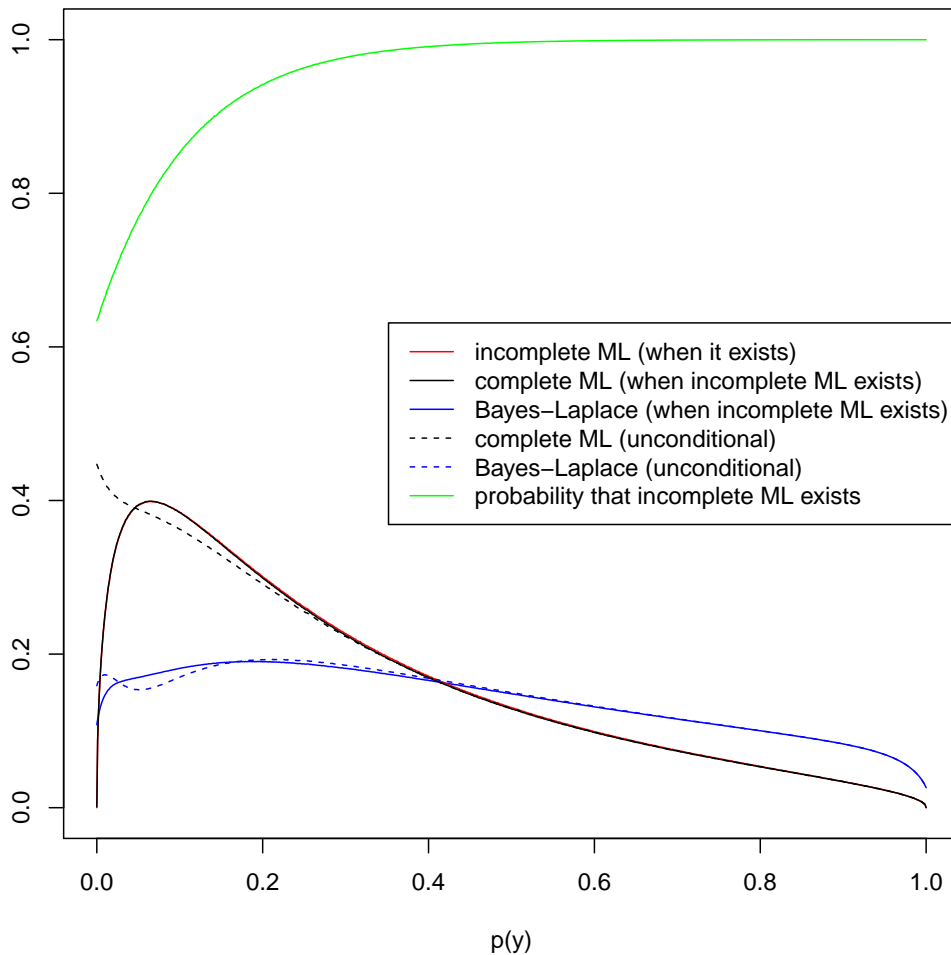
Figure 2: Square root of the (conditional or unconditional) mean squared errors as functions of $p(y)$ in the case with conditional probabilities (21).

References

A. Antonucci, M. Cattaneo, and G. Corani. Likelihood-based naive credal classifier. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 21–30. SIPTA, 2011.

A. Antonucci, M. Cattaneo, and G. Corani. Likelihood-based robust classification with Bayesian networks. In S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, and R. R. Yager, editors, *Advances in Computational Intelligence*, volume 3, pages 491–500. Springer, 2012.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1994.

G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, second edition, 2002.

M. Cattaneo. Likelihood-based inference for probabilistic graphical models: Some preliminary results. In P. Myllymäki, T. Roos, and T. Jaakkola, editors, *PGM 2010, Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, pages 57–64. HIIT Publications, 2010.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

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

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39: 1–38, 1977.

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

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29: 131–163, 1997.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer, second edition, 2007.

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

M. G. Madden. On the classification performance of TAN and general Bayesian networks. *Knowledge-Based Systems*, 22:489–495, 2009.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, second edition, 1989.

K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford, 2001.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, second edition, 2001.

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

H. R. Warner, A. F. Toronto, L. G. Veasey, and R. Stephenson. A mathematical approach to medical diagnosis: Application to congenital heart disease. *JAMA: The Journal of the American Medical Association*, 177:177–183, 1961.

X. Zhu and A. B. Goldberg. *Introduction to Semi-Supervised Learning*. Morgan & Claypool, 2009.