7th International Symposium on Imprecise Probability: Theories and Applications, Innsbruck, Austria, 2011

Likelihood-Based Naive Credal Classifier

Alessandro Antonucci
IDSIA, Lugano
alessandro@idsia.ch

Marco E. G. V. Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

Giorgio Corani
IDSIA, Lugano
giorgio@idsia.ch

Abstract

The naive credal classifier extends the classical naive Bayes classifier to imprecise probabilities, substituting the imprecise Dirichlet model for the uniform prior. As an alternative to the naive credal classifier, we present a likelihood-based approach, which extends in a novel way the naive Bayes towards imprecise probabilities, by considering any possible quantification (each one defining a naive Bayes classifier) apart from those assigning to the available data a probability below a given threshold level. Besides the available supervised data, in the likelihood evaluation we also consider the instance to be classified, for which the value of the class variable is assumed missing-at-random. We obtain a closed formula to compute the dominance according to the maximality criterion for any threshold level. As there are currently no well-established metrics for comparing credal classifiers which have considerably different determinacy, we compare the two classifiers when they have comparable determinacy, finding that in those cases they generate almost equivalent classifications.

Keywords. Classification, naive credal classifier, naive Bayes classifier, likelihood-based learning.

1 Introduction

Classification, understood as the problem of assigning *class* labels to instances described by a set of *features*, is one of the major problems of AI, with lots of important applications, including pattern recognition, prediction, and diagnosis. Bayesian approaches to classification are particularly popular and effective. In particular, the *naive Bayes classifier* (NBC; e.g., see [11, Chap. 17]), assumes the conditional independence of the feature variables given the class; because of this unrealistic assumption, NBC requires the estimation of only a few parameters from the data. Yet, this assumption typically biases the probability computed by NBC which, regarding all the features as independent pieces of evidence, tends to assign a excessively high probability to the most probable class. The problem is emphasised in the presence of many features, among which could easily exist correlations [9]. However, NBC generally achieves a good accuracy under 0-1 loss; this means that, despite the biased probabilities, it produces good ranks among the competing classes [7]. The parameters are typically learned in a Bayesian way with uniform prior. Maximum-likelihood quantification has the advantage of being unbiased and independent from the prior specification, but generally leads to inferior classification performance, especially on data sets where the contingency tables, which contain the counts of the joint occurrences of specific values of the features and the class, are characterised by several zeros [8, 12] (see also Example 3).

The *naive credal classifier* (NCC, [18]), a generalisation of the NBC based on the theory of *imprecise probability* [15], attempts to make classification independent of the choice of the prior in a different way. NCC learns from data through the *imprecise Dirichlet model* (IDM, [16]); this corresponds to adopting a set of priors, which model a condition of near-ignorance about the model parameters. A NCC is equivalent to a collection of NBCs; while NBC returns the single class with highest probability according to the posterior probability mass function, NCC can in some cases suspend the judgment, by returning a set of classes rather than a single one. This provides a cautious and robust classification. A similar approach could be obtained by applying a *rejection option* to NBC, namely by returning more classes when the posterior probability estimated for the most probable class does not exceed a certain threshold. However, the rejection option requires accurate probability estimates to be effective, which is hardly the case for the NBC.

Of course, IDM is not the only technique to learn sets of distributions from data. Among others, *likelihood-based* approaches to the learning of imprecise-proba-

bilistic models from data [3, 14] can be regarded as an alternative to the IDM. Loosely speaking, the idea is to consider, instead of the single maximum-likelihood estimator, all the models whose likelihood is above a certain threshold level.

In this paper we investigate how likelihood-based techniques apply to NCC quantification. To do that, we keep the same independence assumptions of the NBC (and of the NCC), but we change the way the model is quantified. We call the resulting model *likelihood-based naive credal classifier* (LNCC). This model is associated with a classification algorithm which computes the set of unrejected classes according to the *maximality* criterion [15] (exactly as the NCC does) for any threshold level.

A notable feature of our approach is that, in the likelihood evaluation, we do not only consider the available (learning) data set, but also the instance to be classified, whose value of the class variable is assumed to be missing-at-random. This is important to obtain more accurate classification performances when coping with zero counts in the data set.

The paper is organised as follows. We first review some background material about the naive Bayes (Section 2.1) and credal (Section 2.2) classifiers and the likelihood-based approaches to the learning of imprecise-probabilistic models from data (Section 3). Then, in Section 4, we introduce the LNCC and obtain an analytic inference formula to compute the set of candidate optimal classes. Numerical tests are in Section 5. Conclusions and outlooks are finally in Section 6, while the proofs are in the appendix.

2 Naive Classifiers

In this section we review the necessary background information about classifiers developed under the naive assumption (i.e., independence between features given the class). First let us introduce the general problem of classification together with the necessary notation.

We use uppercase for the variables, lowercase for the states, calligraphic for the possibility spaces, and boldface for sets of variables. Let C denote the *class* variable, with generic value c, taking values in a finite set $\mathcal{C}$. Similarly, we have m features, $\mathbf{F} := (F_1, \ldots, F_m)$, each one taking values in the finite set $\mathcal{F}_j$, $j = 1, \ldots, m$.[1] Assume that the available data are d joint observations of these variables, say $\mathcal{D} := \{(c^{(i)}, f_1^{(i)}, \ldots, f_m^{(i)})\}_{i=1}^d$, with $c^{(i)} \in \mathcal{C}$ and $f_j^{(i)} \in \mathcal{F}_j$, for each $i = 1, \ldots, d$ and $j = 1, \ldots, m$. Information associated with the data set $\mathcal{D}$ is described by a *count function* n returning the number of elements of the data set $\mathcal{D}$ satisfying a condition to be specified in its argument. E.g., $n(C = c)$ is the number of instances where the class has value $c \in \mathcal{C}$, while $n(C = c, F_j = f_j)$ is the number of instances where C has value c and the j-th feature has value f_j. For sake of notation, we denote these counts as $n(c)$ and $n(c, f_j)$, and similarly for the others, with $n(\cdot) = d$.

Given an instance of the features $\tilde{\mathbf{f}} = (\tilde{f}_1, \ldots, \tilde{f}_m)$, classification is the problem of assigning it a single class label or, as in the case of Section 2.2, a set of them, all of which are candidates to be the correct category. A classifier always returning a single class is called *precise*, and *credal* otherwise.

2.1 Naive Bayes Classifier

A probabilistic approach to classification consists of learning from the data $\mathcal{D}$ a joint probability mass function for the whole set of variables $(C, \mathbf{F})$. Let the unknown chances of this distribution be denoted by $\theta_{c,\mathbf{f}}$ for each $(c, \mathbf{f}) \in \mathcal{C} \times \mathcal{F}_1 \times \ldots \times \mathcal{F}_m$. Once we learn these chances, we assign to the instance $\tilde{\mathbf{f}}$ the class label maximising the posterior (which is proportional to the joint) probability, i.e.,

$$\arg\max_{c \in \mathcal{C}} \theta_{c,\tilde{\mathbf{f}}}.$$

As the number of parameters specifying the joint distribution grows exponentially with the number of features, such a probabilistic approach is generally too demanding, unless we make some assumption about the independence relations between the variables. A notable example is the so-called *naive* assumption, which says that, given the class variable, the features are conditionally independent from each other.[2] This induces in the joint the following factorisation:

$$\theta_{c,\mathbf{f}} := \theta_c \cdot \prod_{j=1}^{m} \theta_{f_j|c}, \tag{1}$$

where θ_c is the (unconditional) chance for $C = c$, and similarly for the conditional ones. Equation (1) makes it possible to assess the joint distribution, and hence perform classification, by means of a number of parameters which is linear in the number of features and classes. Let θ denote the whole set of chances to be quantified on the right-hand side of (1) and Θ the corresponding set of possible assignments. The parameter θ is quantified in a Bayesian way; given a Dirichlet prior over Θ, we obtain the following poste-

[1] We focus on classification of discrete features. A discussion on the extension to continuous variables is in the conclusions.

[2] We say that A and B are conditionally independent given C if $P(a, b|c) = P(a|c) \cdot P(b|c)$, for each a, b, and c.

rior estimates:

$$\theta_c = \frac{n(c) + s\,t(c)}{n(\cdot) + s}, \quad (2)$$

$$\theta_{f_j|c} = \frac{n(c, f_j) + s\,t(c, f_j)}{n(c) + s\,t(c)}, \quad (3)$$

where Walley's parametrisation of the Dirichlet distribution is employed. In particular, s can be thought of as a number of *hidden instances*, in the usual interpretation of conjugate Bayesian priors as additional samples. The parameters $t(\cdot)$ can be interpreted as the proportion of hidden instances of a given type; for instance, $t(c)$ is the expected proportion of hidden instances for which $C = c$.

In particular, non-informative specifications can be obtained by Perks' prior, which means $t(c) := |\mathcal{C}|^{-1}$ and $t(c, f_j) := |\mathcal{F}_j|^{-1}|\mathcal{C}|^{-1}$ for each $c \in \mathcal{C}$, $f_j \in \mathcal{F}_j$, $j = 1, \ldots, m$, and $s = 1$. In the language of Bayesian networks, this is also known as BDe [11, Chap. 17].

2.2 Naive Credal Classifier

The classification performances of the NBC can be quite sensitive to the choice of the prior. In a situation where different priors return different class labels, a conservative approach consists of taking multiple priors as a model of a condition of prior (near) ignorance about the model parameters, and hence learning a posterior independently for each prior. This can be done by means of the *imprecise Dirichlet model* (IDM, [16]), for which the "precise" specification of the NBC Dirichlet prior is relaxed, and its parameters are free to vary in the following set, with minimal constraints:

$$\mathcal{T} := \left\{ \mathbf{t} \;\middle|\; \begin{array}{l} \sum_{c \in \mathcal{C}} t(c) = 1 \\ \sum_{f_j \in \mathcal{F}_j} t(c, f_j) = t(c), \forall c \in \mathcal{C}, \forall j \\ t(c, f_j) > 0, \forall (c, f_j) \in \mathcal{C} \times \mathcal{F}_j, \forall j \end{array} \right\}. \quad (4)$$

Each $\mathbf{t} \in \mathcal{T}$ corresponds to a different Dirichlet prior and hence a different NBC quantification. The collection of all these NBCs is called *naive credal classifier* (NCC, [17]), and provides a collection of posterior distributions for the class variable given the feature of the instance to be classified. In order to decide which class labels to assign to the instance, the *maximality* criterion [15] is adopted: a class is rejected if there is another class that is more probable according to every distribution. Thus, in order to perform classification with the NCC, for each $c', c'' \in \mathcal{C}$, we have to test whether or not c' *dominates* c'', i.e.,[3]

$$\inf_{\mathbf{t} \in \mathcal{T}} \frac{P_{\mathbf{t}}(c', \tilde{\mathbf{f}})}{P_{\mathbf{t}}(c'', \tilde{\mathbf{f}})} > 1, \quad (5)$$

where $P_{\mathbf{t}}$ is the NBC quantification associated to $\mathbf{t}$. From (1), (2) and (3), we can rewrite the objective function of our optimisation problem in (5) as[4]

$$\left[\frac{n(c') + s\,t(c')}{n(c'') + s\,t(c'')}\right]^{1-m} \prod_{j=1}^{m} \frac{n(c', \tilde{f}_j)}{n(c'', \tilde{f}_j) + s\,t(c'', \tilde{f}_j)},$$

and hence check dominance by solving the corresponding optimisation with the constraints in (4).

Counterintuitive behaviors of NCC take place in presence of zero counts; in particular (a) an attribute F_j such that $n(c', \tilde{f}_j) = 0$ prevents c' from dominating any other class (see Example 3); (b) a class c' such that $n(c') = 0$ is identified as non-dominated for most instances. These behaviors were first observed in [17]; a solution to these problems, which make the NCC unnecessarily imprecise, has been studied in [4], proposing an ϵ-contamination of the IDM prior with the uniform prior of the NBC: this corresponds to a slight modification of the set $\mathcal{T}$, obtained by rewriting the constraints in (4) in the form $\epsilon\,|\mathcal{C}|^{-1} \leq t(c) \leq (1-\epsilon) + \epsilon\,|\mathcal{C}|^{-1}$, and similarly for $t(c, f_j)$. Such a NCC extension is denoted as NCC_ϵ.[5]

3 Likelihood-Based Learning of Imprecise-Probabilistic Models

Coping with multiple priors as in the IDM is not the only possible approach to learn imprecise-probabilistic models from data. In a likelihood-based approach, we can simply start by considering a collection of candidate models, and then only keep those assigning to the available data a probability beyond a certain threshold. We introduce these ideas by means of an example.

Example 1. *Consider a Boolean variable X, for which N observations are available, and n of them report the state true. If $\theta \in [0, 1]$ is the chance that X is true, the likelihood induced by the observed data is $lik(\theta) := \theta^n \cdot (1-\theta)^{N-n}$ and its maximum is attained at $\hat{\theta} = \frac{n}{N}$. For each $\alpha \in [0, 1]$, we can (numerically) compute the values of θ such that $lik(\theta) \geq \alpha\, lik(\hat{\theta})$. Figure 1 depicts the behaviour of these intervals (which can be also interpreted as confidence intervals for θ; e.g., see [10]) for increasing sample size.*

The approach considered in the above example can be easily extended to the general case, and can be interpreted as a way of updating imprecise probabilities [1, 13], in the following sense. Consider a *credal*

[3] Note that the ratio between conditional probabilities can be equivalently described as a ratio between joint probabilities.

[4] Note that a partial optimisation has been already performed in the numerators of the terms in the product.

[5] Note that NCC_0 is the NCC, while NCC_1 is the NBC.

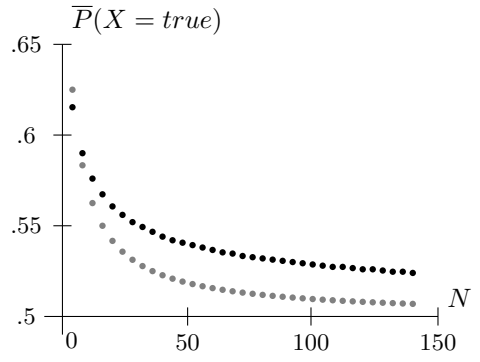

Figure 1: Comparison between probability intervals obtained by likelihood-based learning ($\alpha = .85$, black points) and IDM ($s = 2$, grey points) for Example 1. The plot shows the upper bounds of the interval probability that the variable is true as a function of the sample size N, when $\frac{n}{N} = \frac{1}{2}$. The plot for the lower bounds would be symmetric to this one.

set $\mathbf{P}$, i.e., a collection of probability distributions all over the same variable. Assume the elements of $\mathbf{P}$ are indexed by a parameter θ taking values in a set Θ, i.e., $\mathbf{P} := \{P_\theta\}_{\theta\in\Theta}$. Given the available data $\mathcal{D}$, let us consider the corresponding normalised likelihood:

$$lik(\theta) := \frac{P_\theta(\mathcal{D})}{\sup_{\theta'\in\Theta} P_{\theta'}(\mathcal{D})}. \tag{6}$$

The likelihood-based approach to learning consists of removing from $\mathbf{P}$ the distributions whose normalised likelihood is below some threshold. Thus, given $\alpha \in [0, 1]$, we consider the following (smaller) credal set:

$$\mathbf{P}_\alpha := \{P_\theta\}_{\theta\in\Theta\,:\,lik(\theta)\geq\alpha}. \tag{7}$$

Clearly, $\mathbf{P}_{\alpha=1}$ is typically a "precise" credal set including only the maximum-likelihood distribution, while $\mathbf{P}_{\alpha=0} = \mathbf{P}$. In principle, the original credal set $\mathbf{P}$ can be obtained by means of some other imprecise-probabilistic learning technique, which is indeed refined by the likelihood-based approach. Likelihood-based learning is said to be *pure*, if the credal set $\mathbf{P}$ includes all the possible distributions that can be specified over the variable under consideration (or, as in the next section, at least all those satisfying the structural judgements about symmetry and independence characterising the model under consideration).

4 Likelihood-Based Naive Credal Classifier

Let us consider a pure likelihood-based learning of the model probabilities of the naive classifier. Thus, let $\mathbf{P}$ denote the credal set associated to a NCC with vacuous quantification of the model probabilities (i.e., each chance is only required to belong to the $[0, 1]$ interval). Let the parameter θ with values in Θ denote a parametrisation of this credal set, i.e., $\mathbf{P} := \{P_\theta\}_{\theta\in\Theta}$, where θ is a NBC quantification. Given the available data $\mathcal{D}$, let us consider the normalised likelihood as in (6), and hence the credal set $\mathbf{P}_\alpha \subseteq \mathbf{P}$ as in (7).

We call *likelihood-based naive credal classifier* (LNCC, called *naive hierarchical classifier* in [3]) the collection of NBCs in the credal set $\mathbf{P}_\alpha$. This only provides an implicit specification of the model probabilities.[6] Yet, we can already describe how LNCC-based classification is intended. The same dominance criterion (i.e., maximality) as for the NCC is considered, and we say that c' dominates c'' iff

$$\inf_{\theta\in\Theta\,:\,lik(\theta)\geq\alpha} \frac{P_\theta(c', \tilde{\mathbf{f}})}{P_\theta(c'', \mathbf{f})} > 1. \tag{8}$$

In order to perform classification with the LNCC, we should discuss (8) for each pair of classes $c', c'' \in \mathcal{C}$. This task will be considered in Section 4.1. First, let us note that, when evaluating the likelihood *lik*, we do not only consider the data set $\mathcal{D}$, but also the instance under consideration $\tilde{f}$. The value of the class variable for this instance is unavailable (i.e., missing), no matter what its actual value is. Thus, the probability we should take into account for the overall likelihood evaluation is the product of $P_\theta(\mathcal{D})$ and

$$P_\theta(\tilde{\mathbf{f}}) := \sum_{c\in\mathcal{C}} \left[\theta_c \prod_{i=1}^{m} \theta_{\tilde{f}_i|c}\right]. \tag{9}$$

Note that we perform classification by means of the dominance test in (8) for each $c', c'' \in \mathcal{C}$. Thus, as we cope with the likelihood separately for each pair of classes, a simplification assumption consists of assuming that, when checking whether c' dominates c'', the instance under consideration can only be c' or c''. This basically means to restrict the sum in (9) only to c' and c''. In order to see how this kind of classification works in practice consider the following example.

Example 2. *Consider a LNCC with a Boolean class C and a single Boolean feature F. In this setup, a NBC specification is provided by the three-dimensional parameter $\theta := (\theta_c, \theta_{f|c}, \theta_{f|\neg c})$, taking values in $\Theta := [0, 1]^3$. Apart from (c, f) which appears five times, the other three possible combinations for the class/feature values appear only once in the data set. To decide whether or not $C = c$ dominates $C = \neg c$, when the instance to be classified is $F = f$, we first compute the likelihood of the available (supervised) data:*

$$lik(\theta) = \theta_c^6 \cdot (1-\theta_c)^2 \cdot \theta_{f|c}^5 \cdot (1-\theta_{f|c}) \cdot \theta_{f|\neg c} \cdot (1-\theta_{f|\neg c}).$$

[6] Note that, if regarded as a credal net [6], the LNCC (as the NCC) has non-separately specified credal sets.

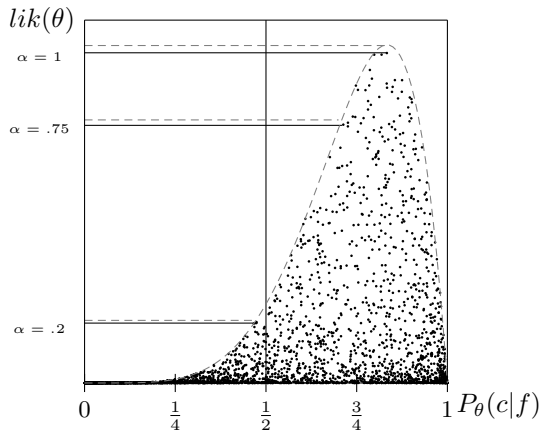

Figure 2: LNCC-based classification. The dominance test in Example 2 is solved by generating a random sample of 3000 NBC quantifications θ and depicting for each θ the posterior and the likelihood as the point $(P_\theta(c|f), lik(\theta))$. Note that in the Boolean case $P_\theta(c|f) > \frac{1}{2}$ is an equivalent dominance condition. The upper envelope of the points in the limit of an infinite sample size (see Section 4.1) is depicted in grey. Horizontal lines describe the cuts for different α-values. Black lines are based on the random sample, while those referred to the upper envelope are grey.

As we also want to consider the instance to be classified, we multiply this likelihood by the chance that $F = f$, which according to (9) is

$$\theta_c\,\theta_{f|c} + (1-\theta_c)\,\theta_{f|\neg c}.$$

For each $\theta \in \Theta$, c dominates $\neg c$ if

$$\frac{\theta_c\,\theta_{f|c}}{(1-\theta_c)\,\theta_{f|\neg c}} > 1.$$

To perform classification with the LNCC, we just have to check whether or not such a dominance relation is satisfied for each θ whose likelihood is not below the maximum likelihood multiplied by α. Figure 2 reports a Monte Carlo solution of this problem. Note that we have dominance for high threshold levels (e.g., $\alpha = .75$), and no dominance for low levels (e.g., $\alpha = .2$).

4.1 Statistical Inference with LNCC

In the previous section we defined the LNCC corresponding to a given α level, and described how we intend to perform inference based on this model. Yet, the sampling-based method considered in Example 2 is not necessary. In this section, we provide a classification algorithm for the LNCC based on a parametric formula for the upper envelope of the likelihood.

Let us therefore, for a generic classification problem, consider the dominance test between c' and c'' for an instance $\tilde{\mathbf{f}}$ to be classified by means of the LNCC for a given threshold α on the basis of the data $\mathcal{D}$. The idea is to parametrise the upper envelope of the likelihood (also called *profile likelihood* [2, 14]) by means of a parameter t ranging on the interval $[a, b]$, where

$$a := -\min_{j=1,\ldots,m} n(c', \tilde{f}_j) - \frac{1}{2},$$
$$b := \min_{j=1,\ldots,m} n(c'', \tilde{f}_j) + \frac{1}{2}.$$

In order to characterise the profile likelihood of the LNCC, we employ the following two results.

Theorem 1. *For each $\theta \in \Theta$ and each pair of classes $c', c'' \in \mathcal{C}$, there is a unique $t \in [a, b]$ such that*

$$\frac{P_\theta(c', \tilde{f})}{P_\theta(c'', \tilde{f})} = \frac{\left[n(c')+\frac{1}{2}+t\right] \prod_{j=1}^{m} \frac{\left[n(c', \tilde{f}_j)+\frac{1}{2}+t\right]}{\left[n(c')+\frac{1}{2}+t\right]}}{\left[n(c'')+\frac{1}{2}-t\right] \prod_{j=1}^{m} \frac{\left[n(c'', \tilde{f}_j)+\frac{1}{2}-t\right]}{\left[n(c'')+\frac{1}{2}-t\right]}}, \quad (10)$$

where $\frac{x}{0}$ is interpreted as $+\infty$ when x is positive, and as 1 when $x = 0$. Moreover, the right-hand side of (10) is a continuous, strictly increasing function of $t \in [a, b]$.

Theorem 1 defines a many-to-one relation between the elements of Θ and those of the interval $[a, b]$. For each $t \in [a, b]$, let Θ_t denote the set of all elements of Θ for which (10) is satisfied.

Theorem 2. *Let L, l', l'', p', p'' be the functions on $[a, b]$ defined by*

$$L(t) = \sup_{\theta \in \Theta_t} lik(\theta),$$
$$l'(t) = \left[n(c')+\tfrac{1}{2}+t\right]^{n(c')} \prod_{j=1}^{m} \frac{\left[n(c', \tilde{f}_j)+\frac{1}{2}+t\right]^{n(c', \tilde{f}_j)}}{\left[n(c')+\frac{1}{2}+t\right]^{n(c')}},$$
$$l''(t) = \left[n(c'')+\tfrac{1}{2}-t\right]^{n(c'')} \prod_{j=1}^{m} \frac{\left[n(c'', \tilde{f}_j)+\frac{1}{2}-t\right]^{n(c'', \tilde{f}_j)}}{\left[n(c'')+\frac{1}{2}-t\right]^{n(c'')}},$$
$$p'(t) = \left[n(c')+\tfrac{1}{2}+t\right] \prod_{j=1}^{m} \frac{\left[n(c', \tilde{f}_j)+\frac{1}{2}+t\right]}{\left[n(c')+\frac{1}{2}+t\right]},$$
$$p''(t) = \left[n(c'')+\tfrac{1}{2}-t\right] \prod_{j=1}^{m} \frac{\left[n(c'', \tilde{f}_j)+\frac{1}{2}-t\right]}{\left[n(c'')+\frac{1}{2}-t\right]},$$

for all $t \in [a, b]$, where both $\frac{0}{0}$ and 0^0 are interpreted as 1. Then

$$L \propto l'\,l''\,(p' + p''). \quad (11)$$

These two theorems can be used to perform LNCC-based classification without sampling. We first evaluate the maximum $\hat{t}$ of $L(t)$. Then, we check whether, for the values $t \in [a, b]$ such that $L(t) \geq \alpha L(\hat{t})$, the ratio on the right-hand side of (10) is always bigger than one. If so, we have that c' dominates c''. To see how this works, consider the classification task in Example 2. When testing whether or not c dominates

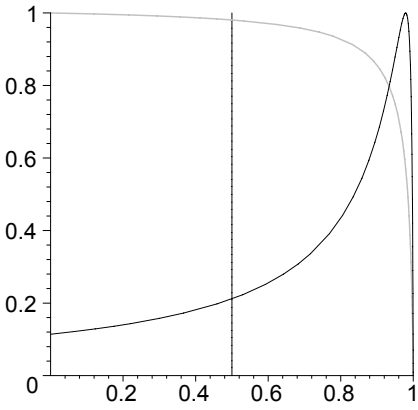

Figure 3: Profile likelihood functions for $P(c|\tilde{f}_1, \tilde{f}_2)$ in Example 3: with and without the probability of the new instance (black and grey curves, respectively).

$\neg c$, we have $[a, b] = [-\frac{11}{2}, \frac{3}{2}]$. For each $t \in [a, b]$, the right-hand side of (10) rewrites as $\frac{11+2t}{3-2t}$, while the likelihood in (11) is proportional to $(11 + 2t)^5 (3 - t)$; the resulting profile likelihood is depicted in Figure 2 (grey curve).

Example 3. *Consider a LNCC with a Boolean class C and two features F_1, F_2. We want to classify a new instance with features $\tilde{f}_1, \tilde{f}_2$, on the basis of a data set $\mathcal{D}$ containing $n(\cdot) = 100$ instances. In the data set $\mathcal{D}$, the class c has been observed $n(c) = 50$ times, always in conjunction with the feature $\tilde{f}_1$, but never with the feature $\tilde{f}_2$; that is, $n(c, \tilde{f}_1) = 50$ and $n(c, \tilde{f}_2) = 0$. Of the $n(\neg c) = 50$ observed instances with class $\neg c$, one had the feature $\tilde{f}_1$, and another one had the feature $\tilde{f}_2$; that is, $n(\neg c, \tilde{f}_1) = 1$ and $n(\neg c, \tilde{f}_2) = 1$. Figure 3 shows the profile likelihood function for $P(c|\tilde{f}_1, \tilde{f}_2)$ (compare with Figure 2) when the probability (9) of the new instance is considered in the likelihood function (black curve), and when it is not considered (grey curve).*

Hence, the LNCC classifies the new instance as c when α is sufficiently large (more precisely, when $\alpha \geq 0.22$); the same classification is obtained by the NBC with uniform prior and by the NCC$_\epsilon$ (for sufficiently large ϵ). By contrast, without using the probability (9) of the new instance in the likelihood function, the classifier would return both classes (if $\alpha \leq 0.98$), as does the standard NCC (that is, NCC$_\epsilon$ with $\epsilon = 0$), while the NBC with maximum-likelihood quantification returns the class $\neg c$ (at least when the usual likelihood function, without the probability of the new instance, is maximised). This is an example of the zero-counts issue discussed at the end of Section 2.2, which is the main reason why the ϵ-modification of NCC has been introduced and why we consider also the probability (9) of the new instance in the likelihood function.

4.2 Computational Complexity

The classification of an instance requires the iteration of the dominance test over all the possible pair of class labels, this task being clearly quadratic in $|\mathcal{C}|$. In order to perform the dominance test, the function $L(t)$ should be evaluated. This requires a number of operations which is linear in the number of attributes m. The same order of magnitude is required to compute the right-hand side of (10). In our preliminary implementation, τ equally spaced points over the interval $[a, b]$ have been considered. The numerical optimisation of the likelihood and identification of the α-cut was therefore simply performed by considering the value of the function $L(t)$ in these points. For the experiments, we adopted $\tau = 250$; empirically, increasing τ beyond this value resulted only in negligible differences in the classifications produced by LNCC. Thus, for practical purposes, we can consider τ as a constant, and we obtain $O(m|\mathcal{C}|^2)$ complexity (as for the NCC, [17]).

5 Experiments

To describe the performance of a credal classifier, we need multiple indicators. In particular, we adopt the following:

- *determinacy* (Det): the percentage of instances classified with a single class;
- *single accuracy* (Sgl-acc): the accuracy over the instances classified with a single class;
- *set-accuracy*: the accuracy over the instances classified with more classes;
- *indeterminate output size*: the average number of classes returned when the classification is indeterminate.

Note that when NCC is determinate, it returns the same class as NBC; this is due to the uniform prior being included in the IDM. This cannot be guaranteed for LNCC; however in our experiments LNCC, when precise, generally returned the same class as NBC. Thus, the single accuracy of NCC [resp. LNCC] is equivalent to the accuracy achieved by NBC on the instances determinately classified by NCC [resp. LNCC]. A credal classifier does a good job at isolating hard-to-classify instances if its Bayesian counterpart has low accuracy on the instances which are indeterminately classified. We denote as *NBC-I* the accuracy of naive Bayes on the instances indeterminately classified by the credal classifier at hand (NCC or LNCC, depending on the context). A large drop between single accuracy and NBC-I means thus that the credal

classifier is effective at isolating instances which are hard to classify.

Unfortunately, there is so far no single indicator which can reliably compare two credal classifiers. The *discounted-accuracy* (D-acc, [5]) has been proposed for this purpose; it is defined as $\frac{1}{n}\sum_{i\in acc}\frac{acc_i}{|output_i|}$, where, with reference to the i-th instance, acc_i denotes whether the set of returned classes contains or not the actual one and $|output_i|$ denotes the number of classes returned. On each instance, the classifier is thus given 0 if inaccurate or $1/|output_i|$ if accurate. Yet, discounted-accuracy sees as equivalent, in the long term, a *vacuous* classifier which returns all classes and a *random* classifier which returns a single class at random. However, the vacuous classifier should be generally preferred over the random one; this is clear if one thinks for instance of the diagnosis of a disease. In a way the vacuous, unlike the random, is aware of being ignorant; yet discounted-accuracy does not capture this point. In fact, the design of metrics to rank credal classifiers is an important *open* problem. Moreover, when dealing with credal classifiers with considerably different determinacy, discounted-accuracy favors the more determinate ones. We thus try to compare LNCC and NCC (in its NCC_ϵ generalisation) when they have the same determinacy. For this purpose we tried different values of ϵ for NCC_ϵ and α for LNCC; more precisely, denoting also the value of α as a subscript, we considered: $\text{NCC}_{0.05}$, $\text{NCC}_{0.15}$, $\text{NCC}_{0.25}$, $\text{NCC}_{0.35}$; $\text{LNCC}_{0.35}$, $\text{LNCC}_{0.55}$, $\text{LNCC}_{0.75}$, $\text{LNCC}_{0.95}$.

5.1 Artificial Data

We generated artificial data sets, considering a binary class and 10 binary features, under a naive data generation mechanism. We set the marginal chances of classes as uniform, while we drew the conditional chances of the features under the constraint $|\theta_{f_j|c'} - \theta_{f_j|c''}| \geq 0.1$ for each $c', c'' \in \mathcal{C}$ and $j = 1, \ldots, m$; the constraint forced each feature to be truly dependent on the class. We drew such chances 20 times uniformly at random and we consider the sample sizes $d \in \{25, 50, 100\}$. For each pair (θ, d) we generated 30 training sets and a huge test set of 10000 instances. For each sample size, we thus perform 20 θ × 30 trials = 600 training/test experiments. Note that, dealing with two classes, set-accuracy is fixed to 100% and indeterminate output size to 2; we do not need thus to consider these indicators.

In Figure 5 we show how the determinacy of NCC and LNCC varies with the sample size, choosing pairs $\{\alpha, \epsilon\}$ which produce reasonably comparable curves. Interestingly, NCC is more sensitive than LNCC to the sample size d; the determinacy of NCC steeply increases with d, unlike that of LNCC. In fact, NCC becomes determinate once the rank of the classes does not change under all the different priors of the IDM; but the smoothing effect of the prior decreases with d. The same is known to happen with likelihood-based methods, but convergence towards the precise model is slower, as shown by the comparison in Figure 1.

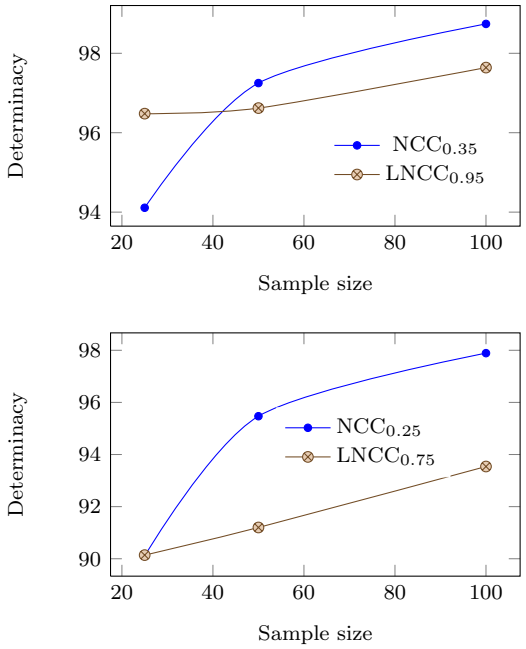

Figure 5: Determinacy of NCC and LNCC as a function of the sample size d.

Classifier	n	Det	Sgl-acc	D-acc	NBC-I
$\text{NCC}_{0.25}$	25	90.1	90.4	86.5	54.5
$\text{LNCC}_{0.75}$	25	90.1	90.4	86.6	52.8
$\text{NCC}_{0.05}$	50	91.7	91.5	88.2	57.4
$\text{LNCC}_{0.75}$	50	91.2	91.8	88.2	55.3
$\text{NCC}_{0.25}$	100	97.9	90.5	89.7	51.2
$\text{LNCC}_{0.95}$	100	97.7	90.6	89.7	51.4

Table 1: Performance indicators for NCC and LNCC, for choices of α and ϵ leading to close determinacies; each number is an average over 600 experiments.

It is interesting to compare LNCC and NCC when they have, for the same sample size, very close determinacy. This is this the case of $\text{NCC}_{0.25}$ and $\text{LNCC}_{0.75}$ for d=25; of $\text{NCC}_{0.05}$ and $\text{LNCC}_{0.75}$ for d=50; of $\text{NCC}_{0.25}$ and $\text{LNCC}_{0.95}$ for d=100. Note that in general it is not possible to predict in advance which choice of ϵ and α will allow to obtain similar determinacy from LNCC and NCC. However, when NCC and LNCC achieve the same determinacy, their performances are very similar also on the remaining indica-

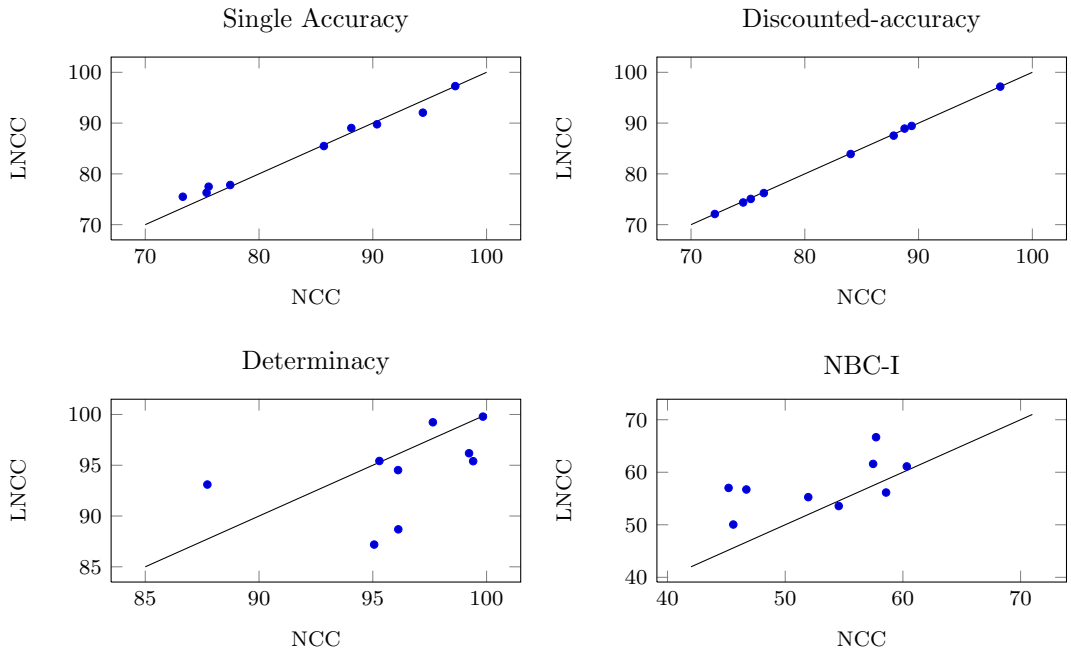

Figure 4: LNCC versus NCC: scatter plots on different UCI binary data sets.

tors, as shown in Table 1. This suggests that, for the same level of determinacy, NCC and LNCC become indeterminate on roughly the same instances, despite the different derivation of their algorithms. Note also the large drop between Sgl-acc and NBC-I for both classifiers, showing that both NCC and LNCC can be seen as extending NBC towards increased reliability.

5.2 Binary Data Sets from UCI

We then considered 9 binary data sets (containing 2 classes) from the UCI repository; the number of instances ranges from 57 to 3000 and the number of features from 8 to 60. Since the data sets are binary, set-accuracy and indeterminate output size can only be respectively 100% and 2; we do not consider thus these indicators. For each credal classifiers we report instead Sgl-acc and NBC-I, namely the accuracy of NBC when the credal classifier is respectively determinate[7] and indeterminate. If there is a large difference between these two indicators, the credal classifier is doing a good job at isolating instances which are difficult to classify for NBC. Moreover, we report determinacy and D-acc to provide a general overview of the classifiers' behavior.

The results in Table 2 show that when LNCC and NCC have close determinacy, they also have very similar performance on the remaining indicators, as in the previous experiments. Also in this case, there is in general a large drop between Sgl-acc and NBC-I, showing that both credal classifiers are effective at isolating instances that are hard to classify for NBC.

Dataset	Classifier	Det	Sgl-acc	D-acc	NBC-I
german	$NCC_{0.05}$	96.1	75.6	74.6	58.6
german	$LNCC_{0.95}$	95.7	75.7	74.6	57.5
haberman	$NCC_{0.05}$	95.1	73.3	72.1	45.6
haberman	$LNCC_{0.95}$	93.9	73.8	71.9	50.2
hepatitis	$NCC_{0.05}$	95.3	85.7	60.3	84.0
hepatitis	$LNCC_{0.75}$	95.4	85.5	61.1	83.9

Table 2: Results for LNCC and NCC on UCI data sets, for choices of α and ϵ leading to close determinacies. We report results only for 3 out of 9 analyzed data sets, because the remaining data sets only show very similar findings: namely that when LNCC and NCC have close determinacy, their performance on all indicators is substantially identical.

However, it is also interesting to see what happens if we set a default choice for ϵ and α. We set ϵ to 0.05 for NCC, thus considering a minimal variation over the NCC of [17], aimed at avoiding issues with zero counts. As for LNCC, we adopted a trial and error approach, from which $\alpha = 0.75$ appeared as a reasonable compromise between determinacy and reliability of the classifier. *On average*, NCC has slightly

[7]This follows from NBC returning the same class as the credal classifier, when the latter is determinate (this is theoretically guaranteed for NCC and only empirically verified for HNCC); Sgl-acc can be thus seen as measuring also the accuracy of NBC when the credal classifier is determinate.

higher determinacy (96.3% vs. 94.3%) and slightly lower single-accuracy (84.2% vs. 85.5%) than LNCC. Moreover, the *area of ignorance* (instances indeterminately classified) of NCC is slightly more difficult to classify for NBC than the area of ignorance of LNCC: the average NBC-I is 53.1% vs. 57.6%. In fact, NCC is slightly more determinate and thus more selective in deciding when to become indeterminate. The average discounted accuracy of the two classifiers is very close (82.8% vs 82.7%). However, averaging indicators over data sets is questionable; we thus also present in Figure 4 the scatter plots of such indicators. On each data set there is little difference between the single-accuracy of NCC and LNCC; the same holds also for the discounted-accuracy. On the other hand, there are sometimes considerable differences between NCC and LNCC as for the determinacy, which tends to be larger for NCC, and as for NBC-I, which tends to be larger for LNCC. In general, when the difference in determinacy between NCC and LNCC increases, so does the difference in NBC-I between LNCC and NCC.

6 Conclusions and Outlooks

We have presented an alternative, likelihood-based, approach to the imprecise-probabilistic quantification of a naive classifier. A numerical comparison with the naive credal classifier (in its modified formulation to cope with zero-count issues) shows that, despite their deeply different derivations, the performance of the two classifiers is very similar when they produce more or less the same amount of indeterminate classifications. When the amount of indeterminacy between the two classifiers is considerably different, a meaningful comparison is difficult: this would require modelling the trade-off between accuracy and informativeness by means of one or more performance indicators, which is currently one of the most important *open* problems in credal classification.

Extensions of the new approach to more complex independence structures (e.g., tree-augmented naive), incomplete data sets, and continuous features seem to be worth of future investigations.

Acknowledgements

The research in this paper has been partially supported by the Swiss NSF grants n. 200020-132252 and by the Hasler foundation grant n. 10030. The authors wish to thank the anonymous referees for their helpful comments.

Appendix

Proof of Theorem 1. Let g be the function assigning to each $t \in [a,b]$ the corresponding right-hand side of (10). We prove the theorem by showing that for each $x \in [0,+\infty]$ there is a unique $t \in [a,b]$ such that $g(t) = x$. When $t \in (a,b)$, all the sums of three terms (of the form $\left[n + \frac{1}{2} \pm t\right]$) in the expression of $g(t)$ are positive. In this case, each fraction

$$\frac{\left[n(c', \tilde{f}_j) + \frac{1}{2} + t\right]}{\left[n(c') + \frac{1}{2} + t\right]} \tag{12}$$

is a continuous, increasing function of t, since it is differentiable with derivative

$$\frac{n(c') - n(c', \tilde{f}_j)}{\left[n(c') + \frac{1}{2} + t\right]^2} \geq 0.$$

Therefore, the numerator of $g(t)$ is a continuous, strictly increasing function of $t \in (a,b)$, since it is the product of m continuous, increasing functions and of the continuous, strictly increasing function $\left[n(c') + \frac{1}{2} + t\right]$. Analogously, we can prove that the denominator of $g(t)$ is a continuous, strictly decreasing function of $t \in (a,b)$, and therefore g is continuous and strictly increasing on (a,b).

In order to prove Theorem 1, it now suffices to show that

$$\lim_{t \downarrow a} g(t) = g(a) = 0 \quad \text{and} \quad \lim_{t \uparrow b} g(t) = g(b) = +\infty. \tag{13}$$

We prove the first expression: the second one can be proved analogously. As t tends to a from above, the denominator of $g(t)$ tends to a positive constant, which is reached when $t = a$. To study the limit of the numerator of $g(t)$, let j_0 be such that $n(c', \tilde{f}_{j_0}) = \min_{j=1,\ldots,m} n(c', \tilde{f}_j)$. We can distinguish two cases: either $n(c', \tilde{f}_{j_0}) = n(c')$, or $n(c', \tilde{f}_{j_0}) < n(c')$. In the first case, $n(c', \tilde{f}_j) = n(c')$ for all j, and the numerator reduces to $\left[n(c') + \frac{1}{2} + t\right]$, since the fractions (12) are all equal 1. Therefore, in this case, the limit of the numerator of $g(t)$ as t tends to a from above is 0, because $a = -\frac{1}{2} - n(c')$. In the second case, $a = -\frac{1}{2} - n(c', \tilde{f}_{j_0})$, and thus the limit of the numerator of $g(t)$ as t tends to a from above is 0 as well, because the fraction (12) with $j = j_0$ tends to 0. Moreover, in both cases, the numerator of $g(t)$ is 0 when $t = a$, since $\frac{0}{0}$ is interpreted as 1. This proves the first expression of (13) and hence the theorem. □

Proof of Theorem 2. Let l_d, π', π'', r be the functions on Θ defined by

$$l_d(\theta) = \prod_{c \in \mathcal{C}} \left(\theta_c^{n(c)} \prod_{j=1}^{m} \prod_{f_j \in \mathcal{F}_j} \theta_{f_j|c}^{n(c,f_j)} \right), \quad r(\theta) = \frac{\pi'(\theta)}{\pi''(\theta)},$$
$$\pi'(\theta) = \theta_{c'} \prod_{j=1}^{m} \theta_{\tilde{f}_j|c'}, \quad \pi''(\theta) = \theta_{c''} \prod_{j=1}^{m} \theta_{\tilde{f}_j|c''},$$

for all $\theta \in \Theta$. Then, up to normalisation, the considered likelihood function lik corresponds to $l_d\,(\pi' + \pi'')$, since $l_d(\theta)$ is the probability of the observed data set $\mathcal{D}$ according to the NBC specified by θ, while $\pi'(\theta)$ and $\pi''(\theta)$ are

the probabilities of the instance under consideration (according to the NBC specified by θ), when its class is c' and c'', respectively. Therefore, in particular, $r(\theta)$ corresponds to the left-hand side of (10).

For each $t \in [a,b]$, consider now the function

$$l_d\,(\pi')^{\frac{1}{2}+t}\,(\pi'')^{\frac{1}{2}-t} = l_d\,\pi'\,r^{t-\frac{1}{2}} = l_d\,\pi''\,r^{t+\frac{1}{2}}. \qquad (14)$$

This function corresponds to the function l_d with modified counts n (which are in general not integer anymore, but still nonnegative), and can be easily maximised. Its maximum is taken in $\hat{\theta}(t)$, where $\hat{\theta}(t)$ is the maximum likelihood quantification of the NBC with respect to the modified counts: that is,

$$\hat{\theta}(t)_{c'} = \frac{n(c') + \frac{1}{2} + t}{n(\cdot) + 1}, \quad \hat{\theta}(t)_{\tilde{f}_j\,|\,c'} = \frac{n(c', \tilde{f}_j) + \frac{1}{2} + t}{n(c') + \frac{1}{2} + t},$$

$$\hat{\theta}(t)_{f_j\,|\,c'} = \frac{n(c', f_j)}{n(c') + \frac{1}{2} + t} \quad \text{for all } f_j \neq \tilde{f}_j,$$

$$\hat{\theta}(t)_{c''} = \frac{n(c'') + \frac{1}{2} - t}{n(\cdot) + 1}, \quad \hat{\theta}(t)_{\tilde{f}_j\,|\,c''} = \frac{n(c'', \tilde{f}_j) + \frac{1}{2} - t}{n(c'') + \frac{1}{2} - t},$$

$$\hat{\theta}(t)_{f_j\,|\,c''} = \frac{n(c'', f_j)}{n(c'') + \frac{1}{2} - t} \quad \text{for all } f_j \neq \tilde{f}_j,$$

$$\hat{\theta}(t)_{c} = \frac{n(c)}{n(\cdot) + 1} \quad \text{and} \quad \hat{\theta}(t)_{f_j\,|\,c} = \frac{n(c, f_j)}{n(c)} \quad \text{for all } f_j,$$

where c is any class different from c', c''. Therefore, in particular, $r(\hat{\theta}(t))$ corresponds to the right-hand side of (10): that is, $\hat{\theta}(t) \in \Theta_t$.

Since $\hat{\theta}(t)$ maximises the function (14) over all $\theta \in \Theta$, it also maximises both functions $l_d\,\pi'$ and $l_d\,\pi''$ over all $\theta \in \Theta$ such that $r(\theta) = r(\hat{\theta}(t))$. That is, $\hat{\theta}(t)$ maximises both functions $l_d\,\pi'$ and $l_d\,\pi''$ over all $\theta \in \Theta_t$, and therefore it also maximises their sum $l_d\,(\pi' + \pi'')$ over all $\theta \in \Theta_t$. Since this last function corresponds, up to normalisation, to the considered likelihood function lik, we obtain the result $L(t) = lik(\hat{\theta}(t))$.

In order to prove Theorem 2, it suffices to show that $lik(\hat{\theta}(\cdot))$ is proportional to $l'\,l''\,(p' + p'')$; that is, it suffices to show that

$$l_d(\hat{\theta}(t))\Big(\pi'(\hat{\theta}(t)) + \pi''(\hat{\theta}(t))\Big) = \gamma\,l'(t)\,l''(t)\,\big(p'(t) + p''(t)\big),$$

where the proportionality constant $\gamma \in (0, +\infty)$ may depend on anything but t. Since

$$\pi'(\hat{\theta}(t)) + \pi''(\hat{\theta}(t)) = \frac{1}{n(\cdot) + 1}\,\big(p'(t) + p''(t)\big)\,,$$

it only remains to show that $l_d(\hat{\theta}(t))$ is proportional to $l'(t)\,l''(t)$. In the expression $l_d(\hat{\theta}(t))$, we can drop all factors for classes c different from c', c'', because $\hat{\theta}(t)_c$ and $\hat{\theta}(t)_{f_j\,|\,c}$ do not depend on t when c is different from c', c''. The desired result follows easily when one considers that

$$n(c) = \sum_{f_j \in \mathcal{F}_j} n(c, f_j)$$

for all $c \in \mathcal{C}$ and all $j \in \{1, \ldots, m\}$. □

References

[1] M. Cattaneo. *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich, 2007.

[2] M. Cattaneo. A generalization of credal networks. In *ISIPTA '09*, pages 79–88. SIPTA, 2009.

[3] M. Cattaneo. Likelihood-based inference for probabilistic graphical models: Some preliminary results. In *PGM 2010*, pages 57–64. HIIT Publications, 2010.

[4] G. Corani and A. Benavoli. Restricting the IDM for classification. In *IPMU 2010*, pages 328–337. Springer, 2010.

[5] G. Corani and M. Zaffalon. Lazy naive credal classifier. In *Proc. 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*, pages 30–37. ACM, 2009.

[6] F.G. Cozman. Credal networks. *Artif. Intell.*, 120:199–233, 2000.

[7] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.*, 29:103–130, 1997.

[8] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Mach. Learn.*, 29:131–163, 1997.

[9] D.J. Hand and K. Yu. Idiot's Bayes—Not so stupid after all? *Int. Stat. Rev.*, 69:385–398, 2001.

[10] D.J. Hudson. Interval estimation from the likelihood function. *J. R. Stat. Soc., Ser. B*, 33:256–262, 1971.

[11] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT press, 2009.

[12] M.G. Madden. On the classification performance of TAN and general Bayesian networks. *Knowl.-Based Syst.*, 22:489–495, 2009.

[13] S. Moral. Calculating uncertainty intervals from conditional convex sets of probabilities. In *UAI '92*, pages 199–206. Morgan Kaufmann, 1992.

[14] Y. Pawitan. *In All Likelihood*. Oxford University Press, 2001.

[15] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[16] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *J. R. Stat. Soc., Ser. B*, 58:3–34, 1996.

[17] M. Zaffalon. Statistical inference of the naive credal classifier. In *ISIPTA '01*, pages 384–393, 2001.

[18] M. Zaffalon. The naive credal classifier. *J. Stat. Plann. Inference*, 105:5–21, 2002.