Discussion

Learning imprecise probability models: Conceptual and practical challenges

Fabio Gagliardi Cozman

Escola Politecnica, Universidade de Sao Paulo, Av. Prof. Mello Moraes 2231 - Cidade Universitaria, Sao Paulo, SP, Brazil

ARTICLE INFO

Article history:
Received 18 November 2013
Received in revised form 22 April 2014
Accepted 23 April 2014
Available online 28 April 2014

Keywords:
Imprecise probabilities
Multinomial learning
Credal networks

ABSTRACT

The paper by Masegosa and Moral, on “Imprecise probability models for learning multinomial distributions from data”, considers the combination of observed data and minimal prior assumptions so as to produce possibly interval-valued parameter estimates. We offer an evaluation of Masegosa and Moral’s proposals.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

The paper by Masegosa and Moral [7], on “Imprecise probability models for learning multinomial distributions from data”, not only deals with technically demanding issues but also with several complex conceptual questions concerning the nature of statistical models. On top of that, they also consider “applications to learning credal networks.” Hence the reader cannot have doubts that Masegosa and Moral have set their eyes on ambitious goals.

Their paper does contribute with several novel ideas, and the balance of their results is certainly positive and welcome. The point of the present commentary is to examine some of their questionable assumptions and proposals, so as to investigate ways in which they can be improved, and to suggest issues that deserve further consideration.

2. Learning with few prior assumptions: the learning principle and the ISSDM

The theme of Masegosa and Moral’s paper is the construction of a data generating model from observed data. One might argue that the construction of a model should assume nothing a priori. But what constitutes “nothing a priori” is a difficult matter. The Bayesian strategy is to express “ignorance” about parameters through a prior distribution. To indicate that nothing is assumed a priori, a Bayesian may adopt flat priors, reference priors, maximum-entropy priors [2]. There is no real agreement on how to encode ignorance in such a way, and some of these efforts seem more desperate than heroic.

Another possibility is to encode the absence of a priori assumptions by adopting a set of prior distributions. The larger the set, the less is assumed about the parameters of interest. This is exactly the strategy followed by Masegosa and Moral.

The first part of their paper looks at the specific case of multinomial distributions. Thus we have to estimate parameters \( \{\theta_i\}_{i=1}^k \), all nonnegative, such that \( \sum \theta_i = 1 \). Each parameter \( \theta_i \) is interpreted as the probability of a particular value of a variable \( X \); that is, \( \theta_i \) means \( P(X=x_i) \). The natural conjugate prior is given by a Dirichlet distribution; that is, a distribution proportional to \( \prod_{i=1}^k \theta_i^{s_i-1} \), where \( s \) and \( \{t_i\}_{i=1}^k \) are parameters. Masegosa and Moral dismiss methods that try to encode...
ignorance by a single Dirichlet distribution, and instead move to models based on sets of prior Dirichlet distributions. As noted in the previous paragraph, their motivation is clear: ignorance cannot be translated into a single prior; rather, ignorance should be expressed by a set of priors.

However, it is not easy to select a set of prior distributions. Suppose one adopts a maximal set of prior distributions (a set containing every possible distribution over a given space), and Bayes rule is applied elementwise. Then the set of posterior distributions is also maximal. That is, if we use a maximal set of distributions to encode a priori ignorance, nothing can ever be learned — ex nihilo nihil fit, mocks the Bayesian. One way to start from complete ignorance and to reach useful inferences is to abandon Bayes rule, for example by adopting Dempster rule [3, Section 6]; but if we are to stay with Bayes rule, we must judiciously construct our sets of prior distributions. Such sets must be large enough to properly encode ignorance, but not so large that they lead to vacuous inferences.

A popular way to combine Dirichlet distributions with sets of prior distributions is to employ the Imprecise Dirichlet Model (IDM). Here we take a set of Dirichlet distributions where parameter $\alpha$ is fixed but parameters $\tau_i$ are left free [15]. The IDM displays the interesting feature that each one of values $x_i$ is assigned the maximal probability interval $[0, 1]$. That is, very little is assumed a priori. The IDM also embodies a principle of symmetry: all values of interest are treated equally. If several variables are considered at once, then it is likely that many combinations of their values are never observed in a reasonably-sized dataset. Hence prior distributions have increased importance in producing estimates. And exactly because many variables are present, we must find ways to specify prior distributions without too many controlling parameters, lest ignorance by a single Dirichlet distribution, and instead move to models based on sets of prior Dirichlet distributions.

Each variable is assumed independent of its nondescendants given its parents. In short, a credal network can be a structure consisting of a graph where each node is a variable, and where edges encode independences. Each variable is assumed independent of its nondescendants given its parents. In short, a credal network can be understood as a set of Bayesian networks [8]. Most applications of credal networks in the literature adopt the assumption that all Bayesian networks represented by a given credal network share the same edges (hence they share the same inde-
pendences). Another common assumption is that the conditional distributions associated with a particular variable do not restrict in any way the conditional distributions associated with any other variable: the credal network is then called separately specified. Masegosa and Moral depart from these assumptions: their priors introduce restrictions between conditional distributions, and, more importantly, then consider a generalized kind of credal network that may encode many different independence relations at once.

There are several challenges in learning this sort of credal network from data, and to do so the authors put together an array of techniques: scores, discretization methods, and Monte Carlo schemes for inference. The latter idea is an interesting contribution, as existing algorithms that operate with credal networks usually focus on optimization, while here the focus is on sampling. The only concern is, again, that the initial beauty of almost-no-assumption-models is blurred by the machinery required for practical operation.

However, in the middle of this battle, Masegosa and Moral have found a true gem. It turns out that by clever selection of parameters, one can guarantee that a single graph is built from data, but a graph where some edges are necessary in that they are selected for all choices of parameters, while others are ambiguous in that they are sensitive to the choice of parameters. This sort of graphical structure is rather intuitive and informative, and should be quite useful not only when learning networks from data, but also when building networks from opinions of a single expert, or from opinions of a sets of experts. The authors must be congratulated for this contribution.

Finally, a word on the evaluation of credal networks. This is always a difficult matter, well discussed in the literature (as noted by the authors). Their experiments reveal interesting features of their methods, and it seems that more practical experience is needed. For instance, how can we really compare a Naive Bayes classifier with a classifier that sometimes outputs set of classes? The authors argue that “it should be better to produce an imprecise result” when a precise answer will be sensitive to parameters, but some may disagree — in practice, one might argue that a precise answer is always better. Thus comparisons with strategies that always yield a precise answer (for instance, minimax strategies) may be useful in the future.

4. Final comments

The problems tackled by Masegosa and Moral are complex and sometimes perplexing; their proposals are bold and it is only fair that they raise further difficult questions. The learning principle surely deserves more discussion. The ISSDM family of models encodes important ideas that should be pursued in earnest, not only by theoretical study but also by practical evaluation. Finally, the insights by Masegosa and Moral concerning credal network learning, particularly in connection with ambiguous links, are quite welcome and merit practical use.

I am grateful to have had the opportunity to comment on this valuable contribution, and can only hope that the authors continue producing similarly stimulating material.

References