Density estimation with ensembles of randomized poly-trees

Sourour Ammar, Philippe Leray, Boris Defourny, Louis Wehenkel

Density estimation with ensembles of randomized poly-trees

Sourour Ammar [1 2] SOUROUR.AMMAR@ETU.UNIV-NANTES.FR
Philippe Leray [1] PHILIPPE.LERAY@UNIV-NANTES.FR
Boris Defourny [3] BORIS.DEFOURNY@ULG.AC.BE
Louis Wehenkel [3] L.WEHENKEL@ULG.AC.BE

[1] Laboratoire d'Informatique de Nantes Atlantique (LINA) UMR 6241, École Polytechnique de l'Université de Nantes, France
[2] Laboratoire d'Informatique, Traitement de l'Information et des Systèmes (LITIS) EA 4108 - Institut National des Sciences Appliquées de Rouen, France
[3] Department of Electrical Engineering and Computer Science & GIGA-Research, University of Liège, Belgium

1. Motivation

Learning of Bayesian networks aims at modeling the joint density of a set of random variables from a random sample of joint observations of these variables (Naïm et al., 2007). Such a graphical model may be used for elucidating the conditional independences holding in the datagenerating distribution, for automatic reasoning under uncertainties, and for Monte-Carlo simulations. Unfortunately, currently available algorithms for Bayesian network structure learning are either restrictive in the kind of distributions they search for, or of too high computational complexity to be applicable in high dimensional spaces.

Ensembles of weakly fitted randomized models have been studied intensively and used successfully in the supervised learning literature during the last two decades. Among the advantages of these methods, let us quote the improved scalability of their learning algorithms thanks to randomization and the improved predictive accuracy the induced models thanks to their higher flexibility in terms of bias/variance trade-off. For example, ensembles of extremely randomized trees have been applied successfully in very complex high-dimensional tasks, such as image and sequence classification (Geurts et al., 2006).

In this work we explore the Perturb and Combine idea celebrated in supervised learning in the context of probability density estimation in high-dimensional spaces. We propose a new family of unsupervised learning methods of mixtures of large ensembles of randomly generated poly-trees. The specific feature of these methods is their scalability to very large numbers of variables and training instances. We explore various variants of these methods empirically on a set of discrete test problems of growing complexity.

2. Methods

2.1. Poly-Tree density models

Let $X = \{X_1, \ldots, X_n\}$ denote a finite set of discrete random variables.

A poly-tree model P for the density over X is defined by a directed acyclic graph which skeleton is acyclic and connected, and the set of vertices of which is in bijection with X and with a set of conditional densities $\mathbb{P}_P(X_i|pa_P(X_i))$, where $pa_P(X_i)$ denotes the set of variables in bijection with the parents of X_i in P. It represents graphically the density factorization

$$\mathbb{P}_P(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}_P(X_i|pa_P(X_i)). \qquad (1)$$

Poly-tree models can be used for probabilistic inference over $\mathbb{P}(X_1, \ldots, X_n)$ with a computational complexity linear in the number of variables n (Pearl, 1986).

One can define nested subclasses of poly-tree density models by imposing constraints on the maximum number p of parents of any node. In these subclasses, not only inference but also parameter learning is of linear complexity in the number of variables. The smallest such subclass is called the tree subspace, in which nodes have exactly one parent ($p = 1$).

2.2. Mixture models of poly-trees

A mixture model of m poly-tree models $(P_1, \ldots, P_m)$ is defined as a convex combination of the elementary poly-tree models, ie.

$$\mathbb{P}_M(X_1, \ldots, X_n) = \sum_{i=1}^{m} \mu_i \mathbb{P}_{P_i}(X_1, \ldots, X_n), \qquad (2)$$

where $\mu_i \in [0, 1]$ and $\sum_i^m \mu_i = 1$.

While single poly-tree models impose restrictions on the kind of densities they can faithfully represent, mixtures of poly-trees are universal approximators.

2.3. Learning a random mixture from data

Let X be a set of discrete random variables, and $D = (x^1, \cdots, x^d)$ be a sample of joint observations $x^i = (x_1^i, \cdots, x_n^i)$ drawn from some datagenerating distribution $\mathbb{P}_G(X)$. Let $\mathcal{P}$ be the space of all possible poly-tree graphical structures defined over X.

Our generic procedure for generating a random mixture of poly-tree models from D is described by Algorithm 1; it receives as inputs X, D, m, and three procedures *DrawPolytree*, *LearnPars*, *ComputeWeight*.

Algorithm 1 (Learning random poly-tree mixtures)

1. Repeat for $i = 1, \cdots, m$:
   (a) $P_i = \mathit{DrawPolytree}(\mathcal{P})$,
   (b) For $j = 1, \cdots, n$:
   $\mathbb{P}_{P_i}(X_j|pa_{P_i}(X_j)) = \mathit{LearnPars}(P_i, X_j, D)$
   (c) $\mu_i = \mathit{ComputeWeight}(P_i, D, m)$
2. Return $\left(\mu_i, (\mathbb{P}_{P_i}(X_j|pa_{P_i}(X_j)))_{j=1}^n\right)_{i=1}^m$.

3. Experiments and preliminary results

In (Ammar et al., 2008) we report some first results with the above algorithm applied to datasets of size $d = 1000$ generated from discrete distributions with $n = 8$, which could be faithfully represented by a chain, a single tree, or a single poly-tree model.

In these simulations we have considered two different instances of *DrawPolytree*, namely a uniform draw over the class $\mathcal{P}$ of all poly-trees, and a uniform draw over the subclass $\mathcal{P}^1$ of trees. In order to achieve this for $m \in \{1, 2, \ldots, 1000\}$, we have used efficient algorithms for sampling trees given in (Quiroz, 1989).

For parameter learning, we used maximum a posteriori values given the dataset and structure, while assuming non-informative priors on the parameters. Concerning the μ_is, we used a uniform weighting strategy, ie. $\mathit{ComputeWeight}(P_i, D, m) = 1/m$.

Overall, these results showed that the quality of the mixture-models converges rather rapidly (ie. for $m \approx 20$), and that the poly-tree mixtures were slightly superior when targeting poly-tree datagenerating distributions, while the mixtures of trees were superior in the other two cases. We also observed a slightly non-monotonic behavior of the model quality with growing values of m, which we suspect to be related to the uniform weighting scheme.

In the immediate future, we will carry out further more systematic experiments on larger problems and spanning different versions of the algorithm.

In particular, we will consider non-uniform weighting schemes, by exploiting the score obtained for a given structure and dataset so as to downweight structures that fit less well to the datagenerating distribution. We will also consider sampling from the spaces $\mathcal{P}^p$ of poly-trees of bounded number of parents.

Experiments will be made over a richer set of datagenerating distributions, in particular ones that can not be represented faithfully by a single poly-tree model. For instance, we will consider general directed acyclic graph models as datagenerating distributions.

We will compare our algorithm in terms of sample and computational efficiency with Bayesian network structure learning and algorithms targeting an *optimal* mixture of tree-models (Meila-Predoviciu, 1999).

Subsequently, we plan to extend our approach to handle continuous variables and incomplete datasets.

Acknowledgments

This work presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.

References

Ammar, S., Leray, P., & Wehenkel, L. (2008). Estimation de densité par ensembles aléatoires de polyarbres. *Proceedings of JFRB*.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3–42.

Meila-Predoviciu, M. (1999). *Learning with mixtures of trees*. Doctoral dissertation, MIT.

Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O., & Becker, A. (2007). *Réseaux bayésiens*. Paris: Eyrolles. 3 edition.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, *29*, 241–288.

Quiroz, A. (1989). Fast random generation of binary, t-ary and other types of trees. *Journal of Classification*, *6*, 223–231. available at http://ideas.repec.org/a/spr/jclass/v6y1989i1p223-231.html.