Maximum Likelihood Markov Hypertrees

Nathan Srebro, David Karger & Tommi Jaakkola

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: We study the problem of finding a maximum likelihood distribution among Markov networks of bounded tree-width.

Motivation: One of the challenges of unsupervised learning, given a sample of observations, is to determine the distribution law from which the samples were drawn. The predicted distribution can be used to make predictions about future, partially observed data. In order for such a prediction to be useful, it must be realizable in a model allowing efficient calculations of marginal probabilities.

One popular class of such models are *Markov networks*, which use an undirected graph to represent dependencies among variables. Markov networks of low *tree-width* (i.e. having a triangulation with small cliques) allow efficient computations, and are useful as learned probability models [8]. A well studied case is that in which the dependency structure is known in advance. In this case the underlying graph is built based on prior knowledge, and a maximum likelihood Markov network over this specific graph is sought [5].

However, in many situation the structure is not known in advance. We we must then learn, from the data, both the graph structure, and the parameters of the Markov network over the graph. Given a target complexity, specified as a maximum allowed tree-width k, and determined by the amount of training data and the available computational resources for using the model, the maximum likelihood Markov network over any graph of tree-width at most k, is sought.

A classic result in the field is that of Chow and Liu [1], who in 1968 gave an efficient algorithm for finding the maximum likelihood Markov tree (trees are graphs of tree-width one). Chow and Liu's algorithm is commonly used for learning a distribution from data. However, in many cases there is enough training data and computational resources available to justify using Markov networks of higher tree-width. We would thus like to extend the work of Chow and Liu to finding maximum likelihood Markov networks of higher tree-width.

Previous Work: Chow and Liu [1], showed how the maximum likelihood Markov tree problem can be cast as a combinatorial problem of finding the maximum weight tree. This combinatorial problem is well-studied and linear-time algorithms for it are known. This leads to very efficient algorithms for the maximum likelihood Markov tree problem.

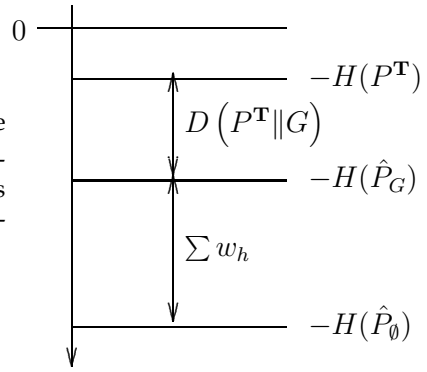
The problem of finding a maximum likelihood Markov network of bounded tree width is discussed in [5]. However, work on the problem has focused mostly on local search heuristics, without any performance guarantees or bounds on the complexity of the problem. Most notably, Malvestuto [4] suggested local search heuristics that employ the notion of a maximal acyclic hypergraph, which we call hypertrees in this work.

Approach: Similar to the work of Chow and Liu, we are able to cast the learning problem as a combinatorial optimization problem on graphs. We show that learning a maximum likelihood Markov network of bounded tree-width is equivalent to finding a maximum weight hypertree. This equivalence gives rise to global, integer-programming based, approximation algorithms with provable performance guarantees, for the learning problem.

The equivalence also allows us to study the computational hardness of the learning problem. We show that learning a maximum likelihood Markov network of bounded tree-width is NP-hard, and are able to provide bounds on the hardness of finding approximate solutions.

Impact: Estimating tractable density models from data is an important and ubiquitous problem. For low tree-width Markov networks, we have provided a combinatorial characterization of the problem structure, established a preliminary hardness result concerning the associated estimation problem, and developed approximation algo-

$D\left(P^{\mathbf{T}} \| G\right) = (\mathbf{E}_{P^{\mathbf{T}}}[\emptyset] - H(P^{\mathbf{T}})) - \sum_{h \in \mathit{Clique}(G)} w_h$: The weight of a hypertree G corresponds to the gain in log-likelihood versus a fully independent model. The weights w_h of *hyperedges*, or subsets of variables, form an interesting information decomposition.



rithms with provable performance guarantees. This is a substantial extension of the work of Chow and Liu on maximum likelihood Markov trees.

We achieved these results by introducing algorithm-theoretic techniques, such as integer programming, randomized rounding and approximation ratio proofs, to a machine learning problem.

Another interesting insight that is gained by casting the maximum likelihood problem as a maximum hypertree problem, is the decomposition of the likelihood in terms of local contributions, or *weights*, of subsets of variables. The resulting information decomposition is not reducible to multivariate mutual information and differs from other recently introduced decompositions [2].

Future Work: Our understanding of the maximum likelihood Markov network problem is not complete. There is a large gap between the performance guarantee for the approximation algorithm and the known complexity bounds for the problem. It is unclear whether we can achieve arbitrarily small approximation ratios, independent of the target tree-width.

The proposed information decomposition in terms of weights over subsets of variables is also not fully understood. We were able to derive a necessary condition for the weights to correspond to some underlying distribution but the condition may not be sufficient.

We are currently also investigating practical algorithms for estimating low tree-width Markov networks. The algorithms, motivated by our theoretical analysis, combine a global initialization heuristic with greedy methods for building hypertrees. The theoretical understanding of the problem structure also enables to define more "global" greedy steps, but further research into special dynamic data structures is needed in order to efficiently implement them.

Research Support: The authors acknowledge support from NSF grant CCR-9624239, a Packard Foundation Fellowship, Nippon Telegraph and Telephone Corporation, ARO MURI grant DAAD19-00-1-0466 and an NIH Genome Training Grant.

References:

[1] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

[2] Shun ichi Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

[3] David Karger and Nathan Srebro. Learning Markov networks: Maximum bounded tree-width graphs. In *Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms*, 2001.

[4] Francesco M. Malvestuto. Approximating discrete probability distributions with decomposable models. *IEEE Transactions on Systems, Man and Cybernetics*, 21(5):1287–1294, 1991.

[5] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, revised second printing edition, 1997.

[6] Nathan Srebro. Maximum likelihood Markov networks: An algorithmic approach. Master's thesis, Massachusetts Institute of Technology, 2000.

[7] Nathan Srebro. Maximum likelihood bounded tree-width Markov networks. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligent*, 2001.

[8] N. Wermuth and S. Lauritzen. Graphical and recursive models of contingency tables. *Biometrika*, 72:537–552, 1983.