Message Passing and Expectation Propagation

Efficient Inference in large scale machine learning

Christoph Dehner
Department of Informatics
Technische Universität München
dehner@in.tum.de

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1. INTRODUCTION

Probabilistic graphical models like Markov Random Fields or Bayesian Networks provide clear and illustrative ways to describe probabilistic processes. In such models, nodes represent random variables and edges their conditional dependencies. The joint distribution of all involved variables can be expressed as a product of factors, which are observed or given specific values during modeling. As an example, a Bayesian network with three variables is given in figure 1. It defines the values of x_1 and x_3 to be conditionally independent given x_2.

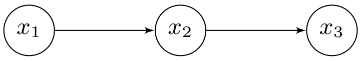

Figure 1: Bayesian network with three variables. x_1 and x_3 are conditionally independent given x_2.

Typical tasks in such a scenario are to do Bayesian inference to extract marginal distribution of specific variables or find maximum apriori estimates. However, with an increasing number of variables these task can become computationally very expensive. The naive approach of marginalizing out all but one variable from the joint distribution for example requires exponentially in the number of variables many evaluations of the joint distribution and is thus not applicable to bigger networks. Therefore, this paper explains more efficient algorithms to do large scale Bayesian inference in graphical models. The next chapter *todo* deals with exact inference and presents two message passing algorithms to calculate marginals and maximum apriori estimates for graphical models. Subsequently, in chapter *todo* expectation propagation as a method of approximate inference is explained.

2. EXACT INFERENCE

The intuition of the algorithms presented in this section is to exploit independence properties of the random variables. Graphical models define the joint probability to be expressed as a product of factors, considering the conditional independences expressed by the graph. For Bayesian networks those factors correspond to conditional distributions, in Markov Random Fields to clique potentials. Assuming appropriately normalized factors, the joint distribution can be written as a product according to equation 1. The index s iterates here over all factors of the graph; $\mathbf{x_s}$ defines the subset of variables, the factor s depends on.

$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x_s}) \tag{1}$$

Expressing the joint distribution as the product defined by the graph allows to exchange summations and multiplications during marginalization. By this way, marginalization can often be done a lot more efficient. Equation 2 demonstrates this transformation with the Bayesian network from figure 1, whose joint distribution $p(x)$ is defined as $p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1)$.

$$\begin{aligned} p(x_2) &= \sum_{x_1}\sum_{x_3} p(x_1)p(x_2|x_1)p(x_3|x_2) \\ &= \sum_{x_1}\Big[p(x_1)p(x_2|x_1)\sum_{x_3}\Big[p(x_3|x_2)\Big]\Big] \\ &= \underbrace{\Big[\sum_{x_1} p(x_1)p(x_2|x_1)\Big]}_{\mu_{x_1\to x_2}} \cdot \underbrace{\Big[\sum_{x_3} p(x_3|x_2)\Big]}_{\mu_{x_3\to x_2}} \end{aligned} \tag{2}$$

Here, the sum of products from the naive approach for marginalization is transformed to a product of sums, reducing the exponential computational complexity in the number of random variables to linear effort.

The brackets in the last line of equation 2 reveal a powerful interpretation of the marginalizing. The marginal distribution of x_2 consists of two messages $\mu_{x_1\to x_2}$ and $\mu_{x_3\to x_2}$ from its neighboring cells x_1 and x_3. For a longer chain of random variables these messages would again be comprised by messages from their neighboring cell. Applied to a chain of arbitrary length, this method gives a recursive pattern in which messages for the marginalization are sent to the marginalized variable node from both ends of the chain.

This message passing idea is the foundation for the two inference algorithms presented in this chapter subsequently. Before that the next section introduces factor graphs, the structure these algorithms operate on.

2.1 Factor graphs

Factor graphs make the factorization of a joint probability distribution explicit. The can be generated from Bayesian networks as well as from Markov random fields and thus allow to define inference algorithms independent of how the underlying probabilistic model was introduced.

Additional to the variable nodes, a factor graph also consists of factor nodes representing the factors of the decomposed joint probability distribution as in equation 1. Edges connect the factor nodes to all variables they depend on. By this way they a bipartite graph with variable nodes (usually visualized by circles) on the one side and and factors (depicted as rectangles) on the other side.

Depending on the exact factorization, a distribution defined by a Bayesian network or Markov random field can be represented by different factor graphs. Figure *todo* depicts a factor graph of the Bayesian network from figure 1, in which all conditional distribution are represented by separate factors.

The inference algorithms on factor graphs of the following sections are valid for trees. Such factor graphs with exactly one path between any pair of nodes can be generated from undirected trees in the case of a Markov random field model as well as from from directed trees and polytrees, if the factor graph is derived from a Bayesian network. More detailed description how to derive factor graph from probabilistic graph models can be found in todo: add reference.

2.2 Marginalization

messages for marginalization sum-product-algorithm

2.3 Maximum Apriori Estimate

Problemformulierung max-sum-algorithms

2.4 Inference in general graphs

loopy belief propagation

3. APPROXIMATE INFERENCE

Known as expectation propagation, similar to variational inference Minimize KL-divergence using moment matching. Interesting properties different from standard VI

3.1 Methodology

3.2 Expectation propagation in graphical models

4. SUMMARY AND OUTLOOK

4.1 Subsection

blabla with 2 sources [2; 1].

5. REFERENCES

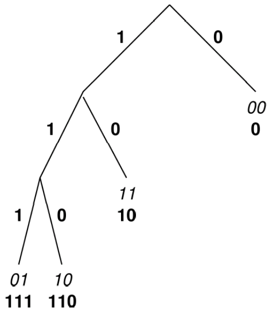

Figure 2: Tree

Table 1: Example table

Column 1	Column 2
0	1

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[2] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.