Thesis Abstract: Approximate Inference, Structure Learning and Feature Estimation in Markov Random Fields

Pradeep Ravikumar
University of California, Berkeley
pradeepr@stat.berkeley.edu

The task of prediction, estimating an output or response given an input, often involves a *statistical model* that describes the probabilistic relationship between the input and the response. An elementary way to represent such a relationship is a random field over the input and the response. When the stochastic system under consideration has many variables of interest, rather than just a single input and response, the random field in turn will have to cover all the variables characterizing the system. Undirected graphical models, or Markov random fields (MRFs), provide an important framework for representing such joint probability distributions, and are thus used in a variety of domains, including statistical physics, natural language processing, image analysis, and spatial statistics, among others.

Let $X = (X_1, \ldots, X_p)$ denote a vector of random variables, with each variable X_s taking values in a corresponding set $\mathcal{X}_s$. A Markov random field over X is then a *family* of probability distributions over X, and is specified by two objects. The first is a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is a set of p vertices so that each variable X_s is associated with a vertex $s \in V$, and $E \subset V \times V$ is the set of edges of the graph. The main advantage of graphical models is that they represent probability distributions compactly. To see the importance of compactness of representation, suppose the variables take values in sets of size k, $|\mathcal{X}_s| = k$. Then the total number of assignments of values to the random vector X is k^p, which is exponential in the graph size p. Thus, for graphical models to be useful, they must reduce this cost of representing a joint distribution. They do this by encoding in the edges E of the graph a set of conditional independence assumptions that any distribution in the family is supposed to satisfy. Termed *Markov properties*, they require that each variable be conditionally independent of its non-neighbors given its neighbors. For instance, if the graph G has no edges, $E = \phi$, then each vertex has no neighbors so that the corresponding graphical model is a family of distributions where the variables $\{X_s\}_{s \in V}$ are all independent. An important consequence of these conditional independence assumptions is provided by the Hammersley-Clifford theorem [1], which states that any positive distribution $\mathbb{P}$ over X that satisfies all the Markov properties of a graph G, also *factorizes* according to the graph. By this is meant the following. Let $\mathcal{C}$ be the set of cliques, or fully connected components, of the graph G, and let $\{\psi_C(X_C)\}_{C \in \mathcal{C}}$ be a set of clique-functions, so that each function $\psi_C(X_C)$ only depends on a subset of the variables $\{X_s,\ s \in C\}$ corresponding to a clique $C \in \mathcal{C}$. Then, a distribution $\mathbb{P}$ that factorizes according to the graph is given as,

$$\mathbb{P}(X) \quad \propto \quad \prod_{C \in \mathcal{C}} \psi_C(X_C).$$

Note that since these clique functions depend only on a subset of the variables, they provide huge savings in the cost of representing probability distributions. We had noted that a graphical model distribution is completely specified by two objects, with the first being the graph G. This set of clique functions is the other characterizing object.

There are two main classes of tasks in the Markov random field framework. The first, naturally, is to estimate the MRF distribution from data. This in turn has two subclasses of tasks, one for each of the two objects that specify an MRF distribution. The first is to estimate the clique functions, also called *features*, from data. This subtask is thus called *feature estimation*. The second is to estimate the underlying graph $G = (V, E)$ of the MRF from data. This subtask is often called *structure learning*. Finally, given the features and the graph structure, which then completely specify an MRF distribution, we arrive at the task of *inference*, which is to answer queries about the probability distribution represented by the MRF. Some basic inference tasks are as follows.

Computing the log partition function: The partition function is the normalization constant of the graphical model distribution, and is required to compute the probabilities of assignments. Unfortunately, when the random vector X is discrete, this is typically intractable for general graphs and feature functions since it involves computing a sum over exponentially many configurations of X. This has motivated the development of "approximate inference" techniques that compute approximations to the partition function.

Event probability estimation: This is the most natural query to a random field—to compute the probability of an event involving the random variables of the graphical model. A common example is a *marginal* probability of setting a subset of nodes to a particular value, e.g. $\mathbb{P}(X_s = 1)$. This too is typically intractable under general settings, so that tractable techniques only give approximations.

Computing upper and lower bounds: Some applications might require guarantees on the approximate estimates

of the event probabilities. However, inference under general MRF settings is so hard that even computing constant factor approximations is intractable. One solution for this is to compute rigorous upper and lower bounds for the event probabilities, so that we obtain an interval in which the true event probability lies. Such an interval serves as an *additive* guarantee.

Inference given moments: This is a setting where we are given only partial information about the MRF distribution. The task is to estimate event probabilities given just the expected values (moments) of the set of feature functions.

Estimating the *MAP* configuration: Given an assignment of values to a subset of the random variables, the maximum a posteriori or MAP configuration is the most probable assignment of values to the rest of the variables. Again, estimating the MAP configuration is intractable for discrete MRFs under general settings.

In this thesis [2], we address all the tasks listed above; the five inference tasks, the structure learning task, and finally the feature estimation task. Together these greatly lighten the load on any domain expert, who is required to merely list the random variables of the system. Given data, the procedures detailed in this thesis, as well as allied procedures in the literature, can then be used to construct a graphical model, and perform efficient approximate inference on those estimated models.

Inference:

(a) To approximate the log partition function, we propose *preconditioner approximations*. The general approach of approximate inference techniques is to "project" the intractable graphical model to a space of tractable models, and perform inference with the projected—and tractable—model. The divergence characterizing this projection is typically a Kullback-Leibler divergence or its approximations. Our preconditioner approximations on the other hand use a new divergence that we call the *graphical model condition number*.

(b) To estimate general event probabilities, we propose variational Chernoff bounds and variational Chebyshev-Chernoff bounds. These involve extending the classical Chernoff and Chebyshev bound framework, which apply to independent and identically distributed random variables, to the graphical model setting. These provide not just approximate estimates, but rigorous upper and lower bounds for general event probabilities. The Chernoff bounds require the complete specification of the distribution, whereas the Chebyshev bounds require just the expected values or moments of the set of feature functions of the MRF.

(c) To compute the MAP configuration, we propose a quadratic programming (QP) relaxation. The state of the art tractable algorithms for the MAP problem were previously based on a linear programming (LP) relaxation, or its dual, instead. While counterintuitive, the QP provides huge savings in computational cost over the LP, due to an order of magnitude reduction in the number of variables. If each of the p variables have arity k, the LP then has $\mathcal{O}(|E|\,k^2)$ variables while our QP has only $\mathcal{O}(n\,k)$ variables.

Structure Learning: For structure learning, we investigate procedures based on edge-appearance parameterizations and ℓ_1-regularized regression. Typical approaches for structure learning involve searching through the combinatorial space of graphs and selecting the graph with the highest score according to some metric. For undirected graphical models, both of the steps involved, computing the score, and searching through the space of graphs is intractable. Using ℓ_1 regularization on the other hand, we transform the search over the discrete space of graphs to a real-valued convex optimization problem, which is thus tractable. The guarantee on our estimate is probabilistic; we show the procedure consistently estimates the true graph even under "high-dimensional" scaling where the number of samples could be merely logarithmic in the number of variables.

Feature Estimation: For feature estimation, we propose additive conditional random fields (aCRFs), a subclass of graphical models which allow efficient nonparametric estimation of feature functions from data given the structure, and sparse additive models (SpAM), a class of models which allow simultaneous variable selection and feature estimation from data.

Acknowledgements

The dissertation owes all to my thesis advisor John Lafferty. Important parts of the thesis were also based on collaborations with Han Liu, Martin Wainwright and Larry Wasserman. I'm also indebted to my thesis committee: John Lafferty, Carlos Guestrin and Eric Xing of Carnegie Mellon University, and Martin Wainwright of University of California, Berkeley.

1. REFERENCES

[1] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[2] P. Ravikumar. Approximate inference, structure learning and feature estimation in markov random fields. *Technical Report CMU-ML-07-115, Carnegie Mellon University*, 2007.