Computational Intelligence Seminar (8 November, 2005, Waseda University, Tokyo, Japan)

From neural networks to graphical models: a brief introduction

Toshiyuki Tanaka [1]

Graduate School of Informatics, Kyoto University,

36-1 Yoshida-honmachi, Sakyo-ku, Kyoto-shi, Kyoto 606-8501, Japan

1 Neural networks and graphical models

One of the most important and interesting aspects of the neural networks is that they exhibit collective information-processing capabilities by connecting processing elements ("neurons"), each of which performs a very simple and limited information processing. The central issue of the neural network research is thus to explore how and why such a system with simple elements exhibits complex information processing functionalities. Being looked at from another direction, by studying the problem, we are asking about the global characterization of the information processing performed by the network, on the basis of the local characterization of information processing (i.e., that performed by each neuron), as well as how they are interconnected.

Essentially the same problem also arises in the study of graphical models (Bayesian networks). The main objective of this section is to give a brief review of the graphical models, with some emphasis on their relation to neural networks. In one important class of graphical models, a system with an N-dimensional random vector $\boldsymbol{x}$ is characterized by a set of local conditional probabilities of the form: $p(x_i|\boldsymbol{x}_{\setminus i})$, where $\boldsymbol{x}_{\setminus i}$ denotes the set of all the elements of $\boldsymbol{x}$ except x_i. If, for all i, x_i is independent of x_j, $j > i$, then the set of the local conditional probabilities defines a global probability distribution

$$p(\boldsymbol{x}) = \prod_{i=1}^{N} p(x_i|x_1, \ldots, x_{i-1}). \tag{1}$$

If, on the other hand, the above independence condition does not hold,

$$\rho(\boldsymbol{x}) = \prod_{i=1}^{N} p(x_i|\boldsymbol{x}_{\setminus i}) \tag{2}$$

does not necessarily define a global probability distribution, but with the proper normalization coefficient $Z_0 = \sum_{\boldsymbol{x}} \rho(\boldsymbol{x})$, one can define a global probability distribution by $p(\boldsymbol{x}) = \rho(\boldsymbol{x})/Z_0$. It should be noted that such a model can be regarded as a generalization of the conventional neural networks with stochastic neurons: In fact, if one assumes that the "state" x_i of the "neuron" i depends on $\boldsymbol{x}_{\setminus i}$ only through their

[1] E-mail: tt@i.kyoto-u.ac.jp

weighted sum $\boldsymbol{w}_i \cdot \boldsymbol{x}_{\setminus i}$, with $\boldsymbol{w}_i$ the weight vector of the neuron i, then the system defined by (2) reduces to a stochastic neural network.

A further generalization is possible by introducing clique functions $p(\boldsymbol{x}_r)$, where a clique r is a subset of the index set $\{1, \ldots, N\}$. We let

$$\rho(\boldsymbol{x}) = \prod_{r \in C} p(\boldsymbol{x}_r), \tag{3}$$

where the product is taken over a set of cliques C. Again, $\rho(\boldsymbol{x})$ does not necessarily define a global probability distribution, but it can be normalized to give a global probability distribution, as $p(\boldsymbol{x}) = \rho(\boldsymbol{x})/Z_0$, where

$$Z_0 = \sum_{\boldsymbol{x}} \rho(\boldsymbol{x}). \tag{4}$$

As a demonstrative example, we would like to note that (3) defines a Boltzmann machine [1] when the elements of $\boldsymbol{x}$ are binary and when one considers "two-body" cliques $r = (i, j) \subset \{1, \ldots, N\}$. See also [2] for an early example of graphical models that appeared in the neural network literature.

2 Expectations, likelihoods and statistical mechanics

With a model such as those described as (2) or (3), it is often the case that one has to deal with the following problems. Let $\boldsymbol{x} = (\boldsymbol{x}_V, \boldsymbol{x}_H)$, where $\boldsymbol{x}_V$ and $\boldsymbol{x}_H$ denote "visible" and "hidden" variables, respectively. Assume that one has an observation $\boldsymbol{a}$ for the visible variables $\boldsymbol{x}_V$. Then, for example, one is interested in evaluating

$$\sum_{\boldsymbol{x}_H} x_i \, p(\boldsymbol{x}_H | \boldsymbol{x}_V = \boldsymbol{a}) = \frac{\sum_{\boldsymbol{x}_H} x_i \, \rho(\boldsymbol{x}_H, \, \boldsymbol{x}_V = \boldsymbol{a})}{Z}, \tag{5}$$

that is, expectation of $x_i \in \boldsymbol{x}_H$ conditioned on the observation $\boldsymbol{x}_V = \boldsymbol{a}$. Here we let

$$Z = \sum_{\boldsymbol{x}_H} \rho(\boldsymbol{x}_H, \, \boldsymbol{x}_V = \boldsymbol{a}). \tag{6}$$

Such quantities are important when one wants to perform inference on the basis of the observation $\boldsymbol{a}$ and the probability model (2) or (3) at hand. Or, one would like to evaluate

$$p(\boldsymbol{x}_V = \boldsymbol{a}) = \sum_{\boldsymbol{x}_H} p(\boldsymbol{x}_H, \, \boldsymbol{x}_V = \boldsymbol{a}) = \frac{Z}{Z_0}, \tag{7}$$

that is, the likelihood of the observation, which measures "goodness" of the model (2) or (3) in explaining the observation $\boldsymbol{a}$, and therefore is useful in learning as well as in model selection. Evaluation of conditional expectations and likelihoods is often computationally hard. It is indeed the case when the elements of $\boldsymbol{x}_H$ and $\boldsymbol{x}$ are discrete, in which case one generally requires computation that scales exponentially in the dimensions of $\boldsymbol{x}_H$ and of $\boldsymbol{x}$ to evaluate Z and Z_0, respectively.

One can notice formal similarity between the formulation just introduced and those found in statistical mechanics. The quantities Z_0 and Z can be regarded as partition functions of properly defined physical

systems. This is the basic observation behind the idea of applying statistical-mechanics notions and tools to evaluate expectations and likelihoods. If no clique function becomes 0, then one can let

$$\rho(\boldsymbol{x}) = \exp\Big[\sum_{r\in C} c_r(\boldsymbol{x}_r)\Big], \tag{8}$$

which is exactly the Gibbs-Boltzmann distribution with "Hamiltonian" $-\sum_{r\in C} c_r(\boldsymbol{x}_r)$.

3 Advanced mean field methods

3.1 Naive mean field approximation

In the following, we assume for simplicity that the ranges of the elements of $\boldsymbol{x}$ are subsets of $\mathbb{R}$, the set of real values. In statistical mechanics, various approximation frameworks to evaluate expectations and likelihoods have been studied. The simplest one is the naive mean field approximation, in which one considers a "test" distribution of the form

$$p_0(\boldsymbol{x};\, \boldsymbol{\theta}) = \exp\Big[\sum_i \theta_i x_i - \psi_0\Big], \tag{9}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ is the parameter of p_0, and where ψ_0 is the logarithm of the normalization coefficient of p_0, which is identified with the free energy in statistical mechanics. One then tries to find the parameter value that gives the "best" approximation to the true distribution $p(\boldsymbol{x})$. In order to obtain the best approximation we minimize the Kullback-Leibler divergence

$$D(p_0\|p) = \sum_{\boldsymbol{x}} p_0(\boldsymbol{x};\, \boldsymbol{\theta}) \log \frac{p_0(\boldsymbol{x};\, \boldsymbol{\theta})}{p(\boldsymbol{x})} = \boldsymbol{\theta}\cdot\boldsymbol{\eta} - \psi_0 - \sum_{r\in C} \langle c_r(\boldsymbol{x}_r)\rangle_0 + \psi, \tag{10}$$

where ψ is the free energy of the true distribution p, where $\langle\cdot\rangle_0$ denotes expectation with p_0, and where $\boldsymbol{\eta} = \langle \boldsymbol{x}\rangle_0$. The stationarity condition gives the so-called mean field equations

$$\theta_i = \sum_{r\in C} \frac{\partial \langle c_r(\boldsymbol{x}_r)\rangle_0}{\partial \eta_i}. \tag{11}$$

With a solution of the mean field equations, one can evaluate expectations with p_0, which hopefully will give a good approximation to the expectations with p. The approximate free energy ψ_0 can also be evaluated from the solution, which is expected to give a good approximation to the true free energy ψ.

An information-geometrical interpretation of the naive mean field approximation is given in [3].

3.2 Belief propagation

In the naive mean field approximation we have considered a single test distribution. Alternatively, we can consider many of them, anticipating that it improves approximation. Specifically, we let

$$p_r(\boldsymbol{x};\, \boldsymbol{\zeta}_r) = \exp\Big[\sum_i \zeta_{ri} x_i + c_r(\boldsymbol{x}_r) - \psi_r\Big], \quad r \in C. \tag{12}$$

We then consider the following algorithm:

1. Initialize $\{\boldsymbol{\xi}_r = (\xi_{r1}, \ldots, \xi_{rN}),\, r \in C\}$.

2. Iterate the following:

   (a) For $r \in C$, let $\boldsymbol{\zeta}_r = \sum_{s \in C \setminus r} \boldsymbol{\xi}_s$.

   (b) Compute the naive mean field approximation of $p_r(\boldsymbol{x};\, \boldsymbol{\zeta}_r)$, to obtain $p_0(\boldsymbol{x};\, \boldsymbol{\theta}_r)$.

   (c) Calculate the "extrinsic information" $\boldsymbol{\xi}_r = \boldsymbol{\theta}_r - \boldsymbol{\zeta}_r$.

The algorithm defined above is the belief propagation [4]. Belief propagation gives the correct answer when the graphical representation [4], induced by the cliques, of the true distribution $p(\boldsymbol{x})$ is a tree. The algorithm is still applicable to cases where the graphical representation has loops, but in such cases we lose theoretical guarantee that it converges to the true solution. Furthermore, it may not converge at all. Nevertheless, the algorithm has been studied extensively, because of its importance in coding theory [5, 6], as well as its relation with the Bethe approximation in statistical mechanics [7, 8].

If the algorithm arrives at an equilibrium, $\boldsymbol{\theta}_r$ becomes independent of r. Let the fixed-point value of $\boldsymbol{\theta}_r$ be $\boldsymbol{\theta}^*$. Then the following properties hold.

- $\boldsymbol{\theta}^*$ is the minimizer of the Kullback-Leibler divergence

$$D(p_0 \| p_r) = \sum_{\boldsymbol{x}} p_0(\boldsymbol{x};\, \boldsymbol{\theta}) \log \frac{p_0(\boldsymbol{x};\, \boldsymbol{\theta})}{p_r(\boldsymbol{x};\, \boldsymbol{\zeta}_r^*)} \tag{13}$$

  for fixed $\boldsymbol{\zeta}_r^*$.

- The true distribution $p(\boldsymbol{x})$ and the test distributions $p_0(\boldsymbol{x};\, \boldsymbol{\theta}^*)$, $\{p_r(\boldsymbol{x};\, \boldsymbol{\zeta}_r^*),\, r \in C\}$ are in a log-linear relation, that is,

$$\log p(\boldsymbol{x}) - \log p_0(\boldsymbol{x};\, \boldsymbol{\theta}^*) = \sum_{r \in C} \bigl[\log p_r(\boldsymbol{x};\, \boldsymbol{\zeta}_r^*) - \log p_0(\boldsymbol{x};\, \boldsymbol{\theta}^*)\bigr] + \text{const.} \tag{14}$$

These properties have an elegant information-geometrical interpretation, which is discussed in [9].

In the context of statistical mechanics, "two-body" cliques are usually considered, but one can use more general cliques, for example, spanning trees of the model [10].

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, vol. 9, pp. 147–169, 1985.

[2] G. E. Hinton and T. J. Sejnowski, "Optimal perceptual inference," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 448–453, 1983.

[3] T. Tanaka, "Information geometry of mean field theory," *IEICE Trans. Fundamentals*, vol. E79-A, no. 5, pp. 709–715, 1996.

[4] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan-Kaufmann, 1988.

[5] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm," *IEEE J. Select. Areas Commun.*, vol. 16, no. 2, pp. 140–152, 1998.

[6] D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Info. Theory*, vol. 45, no. 2, pp. 399–431, 1999; Errata, *ibid.*, vol. 47, no. 5, p. 2101, 2001.

[7] Y. Kabashima and D. Saad, "Belief propagation *vs.* TAP for decoding corrupted messages," *Europhys. Lett.*, vol. 44, no. 5, pp. 668–674, 1998.

[8] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Info. Theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[9] S. Ikeda, T. Tanaka, and S. Amari, "Stochastic reasoning, free energy, and information geometry," *Neural Computation*, vol. 16, no. 9, pp. 1779–1810, 2004.

[10] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," *IEEE Trans. Info. Theory*, vol. 49, no. 5, pp. 1120–1146, 2003.