Out-Tree Dependent Nonparametric Bayesian Inference

Tony Jebara JEBARA@CS.COLUMBIA.EDU
Department of Computer Science, Columbia University

Abstract

A variational Bayesian estimator is proposed that integrates over parameters of a conditional model of a datum given its parent as well as all parent-to-child out-tree connectivity structures. The approach yields Bayesian inference in a nonparametric setting.

1. Introduction

Many paths can upgrade parametric Bayesian inference into nonparametric settings. For instance, Dirichlet processes (Teh et al., 2004; Neal, 2003; Ferguson, 1973) and infinite mixture models (Rasmussen, 1999; Beal et al., 2002) hypothesize latent variables that implicitly group data into clusters which locally obey some stationary parametric form. Instead of hypothesizing a latent clustering, consider a latent tree connectivity between data points where points are sampled from a stationary conditional distribution *given* their parents. This forms a flexible density estimator for manifold structured and hierarchical data (Jebara, 2008). We derive a variational Bayes estimator by integrating over parameters of the conditional between parent and child as well as the latent out-tree sampling structure. As in most nonparametric Bayesian methods, exchangeability is maintained. However, the out-tree yields a richer estimator as each datum is sampled with dependence on an unknown parent allowing potentially complex interactions between points.

2. Extending *iid* sampling to *tdid*

A key quantity in Bayesian inference and density estimation is the *evidence*. Given samples $\mathcal{X} = \{X_1, \ldots, X_T\}$ from a distribution $p(X|\theta)$ with unknown parameters θ, the evidence is the integral $p(\mathcal{X}) = \int_\theta p(\mathcal{X}|\theta)p(\theta)$. In density estimation, a standard assumption is that $X_1, \ldots, X_T$ are sampled *iid*. In other words, the likelihood factorizes as $p(\mathcal{X}|\theta) = \prod_{t=1}^{T} p(X_t|\theta)$. If we assume *iid* sampling, and we assume $p(X_t|\theta)$ lies in the exponential family and we assume a conjugate prior for $p(\theta)$, Bayesian integrals are straightforward (Box & Tiao, 1992). Nonparametric Bayesian inference can be obtained by choosing $p(X_t|\theta)$ outside of the exponential family (i.e. mixtures). This article instead produces nonparametric Bayesian inference by considering non-*iid* assumptions. A more general assumption than *iid* is *tdid* or tree dependent identically distributed sampling (Jebara, 2008). Therein, we assume that data was sampled according to a latent out-tree (West, 1996; Kemp et al., 2003) structure [1] which connects pairs of data points with a parent-child dependence. Sampling *tdid* proceeds as follows. From a uniform prior distribution over out-trees $p(\mathcal{T})$, an undirected tree $\mathcal{T}$ is chosen to connect T nodes. Then, choose a root node from the set of nodes. Form an out-tree by choosing all edges to point away from the root. Then, from a prior over root models $p(\theta_m)$ sample a root model θ_m and then sample the attributes of the root X_r from the model $p(X_r|\theta_m)$. From a prior over conditional mutation models $p_c(\theta_c)$, sample a conditional model θ_c. Then, traversing from parent to child along the out-tree, sample each child's attribute vector X_t according to a conditional (mutation) distribution that depends on its parent $X_{\pi(t)}$ given by $p(X_t|X_{\pi(t)}, \theta_c)$. This sampling structure strictly generalizes the *iid* setting which emerges if the dependence of each datum on its parent is extinguished and $p(X_t|X_{\pi(t)}, \theta_c) \rightarrow p(X_t|\theta_m)$. We will assume that only the observations $\mathcal{X}$ are available for inference and that both the parameters $\theta = \{\theta_m, \theta_c\}$ and the out-tree structure $\mathcal{T}$ are hidden. If we knew $\mathcal{T}$ and θ, the likelihood factorizes as:

$$p(\mathcal{X}|\mathcal{T}, \theta) \quad = \quad \prod_{t=1}^{T} p(X_t|X_{\pi(t)}, \theta).$$

Since $\mathcal{T}$ and θ are unknown, the evidence requires integrating over both. For now, assume that the parameters are known. Also assume a uniform prior over out-trees $p(\mathcal{T}) = \frac{1}{T^{T-1}}$. The *tdid* latent likelihood is given by summing over all out-trees:

$$p(\mathcal{X}|\theta) \quad = \quad \frac{1}{T^{T-1}} \sum_{\mathcal{T}} \prod_{t=1}^{T} p(X_t|X_{\pi(t)}, \theta).$$

[1] An out-tree is an acyclic graph $\mathcal{T}$ with T vertices $\mathcal{X} = X_1, \ldots, X_T$ and directed edges such that each node X_t has one parent node $X_{\pi(t)}$ and the root has zero parents. All directed edges point away from the root.

Since there are T^{T-1} out-trees, this is unwieldy. Instead, break the summation into all possible choices of the root of the out-tree $r = 1 \dots T$ and a summation over the subset $\mathcal{T}_r$ of all T^{T-2} out-trees rooted at node r. The latent *tdid* likelihood simplifies into:

$$p(\mathcal{X}|\theta) \quad = \quad \frac{1}{T^{T-1}} \sum_{r=1}^{T} p(X_r|\theta_m) Z_r$$

where we have used the property that the root has no parent node and defined the following as the contribution of each out-tree rooted at r:

$$Z_r \quad = \quad \sum_{\mathcal{T}_r} \prod_{t \neq r}^{T} p(X_t|X_{\pi(t)}, \theta_c) \; = \; \sum_{\mathcal{T}_r} \prod_{uv \in \mathcal{T}} \beta_{uv}.$$

Above, we also wrote the Z_r term as a product of edges in the out-tree instead of a product of nodes. That formula involves an asymmetric β weight matrix of size $T \times T$ populated by all pairwise conditional probabilities $\beta_{uv} = p(X_u|X_v, \theta_c)$ and where $\beta_{vv} = 0$. Cleverly, Tutte's *Directed Matrix Tree Theorem* (West, 1996) recovers Z_r in cubic time using the determinant:

$$Z_r \quad = \quad \left| [\mathrm{diag}(\beta \mathbf{1}) - \beta]_r \right|.$$

Here, we take $\vec{1}$ to be the ones column vector and the $\mathrm{diag}(\vec{v})$ operator gives a diagonal matrix with $\vec{v}$ on its diagonal. Also, we denote by $[Q]_r$ the matrix cofactor obtained by deleting the r'th row and r'th column of the matrix Q. Reinserting the formula for Z_r gives the likelihood $p(\mathcal{X}|\theta)$ which is now efficient to evaluate. Naively, this requires $\mathcal{O}(T^4)$ total work however Woodbury's formula produces a solution in $\mathcal{O}(T^3)$ (Jebara, 2008). Next, we tackle joint integration over both structure and parameters (Friedman & Koller, 2003; Attias, 1999; Mau et al., 1999).

3. Variational Bayes for *tdid* sampling

The log-evidence $\mathcal{E} = \ln p(\mathcal{X})$ is the integral over *both* parameters and structures. We assume the root and conditional distributions are in the exponential family and the priors on their parameters $p(\theta) = p(\theta_m)p_c(\theta_c)$ are conjugate. Integrating with $p(\mathcal{T}) = \frac{1}{T^{T-1}}$ yields:

$$\mathcal{E} \quad = \quad \ln \int_\theta \sum_{r=1}^{T} p(X_r) \sum_{\mathcal{T}_r} \prod_{t \neq r}^{T} p(X_t|X_{\pi(t)}) \frac{p(\theta)}{T^{T-1}}.$$

However, $\mathcal{E}$ intractable, so we instead manipulate a lower bound on the evidence. This is done by introducing variational distributions, for instance, the distribution $q(r)$ over choices for the root. We also introduce variational distributions over out-trees rooted at each r which we denote $q_r(\mathcal{T}_r)$ and a variational distribution over the parameters $q_c(\theta_c)$. Applying Jensen's inequality produces:

$$\begin{aligned} \mathcal{E} \quad \geq \quad & \sum_r q(r) \ln \int_{\theta_m} p(X_r|\theta_m) p(\theta_m) \\ & + \sum_{r, \mathcal{T}_r} q(r) q_r(\mathcal{T}_r) \sum_{t \neq r}^{T} \int_{\theta_c} q_c(\theta_c) \ln p(X_t|X_{\pi(t)}, \theta_c) \\ & + H(q) - (T-1) \ln T + \sum_r q(r) H(q_r) - D(q_c \| p_c) \end{aligned}$$

Above, H denotes the Shannon entropy and D denotes the Kullback-Leibler divergence. Update rules for each variational distribution iteratively maximize the lower bound by taking derivatives and setting to zero. We update the density over out-trees rooted at node r via:

$$q_r(\mathcal{T}_r) \quad = \quad \frac{1}{Z_r} \prod_{t \neq r}^{T} e^{\int_{\theta_c} q_c(\theta_c) \ln p(X_t|X_{\pi(t)}, \theta_c)}.$$

As in the previous section, this can be rewritten as a product over edges in the out-tree $\mathcal{T}_r$ and summarized simply by a $T \times T$ matrix β whose off diagonal entries are $\beta_{uv} = \int_{\theta_c} q_c(\theta_c) \ln p(X_u|X_v, \theta_c)$. For exponential family $p(X_u|X_v, \theta_c)$, such integrals are easy to solve. Each Z_r is also straightforward to recover using Tutte's theorem. The update for the $q(r)$ distribution is:

$$\begin{aligned} q(r) \quad \propto \quad & e^{H(q_r)} \int_{\theta_m} p(X_r|\theta_m) p(\theta_m) \\ & \times e^{\sum_{\mathcal{T}_r} q_r(\mathcal{T}_r) \sum_{t \neq r}^{T} \int_{\theta_c} q_c(\theta_c) \ln p(X_t|X_{\pi(t)}, \theta_c)} \end{aligned}$$

where the entropy $H(q_r)$ and the expectation over $q_r(\mathcal{T}_r)$ are efficient to compute from the β matrix (Meila & Jaakkola, 2006). Furthermore, the integrals $\int_{\theta_m} p(X_r|\theta_m)p(\theta_m)$ are known for exponential families. We update the distribution over parameters via:

$$\begin{aligned} q_c(\theta_c) \quad & \propto \quad p(\theta_c) e^{\sum_{r, \mathcal{T}_r} q(r) q_r(\mathcal{T}_r) \sum_{t \neq r}^{T} \ln p(X_t|X_{\pi(t)}, \theta_c)}. \\ & \propto \quad p(\theta_c) \prod_{u \neq v} p(X_u|X_v, \theta_c)^{\sum_{r, \mathcal{T}_r} q(r) q_r(\mathcal{T}_r) \delta(uv \in \mathcal{T}_r)}. \end{aligned}$$

This is simply the prior times a product over all pairs of datapoints likelihoods with different weights $q_c(\theta_c) \propto p(\theta_c) \prod_{u \neq v} p(X_u|X_v, \theta_c)^{W_{uv}}$. The weights $W_{uv} = \sum_{r, \mathcal{T}_r} q(r) q_r(\mathcal{T}_r) \delta(uv \in \mathcal{T}_r)$ are recovered easily from the current β matrix.

Thus, a nonparametric variational Bayesian treatment is possible over joint tree structure and parameters. The variational method allows us to refine the lower bound on evidence and permits nonparametric modeling with only $\mathcal{O}(T^2)$ storage and $\mathcal{O}(T^3)$ computation since out-tree distributions can be manipulated using linear algebra on the asymmetric β matrix.

References

Attias, H. (1999). A variational Bayesian framework for graphical models. *NIPS*.

Beal, M., Ghahramani, Z., & Rasmussen, C. (2002). The infinite hidden Markov model. *NIPS*.

Box, G., & Tiao, G. (1992). *Bayesian inference in statistical analysis*. John Wiley & Sons.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209–230.

Friedman, N., & Koller, D. (2003). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, *2*, 95–125.

Jebara, T. (2008). Bayesian out-trees. *Uncertainty in Artifical Intelligence*.

Jebara, T., & Long, P. (2005). *Tree dependent identically distributed learning* (Technical Report CUCS-040-05). Columbia University, Computer Science.

Kemp, C. amd Griffiths, T., Stromsten, S., & Tenenbaum, J. (2003). Semi-supervised learning with trees. *NIPS*.

Mau, B., Newton, M., & Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, *55*, 1–12.

Meila, M., & Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, *16*, 77–92.

Neal, R. (2003). Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, *7*, 619–629.

Rasmussen, C. (1999). The infinite Gaussian mixture model. *NIPS*.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). Hierarchical Dirichlet processes. *NIPS*.

West, D. (1996). *Introduction to graph theory*. Prentice Hall.