

CS281A/Stat241A recitation 13: Variational inference, Gibbs sampling, factorial HMM

November 12, 2007

Percy Liang

Inference and parameter estimation

- Broad goal: represent the unknown variables in the model given the observed variables
- Unknown quantities include latent variables \mathbf{z} and parameters $\boldsymbol{\theta}$
- Running example: factorial HMM
 - Time slices $t = 1, \dots, T$
 - Each latent state takes on values $z_{t,j} \in \{1, \dots, K\}$
 - A priori independent chains $j = 1, \dots, J$
 - Parameters: $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\phi})$
 - * Transition matrix for each slice: $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_J)$
 - * Emission matrix: $\boldsymbol{\phi}$
 - There are many options for this. If \mathbf{x} is continuous, a natural choice is to let \mathbf{x} the sum of many Gaussian variables. If \mathbf{x} is discrete, \mathbf{x} could be drawn from one of K^J multinomials or given by noisy-or.
 - Hyperparameters for specifying prior on parameters: $\boldsymbol{\alpha}, \boldsymbol{\beta}$ (assumed to be fixed)
 - Model:

$$\boldsymbol{\pi}_{j,k} \sim \text{Dirichlet}(\cdot; \boldsymbol{\alpha})$$

$$\boldsymbol{\phi}_{j,k} \sim G(\cdot; \boldsymbol{\beta})$$

$$z_{t,j} \sim \text{Multinomial}(\cdot; \boldsymbol{\pi}_{j,z_{t-1,j}}) = \boldsymbol{\pi}(j, z_{t-1,j}, z_{t,j})$$

$$x_{t,j} \sim F(\cdot; \mathbf{z}_t, \boldsymbol{\phi})$$

This description completely specifies the joint model

$$p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{x}) = p(\boldsymbol{\pi})p(\boldsymbol{\phi})p(\mathbf{z} | \boldsymbol{\pi})p(\mathbf{x} | \mathbf{z}, \boldsymbol{\phi})$$

- Important to remember that conditioned on the data \mathbf{x} , hidden chains \mathbf{z} are not independent
- Approximation schemes
 - Full Bayesian inference: $p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{x})$

- Variational: find best approximation to posterior $\operatorname{argmin}_q \text{KL}(q(\mathbf{z}, \boldsymbol{\theta}) || p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{x}))$
 - * Viterbi EM: \mathbf{z}^* and $\boldsymbol{\theta}^*$
 - * EM: $q(\mathbf{z}) = p(\mathbf{z} | \mathbf{x}, \boldsymbol{\theta}^*)$, $\boldsymbol{\theta}^*$
 - * Variational EM: $q(\mathbf{z}) = \prod_i q(z_i)$, $\boldsymbol{\theta}^*$
 - * Variational Bayes: $q(\mathbf{z}) = \prod q(z_i)$, $q(\boldsymbol{\theta}) = \prod q(\theta_i)$
- Sampling: collect samples of $(\mathbf{z}, \boldsymbol{\theta})$
 - * Gibbs sampling: update one variable at a time
 - * Blocked Gibbs sampling: update many variables at a time

Approach to approximate inference

- General setup: we have a distribution over unknown variables $\mathbf{y} = (\mathbf{z}, \boldsymbol{\theta})$

$$p(\mathbf{y}) = \frac{1}{Z} \prod_c \psi_c(y_c)$$

- What we can do: evaluate $p(\mathbf{y})$ up to a constant, which is good enough to do many things:
 - Comparing two configurations $p(\mathbf{y})/p(\mathbf{y}')$ (useful for Metropolis-Hastings)
 - Computing conditionals over a (small) tractable subset of variables \mathbf{y}_A , for example $p(\mathbf{y}_A | \mathbf{y}_{-A})$
 - * A can be a single variable, as needed for Gibbs sampling
 - * A could represent a chain, as needed for the E-step of EM
- General approach:
 - Have our representation of the unknown variables
 - Use $p(\mathbf{y}_A | \mathbf{y}_{-A})$ to update the representation of \mathbf{y}_A

Factorial HMM

- Gibbs sampling
 - We want to write down the conditional distributions of a variable given all others.
 - It's proportional to the product of the ingoing edges (prior) and the outgoing edges (likelihood).

$$p(z_{t,j} | \dots) \propto \text{Multinomial}(z_{t,j}; \pi_{j,z_{t-1,j}}) G(x_t; \mathbf{z}_t, \boldsymbol{\phi})$$

$$p(\pi_{j,k} | \dots) \propto \text{Dirichlet}(\pi_{j,k}; \boldsymbol{\alpha}) \prod_{t=1}^T \text{Multinomial}(z_{t,j}; \pi_{j,k})^{\mathbb{1}_{[z_{t,j}=k]}}$$

- Blocked Gibbs sampling

$$p(z_{1,j}, \dots, z_{T,j} \mid \dots) = \prod_{t=1}^T p(z_{t,j} \mid z_{t-1,j}, \boldsymbol{\pi}) p(x_t \mid z_{t,1}, \dots, z_{t,J}, \boldsymbol{\phi})$$

- The posterior distribution over such a chain can be computed using the forward-backward (sum-product) algorithm. Given pairwise marginals, one can sample the full chain in one sampling step.
- The distribution over the parameters is the same as in regular Gibbs sampling.

- Mean-field

- Assume we have $q(\mathbf{z}, \boldsymbol{\theta}) = \prod_{j=1}^J \left(\prod_{k=1}^K (q(\boldsymbol{\pi}_{j,k}) q(\boldsymbol{\phi}_{j,k})) \prod_{t=1}^T q(z_{t,j}) \right)$

$$q^*(\pi_{j,k}) \propto \text{Dirichlet}(\pi_{j,k}; \boldsymbol{\alpha}) \prod_{t=1}^T \text{Multinomial}(z_{t,j}; \pi_{j,k})^{q(z_{t,j}=k)} = \text{Dirichlet}(\pi_{j,k}; \boldsymbol{\alpha} + \sum_{t=1}^T q(z_{t,j} = \cdot))$$

- Structured mean-field

- Assume we have $q(\mathbf{z}, \boldsymbol{\theta}) = \prod_{j=1}^J \left(\prod_{k=1}^K (q(\boldsymbol{\pi}_{j,k}) q(\boldsymbol{\phi}_{j,k})) \prod_{t=1}^T q(z_{t,j}) \right)$

- This is the analogue of the blocked Gibbs sampling move described earlier.

- Use general update for structured mean-field:

$$q_A^*(\mathbf{y}_A) \propto \exp\{\mathbb{E}_{q_{-A}(\mathbf{y}_{-A})} \log p(\mathbf{y}_A \mid \mathbf{y}_{-A})\}$$

- If \mathbf{y}_A is a set of discrete variables parameterized by multinomials with Dirichlet priors (as in the fHMM case for $\mathbf{y}_A = (z_{1,j}, \dots, z_{T,j})$), then the optimal $q_A^*(\mathbf{y}_A)$ can be written

$$p(z_{1,j}, \dots, z_{T,j} \mid \dots) = \prod_{t=1}^T W(z_{t,j} \mid z_{t-1,j}, \boldsymbol{\pi}) W(x_t \mid z_{t,1}, \dots, z_{t,J}, \boldsymbol{\phi}),$$

where each weight W is the result of applying $\exp \mathbb{E} \log$ to the corresponding factor p .

- Therefore, inferring the best q_A^* for discrete variables involves running the same sum-product inference as in blocked Gibbs sampling, but only using W instead of p .
- As a self-contained example, suppose that we want to compute the corresponding $W(z \mid \theta)$ of $p(z \mid \theta) = \theta_z$, where $q(\theta) = \text{Dir}(\boldsymbol{\alpha})$. Then we have

$$W(z \mid \theta) = \exp \mathbb{E}_{\theta \sim \text{Dir}(\boldsymbol{\alpha})} \theta_z = \frac{\exp \Psi(\alpha_z)}{\exp \Psi(\sum_{z'} \alpha_{z'})},$$

which is based on a result on problem 1(c) of homework 3.