CS281: Advanced ML **November 1, 2017**

Lecture 16: Variational Inference Part 2

Lecturer: Sasha Rush *Scribes: Denis Turcu,Xu Si,Jiayu Yao*

16.1 Announcements

- T4 out, due 9/13 at 5pm
- Exams available in office
- OH - today 2:30-4pm (Wednesdays)
- Follow formatting for the abstract of the final project. There were many inconsistencies with the formatting requirements for the initial proposal.

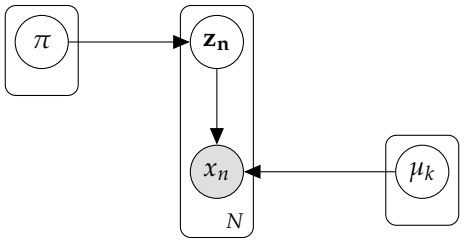
16.2 Introduction

Last class, we talked about variational inference. This class, we gonna talk about a very different type of VI. We also will talk about some other types of VI but will not go too much into the details.

Murphy's book, especially Chapter 22, covers many details on the theory side. The other text, Murphy referred as "The Monster", we put online as a textbook written by Michael Jordan.

16.3 Bayesian GMM

We are going to talk more about variational inference. We also put another reference online called VI: A Review for Statisticians. It covers in great detail of Bayesian GMM, so let's write down that model:



We assume:

$$\mu_k \sim \mathcal{N}(0, \sigma^2) \; \forall \; k$$

$$z_n \sim Cat(\frac{1}{k}, \ldots, \frac{1}{k}) \; \forall \; k$$

$$x_n | z_n, \mu \sim \mathcal{N}(\mu_{z_n}, 1) \; \forall \; n.$$

Then we write:

$$p(\{x_n\}, \{z_n\}, \mu) = p(\mu) \prod_n p(z_n) p(x_n | z_n, \mu).$$

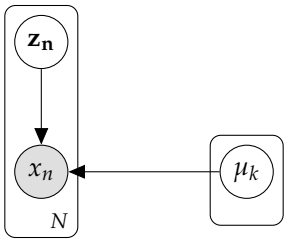
And we get:

$$p(x) = \int_{z,\mu} p(x, z, \mu) = \int p(\mu) \prod_n \sum_{z_n} p(z_n) p(x_n | z_n, \mu) \mathrm{d}\mu.$$

Variation setup. Goal:

$$\min_{q \in EASY} KL(q||p)$$

reverse KL.
We pick $EASY$ as mean field.



p

q

Variational parametrization:

$$q(\mu, z) = \prod_k q_k(\mu_k) \prod_n q_n(z_n).$$

$$q_n(z_n; \lambda_n^z) \qquad Cat(\lambda_n^z).$$

$$q_k(\mu_k; \lambda_k^\mu, \lambda_k^{\sigma^2}) \qquad \mathcal{N}(\lambda_k^\mu, \lambda_k^{\sigma^2}).$$

$$\arg\min_{q \in EASY} KL(q||p) = \arg\min_{\lambda} KL\left(\prod_k q_k(\mu_k; \lambda_k^\mu, \lambda_k^{\sigma^2}) \prod_n q_n(z_n; \lambda_n^z) \;||\; p\right)$$

"When we do *mean field*?"

$$q_i \sim \exp[\mathbb{E}_{-q_i} \log(p(z, x))]$$

Brief interlude: coordinate ascent $\rightarrow$ CAVI (coordinates ascent variational inference). Doing each individual one at a time.

- Bound we are optimizing is non-convex.
- This method is monotonically increasing.
- Sensitive to initialization $\rightarrow$ common for random restarts.

Example, deriving the math for GMM can be useful to understand how it works and how we will do mean field updates. Start from the above, setup the problem:

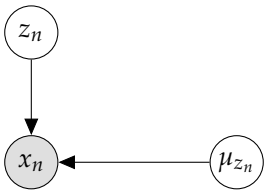
$$\mu_k \sim \mathcal{N}(0, \sigma^2) \; \forall \; k$$

$$z_n \sim Cat(\frac{1}{k}, \ldots, \frac{1}{k}) \; \forall \; k$$

$$x_n|z_n, \mu \sim \mathcal{N}(\mu_{z_n}, 1) \; \forall \; n$$

where we also have λ_n^z for hidden switch variable, and $\lambda_k^m, \lambda_k^{s^2}$ for the Gaussians.

$$\begin{aligned} q_n(z_n; \lambda_n^z) &\propto exp[\mathbb{E}_{-q_n} \log(p(\mu, z, x))] \\ &\propto exp[\mathbb{E}_{-q_n} \log(p(x_n|z_n, \mu_{z_n}))] \\ &\propto exp[\mathbb{E}_{-q_n} - (x_n - \mu_{z_n})^2/z] \\ &\propto exp[\mathbb{E}_{-q_n}(x_n\mu_{z_n} - {\mu_{z_n}}^2/2)] \\ &\propto exp[x_n\mathbb{E}_{-q_n}(\mu_{z_n}) - \mathbb{E}_{-q_n}({\mu_{z_n}}^2)/2] \end{aligned}$$



So we identify $\mathbb{E}_{-q_n}(\mu_{z_n})$ with $\lambda^m_{k=z_n}$ and $\mathbb{E}_{-q_n}({\mu_{z_n}}^2)$ with $\lambda_k^{s^2}$, and then we can write:

$$\begin{aligned} q_k(\mu_k; \lambda^m_{k=z_n}, \lambda_k^{s^2}) &\propto exp[\mathbb{E}_{-q_n}(\log p(\mu_k) + \sum_n \log p(x_n|z_n, \mu)] \\ &= -{\mu_k}^2/(2\sigma^2) + \sum_n \mathbb{E}(\log p(x_n|z_n, \mu)) \\ &= -{\mu_k}^2/(2\sigma^2) + \sum_n \mathbb{E}(z_{nk}(\log p(x_n|\mu_k)) \\ &= -{\mu_k}^2/(2\sigma^2) + \sum_n \mathbb{E}_{-q_n}(z_{nk})(\log p(x_n|\mu_k)) \\ &= -{\mu_k}^2/(2\sigma^2) + \sum_n \lambda^z_{nk}(-(x_n - \mu_{z_n}))^2/2 + const. \\ &= (\sum_k \lambda^z_{nk} x_n)\mu_k - (\frac{\sigma^2}{2} + \sum \lambda^z_{nk}/2){\mu_k}^2 + const. \end{aligned}$$

Then:

$$q_k(\mu_k) = exp[\theta^T \phi - A + \dots],$$

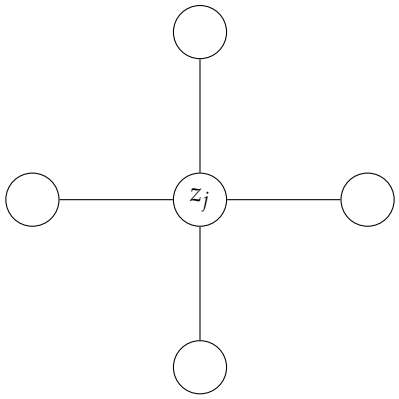
where $\theta_1 = \sum_n \lambda^z_{nk} x_n, \; \theta_2 = -(\frac{\sigma^2}{2} + \sum \lambda^z_{nk}/2)$ and $\phi_1 = \mu_k, \; \phi_2 = {\mu_k}^2$, as in GLM.
For normal distribution, we have:

$$\lambda_k^m = \frac{\sum_n \lambda^z_{nk} x_n}{1/\sigma^2 + \sum_n \lambda^z_{nk}}, \text{ and } \lambda_k^{s^2} = \frac{1}{1/\sigma^2 + \sum_n \lambda^z_{nk}}.$$

16.4 Exponential Family

$$p(z_j|z_{-j}, x) = h(z_j) exp(\theta^T \phi(z_j) - A(\theta)),$$

where θ are function of z_{-j}, x. One nice case is *UGM*:



blanket

$$q(z) = \prod_j q(z_j)$$

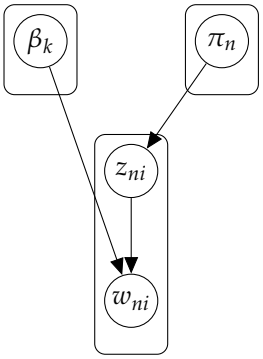
$$\begin{aligned} q(z) &\propto exp[\mathbb{E}_{-q_j} \log p(z_j | z_{-j}, x)] \\ &\propto exp[\log(h) + \mathbb{E}(\theta)_{z_j}^T - \mathbb{E}(A(\theta))] \\ &\propto h(z_j) exp[\mathbb{E}(\theta)^T z_j] \end{aligned}$$

where $\mathbb{E}(\theta)^T$ are the natural parameters of the variational approximation

$$\lambda_j = \mathbb{E}[\theta(z_{-j}, x)]$$

16.5 Latent Dirichlet Allocation

- Widely used generative latent variable model
- generative model set up



  where $\beta_k \sim Dir(\eta)$, $\pi_n \sim Dir(\alpha)$, $z_{ni} \sim Cat(\pi_n)$, $w_{ni} \sim Cat(\beta_{z_{ni}})$

- Topic molding story:
  - n - documents
  - i - words

- π_n - document topic distribution
- β_k - topic-word distribution
- z_{ni} - topic selected for word i of document n
- w_{ni} - word selected for ni.
- λ_{ni}^{z} - probability for the topic of word i in document n

16.6 Demo

We did an example iPython notebook (`TopicModeling.ipynb`) in class.