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.

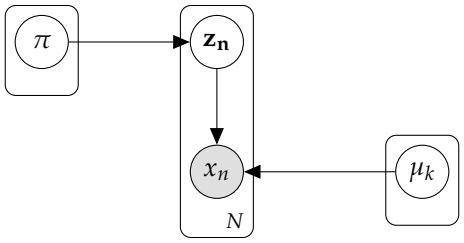
16.2 Introduction

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 [BKM17]. 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:

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

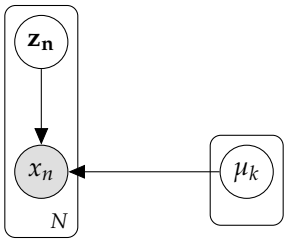
And we get:

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

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:

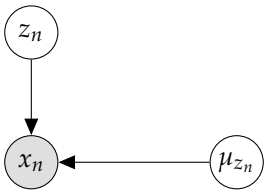
$$\mu_k \sim \mathcal{N}(0, \sigma^2) \; \forall \; k$$

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

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

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:

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

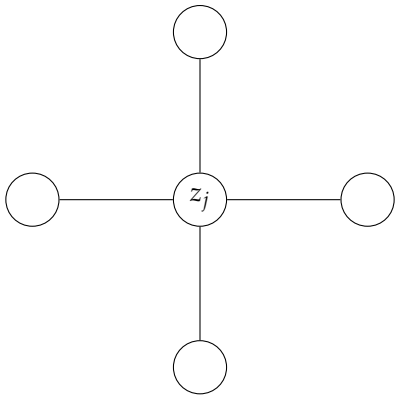
where $\theta_1 = \sum_n \lambda^z_{nk} x_n, \; \theta_2 = -(\frac{\sigma^2}{2} + \sum \lambda^z_{nk}/2)$ and $\phi_1 = \mu_k, \; \phi_2 = {\mu_k}^2$, as in GLM.
For normal distribution, we have:

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

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

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

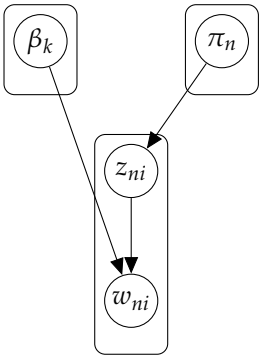
$$\begin{aligned} q(z) &\propto exp[\mathbb{E}_{-q_j} \log p(z_j | z_{-j}, x)] \\ &\propto exp[\log(h) + \mathbb{E}(\theta)_{z_j}^T - \mathbb{E}(A(\theta))] \\ &\propto h(z_j) exp[\mathbb{E}(\theta)^T z_j] \end{aligned}$$

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^z_{ni} - probability for the topic of word i in document n

16.6 Demo

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

16.7 Exercise: Variational Bayes EM

In lecture we discussed variational inference for Gaussian mixture models.[1] As covered by the notes "Variational Inference: A Review for Statisticians" [BKM17], we have the following formal description of the update steps, where the notation is as follows:

- z_i are the latent cluster labels with prior $\text{Cat}(\frac{1}{K}, \ldots, \frac{1}{K})$.
- y_i are the data.
- μ_k are the latent cluster means with prior $\mathcal{N}(0, \sigma^2)$.
- $\widetilde{\mu}, \widetilde{\sigma}^2$ are variational parameters for $q(\mu_k) \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}^2_k)$.
- $\widetilde{\alpha}$ are variational parameters for $q(z_i) \sim \text{Cat}(\widetilde{\alpha}_i)$.

Algorithm 1 CAVI for a Gaussian mixture model

Input: Data $y_1, \ldots, y_n$, number of components K, prior variance of component means σ^2
Output: Variational densities $q(\mu_k) \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}^2_k)$ and $q(z_i) \sim \text{Cat}(\widetilde{\alpha}_i)$

while the ELBO has not converged **do**
  for $i \in \{1, \ldots, n\}$ **do**
    Set $\widetilde{\alpha}_{ik} \propto \exp\{\widetilde{\mu}_k y_i - \frac{1}{2}(\widetilde{\mu}^2_k + \widetilde{\sigma}^2_k)\}$
  for $k \in \{1, \ldots, K\}$ **do**
    Set $\widetilde{\mu}_k \leftarrow \dfrac{\sum_i \widetilde{\alpha}_{ik} y_i}{1/\sigma^2 + \sum_i \widetilde{\alpha}_{ik}}$
    Set $\widetilde{\sigma}^2_k \leftarrow \dfrac{1}{1/\sigma^2 + \sum_i \widetilde{\alpha}_{ik}}$
  Compute ELBO$(\widetilde{\mu}, \widetilde{\sigma}^2, \widetilde{\alpha})$
Return $q(\widetilde{\mu}, \widetilde{\sigma}^2, \widetilde{\alpha})$

In fact, this algorithm can be seen as a special case of the *variational Bayes EM* algorithm. This algorithm is really only tractable for conjugate-exponential models, and we shall state the result for that case below. The result is found in [Beal03].

Theorem 1 (Variational Bayes EM for Conjugate-Exponential Models). Suppose we have a model with parameters μ giving rise to an i.i.d. data set $y = \{y_1, \ldots, y_n\}$ with corresponding hidden variables $z = \{z_1, \ldots, z_n\}$. Suppose further that the complete-data likelihood is in the exponential family:

$$p(y_i, z_i \mid \mu) \propto f(y_i, z_i) \exp\left[\theta(\mu)^\top u(y_i, z_i)\right],$$

[1] This section and exercise written by Jeffrey Cai.

and that the parameter prior is conjugate to the complete-data likelihood:

$$p(\mu \mid \eta, \nu) \propto g(\mu)^{\eta} \exp\left[\theta(\mu)^{\top} \nu\right]$$

where η, ν are hyperparameters, and "conjugate" means that the posterior

$$p(\mu \mid \eta', \nu') \propto p(\mu \mid \eta, \nu) p(y, z \mid \mu)$$

has the same parametric form as the prior.
Suppose our variational family consists of factored distributions

$$q(z, \mu) = q(\mu) \cdot \prod_i q(z_i).$$

Then iterating VBE and VBE steps, given below, causes the ELBO to converge to a local maximum.
VBE Step: Given $q(\mu)$ from the previous epoch, define $q(z)$ for the VBM step by:

$$q(z_i) \propto f(y_i, z_i) \exp\left[\overline{\theta}^{\top} u(y_i, z_i)\right]$$
$$\text{where } \overline{\theta} = E_{q(\mu)}[\theta(\mu)].$$

VBM Step: Given $q(z)$ from the VBE step, define $q(\mu)$ for the next epoch by:

$$q(\mu) \propto g(\mu)^{\widetilde{\eta}} \exp\left[\theta(\mu)^{\top} \widetilde{\nu}\right]$$
$$\text{where } \widetilde{\eta} = \eta + n$$
$$\widetilde{\nu} = \nu + \sum_{i=1}^{n} \overline{u}(y_i)$$
$$\text{where } \overline{u}(y_i) = E_{q(z_i)}[u(y_i, z_i)].$$

Exercise 16.1. Here you will show the claim that the "CAVI for a Gaussian mixture model" algorithm is in fact a special case of VBEM for conjugate-exponential models.
(a) Express the complete data likelihood $p(y_i, z_i \mid \mu)$ for the Gaussian mixture model in exponential family form. That is, give the natural parameters $\theta(\mu)$, the sufficient statistics $u(y_i, z_i)$, and the function $f(y_i, z_i)$. *Hint:* There should be $2K + 1$ sufficient statistics.
(b) Express the parameter prior in exponential family form. (Note that the natural parameters $\theta(\mu)$ should be the same as those you derived in (a).)
(c) Assume that from the previous epoch, $q(\mu_k) \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2)$. Show that after applying the VBE updates in Theorem 1, one obtains $q(z_i) \sim \text{Cat}(\widetilde{\alpha}_i)$ where the resulting $\widetilde{\alpha}_i$ is the same as the Algorithm 1 update.
(d) Assume that from the VBE step, $q(z_i) \sim \text{Cat}(\widetilde{\alpha}_i)$. Show that after applying the VBM updates in Theorem 1, one obtains $q(\mu_k) \sim \mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2)$ where the resulting $\widetilde{\mu}_k, \widetilde{\sigma}_k^2$ is the same as the Algorithm 1 update.

Solution. (a) We have,

$$\begin{aligned} p(y_i, z_i \mid \mu) &= \frac{1}{K} \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(y_i - \mu_{z_i})^2\right] \\ &= \frac{1}{\sqrt{2\pi} K} \exp\left[-\frac{1}{2} y_i^2 + \sum_{k=1}^{K} \mathbf{1}_{z_i = k} \left(y_i \mu_k - \frac{1}{2}\mu_k^2\right)\right]. \end{aligned}$$

From this we deduce:

$$\theta(\mu) = \begin{bmatrix} (-\mu_k^2/2)_{k=1}^K \\ (\mu_k)_{k=1}^K \\ -1/2 \end{bmatrix} \qquad u(x_i, y_i) = \begin{bmatrix} (\mathbf{1}_{z_i = k})_{k=1}^K \\ (\mathbf{1}_{z_i = k}\, y_i)_{k=1}^K \\ y_i^2 \end{bmatrix} \qquad f(x_i, y_i) = 1.$$

(b) We have,

$$p(\theta) = \prod_k p(\theta_k) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^K \exp\left[\sum_{k=1}^K \left(-\frac{\theta_k^2}{2\sigma^2}\right)\right].$$

Consequently,

$$\nu = \begin{bmatrix} (1/\sigma^2)_{k=1}^K \\ (0)_{k=1}^K \\ 0 \end{bmatrix}.$$

(c) We have,

$$\overline{\theta} = E_{q(\mu)}[\theta(\mu)] = \begin{bmatrix} (-E_{q(\mu)}[\mu_k^2]/2)_{k=1}^K \\ (E_{q(\mu)}[\mu_k])_{k=1}^K \\ -1/2 \end{bmatrix} = \begin{bmatrix} ((\widetilde{\mu}_k^2 + \widetilde{\sigma}_k^2)/2)_{k=1}^K \\ (\widetilde{\mu}_k)_{k=1}^K \\ -1/2 \end{bmatrix}$$

and therefore

$$\begin{aligned} q(z_i) &\propto \exp\left[-\frac{1}{2}\sum_{k=1}^K \mathbf{1}_{z_i=k}(\widetilde{\mu}_k^2 + \widetilde{\sigma}_k^2) + \sum_{k=1}^K \mathbf{1}_{z_i=k}\, y_i \widetilde{\mu}_k - \frac{1}{2}y_i^2\right] \\ &= \exp\left[-\frac{1}{2}(\widetilde{\mu}_{z_i}^2 + \widetilde{\sigma}_{z_i}^2) + y_{z_i}\widetilde{\mu}_{z_i} + \text{const.}\right] \end{aligned}$$

which is the desired form.
(d) We have,

$$\overline{u}(y_i) = \begin{bmatrix} (\widetilde{\alpha}_{ik})_{k=1}^K \\ (\widetilde{\alpha}_{ik} y_i)_{k=1}^K \\ y_i^2 \end{bmatrix}$$

and therefore

$$\overline{v} = \nu + \sum_{i=1}^n \overline{u}(y_i) = \begin{bmatrix} (\frac{1}{\sigma^2} + \sum_{i=1}^N \widetilde{\alpha}_{ik})_{k=1}^K \\ (\sum_{i=1}^N \widetilde{\alpha}_{ik} y_i)_{k=1}^K \\ y_i^2 \end{bmatrix}$$

and then

$$\begin{aligned} q(\mu) &\propto \exp\left[\sum_{k=1}^K \left(-\frac{1}{2}\mu_k^2\left(\frac{1}{\sigma^2} + \sum_{i=1}^N \widetilde{\alpha}_{ik}\right) + \mu_k \sum_{i=1}^N \widetilde{\alpha}_{ik} y_i\right) - \frac{1}{2}y_i^2\right] \\ &= \exp\left[\sum_{k=1}^K -\frac{1}{2\widetilde{\sigma}_k^2}(\mu_k^2 - 2\widetilde{\mu}_k\mu_k) + \text{const.}\right] \\ &= \exp\left[\sum_{k=1}^K -\frac{(\mu_k - \widetilde{\mu}_k)^2}{2\widetilde{\sigma}_k^2} + \text{const.}\right] \end{aligned}$$

where $\widetilde{\sigma}_k^2 = (\sigma^{-2} + \sum_i \widetilde{\alpha}_{ik})^{-1}$ and $\widetilde{\mu}_k = \widetilde{\sigma}_k^2 \sum_i \widetilde{\alpha}_{ik} y_i$ are the desired parameter updates, and in the last step we completed the square by pulling a missing term from the constant. Hence $q(\mu)$ has the desired form. □

References

[Beal03] Beal, M. Variational methods for approximate Bayesian inference. PhD thesis, University of London, 2003.

[BKM17] Blei, M., Kucukelbir, A., and McAuliffe, J. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518), 2017.