DD2434/FDD3434 Machine Learning, Advanced Course
Module 3 Exercise

November 2023

Contents

3 Stochastic Variational Inference – Exercises

3.1 Exponential family

Show that the following distributions are in the exponential family:
a) Beta-distribution
b) Binomial-distribution
c) Dirichlet-distribution
d) Multinomial-distribution
e) LogNormal-distribution

3.1.1 Solution

Exponential family

Wednesday, 15 November 2023 14:12

A distribution $p(x|\theta)$ is in the exponential family

if $p(x|\theta) = h(x) \exp\{\eta(\theta) t(x) - A(\theta)\}$

c) Dirichlet-distribution

$$p(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} = \exp\left\{\log \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}\right\}$$

$$= \exp\left\{-\log B(\alpha) + \sum_{i=1}^{K} (\alpha_i - 1) \log x_i\right\} = \exp\left\{\sum_{i=1}^{K} \alpha_i \log x_i - \underline{\sum_{i=1}^{K}} \log x_i - \log B(\alpha)\right\}$$

$$= \underbrace{\frac{1}{\prod_{i=1}^{K} x_i}}_{h(x)} \cdot \exp\left\{\underbrace{\sum_{i=1}^{K} \alpha_i \log x_i}_{= [\alpha_1, \dots, \alpha_K] \cdot \begin{bmatrix} \log x_1 \\ \vdots \\ \log x_K \end{bmatrix}} - \underbrace{\log B(\alpha)}_{A(\alpha)}\right\} = h(x) \exp\{\eta(\alpha) t(x) - A(\alpha)\}$$

$[\alpha_1, \dots, \alpha_K] \to \eta(\alpha)$, $\begin{bmatrix} \log x_1 \\ \vdots \\ \log x_K \end{bmatrix} \to t(x)$

<u>Note</u> $\eta(\alpha) = \alpha$ above. This is called the natural parameter form. Below is another solution which also shows that $p(x|\alpha)$ is in the exponential family, but not on natural parameter form.

Alternative 2 : $p(x|\alpha) = \exp\{\underbrace{-\log B(\alpha)}_{A(\alpha)} + \underbrace{\sum_{i=1}^{K} (\alpha_i - 1)\log x_i}\}$

$$\underbrace{[\alpha_1 - 1, \alpha_2 - 1, \ldots \alpha_K - 1]}_{\eta(\alpha)} \cdot \begin{bmatrix} \log x_1 \\ \vdots \\ \log x_K \end{bmatrix}$$

$$\eta(\alpha) = \alpha - 1$$

3.2 Metrics for distributions

Let $p_1 = p(X|\mu_1, \sigma_1^2) = Normal(\mu_1, \sigma_1^2)$ and $p_2 = p(Y|\mu_1, \sigma_1^2) = Normal(\mu_2, \sigma_2^2)$. Write a function that calculates the euclidean distance, d, between the pdf's and another function that calculates symmetric KL-divergence, D_{KL}^{sym}, between the pdf's.
Visualize p_1 and p_2 for the cases:
1) $\mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = \sigma_2^2 = 0.01$
2) $\mu_1 = 0, \mu_2 = 10, \sigma_1^2 = \sigma_2^2 = 10^4$
Calculate and compare $d((\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2))$ and $D_{KL}^{sym}((\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2))$ for each scenario. Which metric do you think best distinguishes p_1 and p_2?

3.2.1 Solution

See notebook in Canvas.

3.3 Mixture Model with Bernoulli observations - SVI

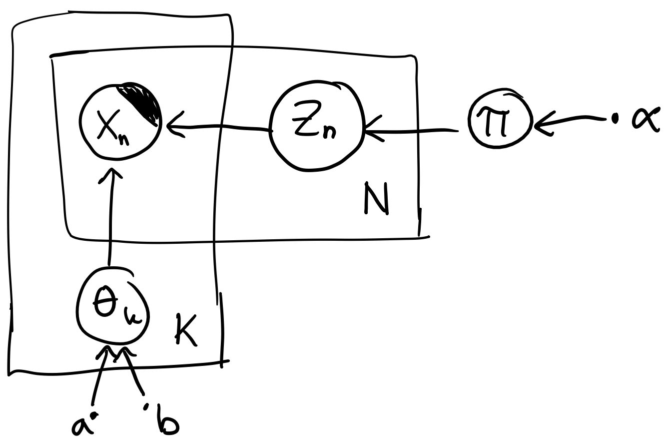
Consider the Bernoulli Mixture model (MM) of module 3.
a) Which are the local and global latent variables of the model?
b) Derive the full conditionals of the model.
c) Show that these distributions are in the exponential family.
d) Assume the mean-field factorization: $q(Z, \pi, \theta) = q(\pi) \prod_n q(Z_n) \prod_k q(\theta_k)$ and that each variational distribution is in the same parametric family as the associated complete conditional. Use the natural parameters from c) and equations 15 and 16 of Hoffman to derive the updates of the global and local latent variables. How do these updates compare to that of the Bernoulli-MM of Module 3?
e) Use the code from the Bernoulli-MM of exercises of module 3 and rewrite the CAVI algorithm to the SVI algorithm described in Figure 4 of Hoffman. Compare the outputs of the SVI algorithm and CAVI algorithm for different parameters. How do they compare? Compare the run time for a fixed number of iterations, large N ($N \geq 10^6$) and large D ($D \geq 10^3$).

3.3.1 Solution

Bernoulli Mixture model - Module 4

Wednesday, 15 November 2023 14:12

Consider the Bernoulli mixture model from the exercises of Module 3.



Final CAVI updates:

$$q(\theta) = \prod_{k=1}^{K} q(\theta_k), \quad q(\theta_k) = \text{Beta}(\hat{a}_k, \tilde{b}_k), \quad \begin{cases} \hat{a}_k = \sum_{n=1}^{N} \left[q(z_n = k)\, x_n^D \right] + a \\ \tilde{b}_k = \sum_{n=1}^{N} \left[q(z_n = k)\, (D - x_n^D) \right] + b \end{cases}$$

$$q(z) - \prod_{n=1}^{N} q(z_n), \quad q(z_n) = \text{Cat}(r_{n1}, \dots, r_{nK}), \quad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$$

$$\log \rho_{nk} = x_n^D \left(\psi(\tilde{a}_k) - \psi(\tilde{a}_k + \tilde{b}_k) \right) + (D - x_n) \left(\psi(\tilde{b}_k) - \psi(\tilde{b}_k + \tilde{a}_k) \right) + \psi(\tilde{\alpha}_k) - \psi(\tilde{\alpha}_0)$$

$$q(\pi) = \text{Dir}(\tilde{\alpha}), \quad \tilde{\alpha} = \sum_{n=1}^{N} \underbrace{q(z_n = k)}_{= r_{nk}} + \alpha_k$$

a) Local and global parameters:

Local z_n as $p(x_n, z_n \mid z_{-n}, \theta, \pi, a, b, \alpha) = \{$ Study graph for v-separations $\}$

Global θ, π $\qquad\qquad = p(x_n, z_n \mid \theta, \pi)$

I write the prior parameters here for clarity. They are assumed given for the rest of the solution.

b) Full conditionals:

study DGM

$$p(\theta_k \mid \theta_{-k}, X, Z, \pi) = p(\theta_k \mid X, Z) \propto$$

$$\propto p(\theta_k, X, Z) = p(X \mid Z, \theta_k)\, p(Z)\, p(\theta_k) \propto p(X \mid Z, \theta_k)\, p(\theta_k)$$

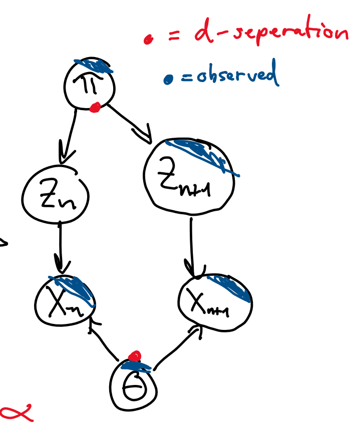
$$= \prod_{n=1}^{N} \prod_{k'=1}^{K} p(x_n \mid z_n = k', \theta_{k'})^{\mathbb{1}\{z_n = k'\}}\, p(\theta_k) \propto \prod_{n=1}^{N} p(x_n \mid z_n = k, \theta_k)^{\overbrace{\mathbb{1}\{z_n = k\}}^{\equiv z_{nk}}}\, p(\theta_k)$$

$$= \prod_{n=1}^{N} \prod_{d=1}^{D} \mathrm{Ber}(\theta_k)^{z_n^k} \cdot \mathrm{Beta}(\theta_k) \propto$$

$$\propto \prod_{n=1}^{N} \prod_{d=1}^{D} \theta_k^{x_{nd} z_{nk}} \cdot (1-\theta_k)^{(1-x_{nd}) z_{nk}} \cdot \theta_k^{a-1} \cdot (1-\theta_k)^{b-1} =$$

$$= \theta_k^{\overbrace{\sum_{n=1}^{N}\sum_{d=1}^{D} z_n^k x_{nd}}^{\equiv \gamma^k}} \cdot (1-\theta_k)^{\overbrace{\sum_{n=1,d=1}^{N,D} z_n^k}^{\phi^k} - \gamma^k} \cdot \theta_k^{a-1} \cdot (1-\theta_k)^{b-1} = \theta_k^{\gamma^k + a - 1} (1-\theta_k)^{\phi^k - \gamma^k + b - 1}$$

$$\Rightarrow \quad p(\theta_k \mid X, Z) = \mathrm{Beta}(\gamma^k + a, \phi^k - \gamma^k + b)$$



Similar for $p(\pi \mid X, Z, \theta)$ and $p(Z_n \mid Z_{-n}, X, \theta, \pi)$

see DGM

$$p(Z_n \mid Z_{-n}, X, \theta, \pi) = p(Z_n \mid X_n, \pi, \theta)$$

$$\propto p(Z_n, X_n, \pi) = p(X_n \mid Z_n, \pi)\, p(Z_n \mid \pi)\, \cancel{p(\pi)}$$

$$= \prod_{d=1}^{D} \prod_{k=1}^{K} \mathrm{Ber}(\theta_k)^{\mathbb{1}\{z_n = k\}}\, \mathrm{Cat}(\pi) = \prod_{d=1}^{D} \prod_{k=1}^{K} \left(\theta_k^{x_{nd}} (1-\theta_k)^{1-x_{nd}}\right)^{z_n^k} \prod_{k=1}^{K} \pi_k^{z_n^k}$$

$$= \prod_{k=1}^{K} \Big(\underbrace{\theta_k^{x_n^D} (1-\theta_k)^{1-x_n^D} \cdot \pi_k}_{\equiv \tilde{\rho}_{nk}} \Big)^{z_n^k} \Rightarrow$$

$$p(z_n \mid x_n, \pi, \theta) = \mathrm{Cat}\Big(\tilde{\rho}_{nk} \Big/ \sum_{k=1}^{K} \tilde{\rho}_{nk} \Big)$$

c) We want to rewrite $p(\theta_k \mid X, z_n = k)$ on the form:

$$p(\theta_k \mid X, Z, \underbrace{a, b}) = h(\theta_k) \cdot \exp\{ \eta(X, Z, a, b) \cdot t(\theta_k) - A(x_n, z_n) \}$$

include again for clarity

$$p(\theta_k \mid X, Z) = \frac{1}{B(\gamma^k + a, \phi^k - \gamma^k + b)} \cdot \theta_k^{\gamma^k + a - 1} \cdot (1-\theta_k)^{\phi^k - \gamma^k + b - 1}$$

$$= \exp\{ -\log B(\gamma^k + a, \phi^k - \gamma^k + b) \} \cdot \exp\{ (\gamma^k + a - 1) \log \theta_k + (\phi^k - \gamma^k + b - 1) \log (1-\theta_k) \}$$

$$= e^{-(\log \theta_k + \log 1 - \theta_k)} = \frac{1}{\theta_k (1-\theta_k)}$$

$$= \underbrace{\frac{1}{\theta^k (1-\theta^k)}}_{= h(\theta^k)} \exp\{ \underbrace{-\log B(\gamma^k + a, \phi^k - \gamma^k + b)}_{= A(x_n, z_n)} \} \cdot \exp\{ \underbrace{(\gamma^k + a) \log \theta_k + (\phi^k - \gamma^k + b) \log (1-\theta_k)}_{= \underbrace{[\gamma^k + a, \ \phi^k - \gamma^k + b]}_{\eta(X, Z, a, b)} \cdot \underbrace{\begin{bmatrix} \log \theta_k \\ \log 1 - \theta_k \end{bmatrix}}_{= t(\theta^k)}} \}$$

remember ϕ^k and γ^k are functions of x_n and z_n

Similar for $p(\pi \mid X, Z, \theta)$

Locals:

$$p(z_n \mid x_n, \pi, \theta) = \prod_{k=1}^{K} \left(\frac{\theta_k^{x_n^D} (1 - \theta_k^{D - x_n^D}) \pi_k}{\sum_{k=1}^{K} \tilde{\rho}_{nk}} \right)^{z_n^k} = \exp\Big\{ \sum_{k=1}^{K} \underbrace{\frac{\left(x_n^D \log \theta_k + (D - x_n^D) \log (1-\theta_k) + \log \pi_k \right)}{\sum_{k=1}^{K} \tilde{\rho}_{nk}}}_{\eta(x_n, \theta, \pi)} \underbrace{z_n^k}_{t(z_n)}$$

d) $\lambda = (a^*, b^*) = \mathbb{E}_{\theta_k, \theta_{-k}, Z, \pi}\left[\eta(x, z, a, b) \right] =$

$$= \mathbb{E}_{\theta_k, \theta_{-k}, Z, \pi}\left[(\gamma^k + a, \phi^k - \gamma^k + b) \right] = \left(\mathbb{E}_Z[\gamma^k] + a, \mathbb{E}_Z[\phi^k] - \mathbb{E}_Z[\gamma^k] + b \right)$$

$$\Rightarrow \begin{cases} a^* = \sum_{n=1}^{N} \mathbb{E}_Z\Big[z_n^k \underbrace{\sum_{d=1}^{D} x_{nd}}_{\equiv x_n^D} \Big] + a = \sum_{n=1}^{N} q(Z_n = k)\, x_n^D + a \\ b^* = \sum_{n=1}^{N} \mathbb{E}_Z\Big[\underbrace{\sum_{d=1}^{D} z_n^k - z_n^k x_{nd}}_{\substack{= z_n^k \cdot D - z_n^k x_n^D \\ = z_n^k (D - x_n^D)}} \Big] + b = \sum_{n=1}^{N} q(Z_n = k)(D - x_n^D) + b \end{cases}$$

Note that these updates are identical to the CAVI updates

Similar for our other global hidden variable π

Local variable update

$$r_{nk}^* = \mathbb{E}_{\pi,\theta,Z}\left[\eta(x_n, \theta, \pi) \right] = \mathbb{E}_{\pi,\theta,Z}\left[\frac{x_n^D \log \theta_k + (D - x_n^D) \log(1 - \theta_k) + \log \pi_k}{\sum_{k=1}^{K} \tilde{\rho}_{nk}} \right]$$

$$= \left(x_n^D\, \mathbb{E}_{\theta_k}[\log \theta_k] + (D - x_n^D)\, \mathbb{E}_{\theta_k}[\log 1 - \theta_k] + \mathbb{E}_{\pi}[\log \pi_k] \right) \Big/ \sum_{k=1}^{K} \tilde{\rho}_{nk}$$

Note that this is the same as our CAVI update for r_{nk}

e) Implementation!