Tutorial Week 5

COMP9418 – Advanced Topics in Statistical Machine Learning, 17s2, UNSW Sydney

Instructor: Edwin V. Bonilla

Last Update: Saturday 19th August, 2017 at 12:43

Following the feedback at the lecture, the exercises below are in order of priority.

1. **Variational Inference for Bayesian GMMs.** Let $\mathbf{X} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ be the observed data and $\mathbf{Z} = \{\mathbf{z}^{(n)}\}_{n=1}^{N}$ the corresponding latent variables, with each $\mathbf{x}^{(n)} \in \mathbb{R}^D$ and each $\mathbf{z}^{(n)}$ is a categorical variable encoded using one-hot-encoding.

   We can define the joint distribution of a Bayesian GMM as follows:

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}), \tag{1}$$

   where

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{k=1}^{K} \pi_k^{\alpha_k - 1}, \text{ with } \alpha_k = \alpha/K, \tag{2}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0), \tag{3}$$

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k{}^{z_k^{(n)}}, \tag{4}$$

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_k^{(n)}}. \tag{5}$$

   Assuming an approximate posterior distribution of the form:

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}), \tag{6}$$

   show that the optimal variational distribution is given by:

$$q^\star(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \tilde{r}_{nk}^{z_k^{(n)}}, \tag{7}$$

$$q^\star(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q^\star(\boldsymbol{\pi})q^\star(\boldsymbol{\mu}, \boldsymbol{\Lambda}), \tag{8}$$

$$q^\star(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\tilde{\boldsymbol{\alpha}}), \text{ with } \tilde{\alpha}_k = \tilde{r}_k + \alpha_k, \tag{9}$$

$$q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\tilde{\mathbf{m}}_k, (\tilde{\beta}_k\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\widetilde{\mathbf{W}}_k, \tilde{\nu}_k), \tag{10}$$

where

$$\tilde{r}_{nk} \propto \bar{\pi}_k \bar{\lambda}_k^{1/2} \exp\left\{ -\frac{D}{2\tilde{\beta}_k} - \frac{\tilde{\nu}_k}{2} (\mathbf{x}^{(n)} - \tilde{\mathbf{m}}_k)^T \widetilde{\mathbf{W}}_k (\mathbf{x}^{(n)} - \tilde{\mathbf{m}}_k) \right\}, \tag{11}$$

$$\tilde{\beta}_k = \beta_0 + \tilde{r}_k \tag{12}$$

$$\tilde{\mathbf{m}}_k = \frac{1}{\tilde{\beta}_k} \left(\beta_0 \mathbf{m}_0 + \tilde{r}_k \tilde{\boldsymbol{\mu}}_k \right) \tag{13}$$

$$\widetilde{\mathbf{W}}_k^{-1} = \mathbf{W}_0^{-1} + \tilde{r}_k \tilde{\boldsymbol{\Sigma}}_k + \frac{\beta_0 \tilde{r}_k}{\beta_0 + \tilde{r}_k} (\tilde{\boldsymbol{\mu}}_k - \mathbf{m}_0)(\tilde{\boldsymbol{\mu}}_k - \mathbf{m}_0)^T, \tag{14}$$

$$\tilde{\nu}_k = \nu_0 + \tilde{r}_k, \tag{15}$$

$$\log \bar{\lambda}_k = \mathbb{E}[\log |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D} \psi\left(\frac{\tilde{\nu}_k + 1 - i}{2} \right) + D \log 2 + \log \left| \widetilde{\mathbf{W}}_k \right|, \tag{16}$$

$$\log \bar{\pi}_k = \mathbb{E}[\log \pi_k] = \psi(\tilde{\alpha}_k) - \psi(\tilde{\alpha}_0), \text{ with } \tilde{\alpha}_0 = \sum_{k=1}^{K} \tilde{\alpha}_k, \tag{17}$$

where $\psi(\cdot)$ is the digamma function and the required expected sufficient statistics are given by:

$$\tilde{r}_k = \sum_{n=1}^{N} \tilde{r}_{nk}, \tag{18}$$

$$\tilde{\boldsymbol{\mu}}_k = \frac{1}{\tilde{r}_k} \sum_{n=1}^{N} \tilde{r}_{nk} \mathbf{x}^{(n)}, \text{ and} \tag{19}$$

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{1}{\tilde{r}_k} \sum_{n=1}^{N} \tilde{r}_{nk} (\mathbf{x}^{(n)} - \tilde{\boldsymbol{\mu}}_k)(\mathbf{x}^{(n)} - \tilde{\boldsymbol{\mu}}_k)^T. \tag{20}$$

Explain how a variational inference algorithm would work using the updates above.

2. **Gibbs' Inequality.** Prove that the relative entropy (or KL divergence) between two distributions $p(X)$ and $q(X)$ with $X \in \mathcal{X}$ is non-negative:

$$\mathrm{KL}(p(X) \| q(X)) \geq 0,$$

with equality if and only if $p(x) = q(x)$ for all x. HINT: Use Jensen's inequality.

3. **Mutual Information.** Show that the mutual information between X and Y is the average reduction in uncertainty of X due to the knowledge of Y, i.e. $I(X;Y) = H(X) - H(X|Y)$.

4. **Joint entropy of independent random variables.** Show that if X and Y are statistically independent discrete random variables then $H(X,Y) = H(X) + H(Y)$.

5. **Computation of Joint, Marginal and Conditional Entropies.** Consider the following joint distribution over (X,Y):

	$p(X,Y)$	X			
		1	2	3	4
Y	1	1/8	1/16	1/32	1/32
	2	1/16	1/8	1/32	1/32
	3	1/16	1/16	1/16	1/16
	4	1/4	0	0	0

Compute $H(X)$, $H(Y)$, $H(X|Y)$, $H(X,Y)$, $H(Y) - H(Y|X)$.