Tutorial Week 12

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

Instructor: Edwin V. Bonilla

Last Update: Friday 13th October, 2017 at 10:50

1. Consider the observations $\mathbf{X}$, latent variables $\mathbf{Z}$ and model parameters $\boldsymbol{\theta}$. Recall that the *Kullback-Leibler* divergence between distributions $q(\mathbf{Z}|\mathbf{X})$ and $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ is given by:

$$\mathrm{KL}\left(q(\mathbf{Z}|\mathbf{X}) \| p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})\right) \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}\left[\log \frac{q(\mathbf{Z}|\mathbf{X})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}\right] \geq 0, \tag{1}$$

where $\mathbb{E}_{p(x)}[g(x)]$ computes the expectation of $g(x)$ over $p(x)$; $q(\mathbf{Z}|\mathbf{X})$ is an approximating distribution and $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ is the true but unknown posterior distribution; and with the equality occurring iff $q(\mathbf{Z}|\mathbf{X}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$. Given the objective function:

$$\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}\left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z}|\mathbf{X})}\right], \tag{2}$$

Show that the objective used by variational inference $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$, in Equation (2) above, can be expressed as a sum of a KL (Kullback-Leibler divergence) term and a ELL (expected log likelihood) term. The KL term is the negative KL divergence between the approximate posterior $q(\mathbf{Z}|\mathbf{X})$ and the prior $p(\mathbf{Z}|\boldsymbol{\theta})$ and the ELL term is the expectation of the log conditional likelihood $\log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})$ over the approximate posterior $q(\mathbf{Z}|\mathbf{X})$.

2. Consider the supervised learning problem where we are given a dataset $\mathcal{D} = \{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^{N}$, where $\mathbf{x}^{(n)}$ is a D-dimensional input vector and $\mathbf{y}^{(n)}$ is a P-dimensional output and the goal is to learn the mapping from inputs to outputs. A possible approach to this problem is to assume that there are Q latent functions $\{f_j\}$ drawn from Q zero-mean Gaussian processes $f_j \sim \mathcal{GP}\left(0, \kappa_j(\cdot, \cdot; \boldsymbol{\theta}_j)\right)$, with $j = 1, \ldots Q$. Then our prior model is:

$$p(\mathbf{f}|\boldsymbol{\theta}) = \prod_{j=1}^{Q} p(\mathbf{f}_{\cdot j}|\boldsymbol{\theta}_j) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{f}_{\cdot j}; \mathbf{0}, \mathbf{K}_{\mathbf{xx}}^{j}), \tag{3}$$

where $\mathbf{f}$ is the set of all latent function values; $\mathbf{f}_{\cdot j} = \{f_j(\mathbf{x}_n)\}_{n=1}^{N}$ denotes the values of latent function j; $\mathbf{K}_{\mathbf{xx}}^{j}$ is the covariance matrix induced by the covariance function $\kappa_j(\cdot, \cdot; \boldsymbol{\theta}_j)$ evaluated at every pair of inputs; and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j\}$ are the covariance hyperparameters.

Along with the prior in Equation (3), we can also assume that our multi-dimensional observations $\{\mathbf{y}^{(n)}\}$ have the likelihood:

$$p(\mathbf{y}|\mathbf{f}, \boldsymbol{\phi}) = \prod_{n=1}^{N} p(\mathbf{y}^{(n)}|\mathbf{f}_{n\cdot}, \boldsymbol{\phi}), \tag{4}$$

where $\mathbf{y}$ is the set of all output observations; $\mathbf{y}^{(n)}$ is the nth output observation; $\mathbf{f}_{n\cdot} = \{f_j(\mathbf{x}^{(n)})\}_{j=1}^{Q}$ is the set of latent function values which $\mathbf{y}^{(n)}$ depends upon; and $\boldsymbol{\phi}$ are the conditional likelihood parameters.

(a) Explain the main statistical independence assumptions implied by the prior and the likelihood in Equations (3) and (4), respectively.

(b) If your problem is multi-class classification with C classes, what conditional likelihood model $p(\mathbf{y}^{(n)}|\mathbf{f}_{n\cdot}, \boldsymbol{\phi})$ would you use? what would Q and P be?

3. Now consider the prior in Equation (3) augmented with inducing variables:

$$p(\mathbf{u}) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{u}_{\cdot j}; \mathbf{0}, \mathbf{K}_{\mathbf{zz}}^{j}), \qquad p(\mathbf{f}|\mathbf{u}) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{f}_{\cdot j}; \tilde{\boldsymbol{\mu}}_j, \widetilde{\mathbf{K}}_j), \text{ where} \tag{5}$$

$$\tilde{\boldsymbol{\mu}}_j = \mathbf{K}_{\mathbf{xz}}^{j}(\mathbf{K}_{\mathbf{zz}}^{j})^{-1}\mathbf{u}_{\cdot j}, \text{ and} \tag{6}$$

$$\widetilde{\mathbf{K}}_j = \mathbf{K}_{\mathbf{xx}}^{j} - \mathbf{A}_j \mathbf{K}_{\mathbf{zx}}^{j} \text{ with } \mathbf{A}_j = \mathbf{K}_{\mathbf{xz}}^{j}(\mathbf{K}_{\mathbf{zz}}^{j})^{-1}, \tag{7}$$

and an approximate posterior:

$$q(\mathbf{f}, \mathbf{u}|\boldsymbol{\lambda}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}|\boldsymbol{\lambda}), \tag{8}$$

$$q(\mathbf{u}|\boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k q_k(\mathbf{u}|\mathbf{m}_k, \mathbf{S}_k) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{Q} \mathcal{N}(\mathbf{u}_{\cdot j}; \mathbf{m}_{kj}, \mathbf{S}_{kj}), \tag{9}$$

where $\boldsymbol{\lambda} = \{\pi_k, \mathbf{m}_{kj}, \mathbf{S}_{kj}\}$ are the variational parameters: the mixture proportions $\{\pi_k\}$, the posterior means $\{\mathbf{m}_{kj}\}$ and posterior covariances $\{\mathbf{S}_{kj}\}$ of the inducing variables corresponding to mixture component k and latent function j. We also note that that $q_k(\mathbf{u}|\mathbf{m}_k, \mathbf{S}_k)$ is a Gaussian with mean $\mathbf{m}_k$ and block-diagonal covariance $\mathbf{S}_k$.

(a) Show that the prior defined in Equations (5)–(7) is equivalent to that in Equation (3).

(b) Show that

$$\mathcal{L}_{\text{kl}}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} -\text{KL}(q(\mathbf{f}, \mathbf{u}|\boldsymbol{\lambda}) \| p(\mathbf{f}, \mathbf{u})) = -\text{KL}(q(\mathbf{u}|\boldsymbol{\lambda}) \| p(\mathbf{u})). \tag{10}$$

(c) Show that the expected likelihood term $\mathcal{L}_{\text{ell}}$ in the variational objective for this augmented model is given by:

$$\mathcal{L}_{\text{ell}}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathbb{E}_{q_{k(n)}(\mathbf{f}_{n\cdot}|\boldsymbol{\lambda}_k)} \left[\log p(\mathbf{y}^{(n)}|\mathbf{f}_{n\cdot}, \boldsymbol{\phi})\right], \tag{11}$$

where $q_{k(n)}(\mathbf{f}_{n\cdot}|\boldsymbol{\lambda}_k)$ is a Q-dimensional Gaussian with:

$$q_{k(n)}(\mathbf{f}_{n\cdot}|\boldsymbol{\lambda}_k) = \mathcal{N}(\mathbf{f}_{n\cdot}; \mathbf{b}_{k(n)}, \boldsymbol{\Sigma}_{k(n)}), \tag{12}$$

where $\boldsymbol{\Sigma}_{k(n)}$ is a *diagonal* matrix. The jth element of the mean and the (j, j)th entry of the covariance of the above distribution are given by:

$$[\mathbf{b}_{k(n)}]_j = \mathbf{a}_{jn}^T \mathbf{m}_{kj}, \qquad [\boldsymbol{\Sigma}_{k(n)}]_{j,j} = [\widetilde{\mathbf{K}}_j]_{n,n} + \mathbf{a}_{jn}^T \mathbf{S}_{kj} \mathbf{a}_{jn}, \tag{13}$$

where $\mathbf{a}_{jn} \stackrel{\text{def}}{=} [\mathbf{A}_j]_{:,n}$ denotes the M-dimensional vector corresponding to the nth column of matrix $\mathbf{A}_j$; $\widetilde{\mathbf{K}}_j$ and $\mathbf{A}_j$ are given in Equation (7); and, as before, $\{\mathbf{m}_{kj}, \mathbf{S}_{kj}\}$ are the variational parameters corresponding to the mean and covariance of the approximate posterior over the inducing variables for mixture component k and latent process j.

(d) Discuss the computational complexity of posterior estimation by optimisation of the evidence lower bound:

$$\mathcal{L}_{\text{elbo}}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \mathcal{L}_{\text{kl}}(\boldsymbol{\lambda}) + \mathcal{L}_{\text{ell}}(\boldsymbol{\lambda}). \tag{14}$$