Variational inference questions

Hugh Salimbeni

February 2016

1. In the lectures it was claimed that

$$\log \int p(X,Z)dZ = \mathbb{E}_{q(Z)} \log \frac{p(X,Z)}{q(Z)} + \mathbb{E}_{q(Z)} \log \frac{q(Z)}{p(Z|X)}$$

   Derive this result

2. You are given that $q(Z)$ factorises, so

$$q(Z) = q_1(Z_1)q_2(Z_2)...q_M(Z_M).$$

   Show that the functional derivative of the ELBO wrt one of the factors $q_i(Z_i)$ is given by

$$\frac{\delta}{\delta q_i}\text{ELBO} = \int q_1 q_2 ... q_{i-1} q_{i+1} ... q_M \log p(X|Z) dZ_1 dZ_2 ... dZ_{i-1} dZ_{i+1} ... dZ_M - \log q_i(Z_i') + \text{const.}$$

3. In this question you are going to derive the results for the bi-variate Gaussian, with a factored variational approximation.

   You are given that $\mathbf{z} \sim N\left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} | \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1} \right)$

   (a) Write down $\log p(Z)$ in full. Note that you do not need to invert the matrix to calculate the determinant. Leave your answer in terms of things like $(z_1 - \mu_1)$ etc

   (b) Write down all the terms of your previous answer that depend on z_1. This time expand all the brackets. You should have only 4 terms.

   (c) Show that

   $$\log q_1^*(z_1) = -\tfrac{1}{2} z_1^2 \Lambda_{11} + z_1 \mu_1 \Lambda_{11} - z_1 \Lambda_{12} \mathbb{E}(z_2) + z_1 \Lambda_{12} \mu_2 + \text{constant terms}$$

   (d) Complete the square in the expression above

(e) Explain why

$$q_1^* \sim N(z_1|m_1, \Lambda_{11}^{-1})$$

where

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}z_2 - \mu_2)$$

(f) Use a symmetry argument to write down the other factor. Hence write down the expectations and solve the equations to show the result given in the lecture notes.

4. (a) Write down the definition of $KL(q(Z), p(Z))$

   (b) Consider a region where $p(Z) = 0$. If you are free to choose q to minimize $KL(q(Z), p(Z))$, explain why $q(Z)$ must also equal zero in this region

   (c) Use your previous answer to explain why the variational method is referred to as "zero forcing" and why it typically underestimates the support of p

5. In this question you will derive $q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ for the variational GMM. For convenience we define the following:

$$\begin{aligned} N_k &= \sum\nolimits_{i=1}^{N} r_{ik} \\ \bar{\mathbf{x}} &= \frac{1}{N_k}\sum\nolimits_{i=1}^{N} r_{ik}\mathbf{x}_i \\ \mathbf{S}_k &= \frac{1}{N_k}\sum\nolimits_{i=1}^{N} r_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \end{aligned}$$

Note that the second two quantities are just the empirical mean and covariance, weighted by the responsibilities, and the first is just the expected number of points in each cluster.

Using the definitions for the variational GMM in slides, show that

$$q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k, \mathbf{m}_k, (\beta_k\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k)$$

where

$$\begin{aligned} &\beta_k = \beta_0 + N_k \\ &\mathbf{m}_k = \tfrac{1}{\beta_k}(\beta_0\mathbf{m}_0 + N_k\bar{\mathbf{x}}_k) \\ &\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_K}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^T \\ &\mu_k = \nu_0 + N_k + 1 \end{aligned}$$

You may like to use:

- The cyclic properties of trace: $Tr(ABC) = Tr(BCA) = Tr(CAB)$
- The fact that a scalar can be written as a trace of a 1x1 matrix
- The identity:

  $$\sum_{i=1}^{N}(a_i - b)^2 = N(\bar{a} - b)^2 + \sum_{i=1}^{N}(a_i - \bar{a})^2$$

  with $\bar{a} = \frac{1}{N}\sum a_i$
- Completing the square (for scalars)

  $$A(x-a)^2 + B(x-b)^2 = (A+B)\left(x - \frac{Aa + Bb}{A+B}\right)^2 + \frac{AB}{A+B}(a-b)^2$$

  and the equivalent expression for matricies