Tutorial Week 4

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

Instructor: Edwin V. Bonilla

Last Update: Sunday 13th August, 2017 at 21:36

1. **Expectation-Maximisation (EM) Algorithm.** Consider the *Kullback-Leibler* divergence between distributions $q(\mathbf{Z}|\mathbf{X})$ and $p(\mathbf{Z}|\mathbf{X})$:

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

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:

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

   (a) show that the this objective function is a lower bound on the log marginal likelihood, i.e. $\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$;

   (b) show the optimal setting for the approximating distribution (i.e. the distribution that maximizes the above lower bound) is: $q(\mathbf{Z}|\mathbf{X}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$;

   (c) explain how (a) and (b) justify the EM algorithm.

2. [Murphy, MLaPP, 2012] **EM for GMMs.** Show that the M-step for maximum-likelihood estimation of a mixture of Gaussians is given by:

$$\hat{\pi}_k = \frac{r_k}{N} \qquad \hat{\boldsymbol{\mu}}_k = \tfrac{1}{r_k} \textstyle\sum_{i=1}^N r_k^{(i)} \mathbf{x}^{(i)} \quad \hat{\boldsymbol{\Sigma}}_k = \frac{1}{r_k} \sum_{i=1}^N r_k^{(i)} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_k)^T, \tag{3}$$

where $r_k \stackrel{\text{def}}{=} \sum_{i=1}^N r_k^{(i)}$.

3. [Murphy, MLaPP, 2012] **Moments of a GMM**. Consider a GMM with K components,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{4}$$

   (a) Show that

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k. \tag{5}$$

   (b) Show that

$$\mathbb{C}\mathrm{ov}[\mathbf{x}] = \sum_{k=1}^K \pi_k \left(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\right) - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T. \tag{6}$$