T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 18th of April, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

1. Consider the probabilistic principal component analysis (PPCA) model

$$\mathbf{x}(t) = \mathbf{A}s(t) + \mathbf{n}(t) \tag{1}$$

with 3-dimensional data $\mathbf{x}(t)$, 1-dimensional source $s(t)$, and Gaussian noise $p(n_i(t)) = N(n_i(t) \mid 0, \sigma_i^2)$. We would like to estimate the noise level σ_i^2 for each data dimension $i = 1, 2, 3$ using the maximum likelihood estimator. What happens when the weight matrix $\mathbf{A}$ goes to $\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$, and the source $s(t)$ copies the first dimension of the data: $s(t) = x_1(t)$?

2. Variational Bayesian cost function C_{vb} is often a sum of simple terms, one for each variable. Consider the variable s whose model is $p(s \mid m, v) = N(s \mid m, \exp(-v))$. Assume that s, m, v are independent aposteriori, that is, $q(s, m, v) = q(s)q(m)q(v)$ and that $q(s) = N(s \mid \overline{s}, \widetilde{s})$. Show that the term of the cost for s, $C_{\text{vb}}(s) = E_q \left\{ \ln \frac{q(s)}{p(s|m,v)} \right\}$, is

$$\begin{aligned} C_{\text{vb}}(s) = &\frac{1}{2} \left(E_q\{\exp v\} \left[(\overline{s} - E_q\{m\})^2 + \operatorname*{Var}_q\{m\} + \widetilde{s} \right] - E_q\{v\} + \ln(2\pi) \right) \\ &- \frac{1}{2} \ln(2\pi e \widetilde{s}). \end{aligned} \tag{2}$$

3. Consider the following model:

$$p(x(t) \mid m, v) = N(x(t) \mid m, \exp(-v)) \tag{3}$$

$$p(m) = N(m \mid 0, \exp(5)) \tag{4}$$

$$p(v) = N(v \mid 0, \exp(5)), \tag{5}$$

where $x(t), t = 1, \ldots, T$ are the observed data and m and v are latent variables. Use the posterior approximation $q(m, v) = q(m)q(v) = N(m \mid \overline{m}, \widetilde{m})N(v \mid \overline{v}, \widetilde{v})$. Assuming that $q(v)$ is fixed, find $q(m)$ that minimizes $C_{\text{vb}} = E_q \left\{ \ln \frac{q(m,v)}{p(\{x(t)\}_{t=1}^T, m, v)} \right\}$.

4. Show that the EM algorithm is a special case of the VB-EM algorithm where the family of approximate distributionis $q(\boldsymbol{\theta})$ for the parameters is restricted to delta distributions (distributions where the whole probability mass is concentrated on a single point). Some assumptions have to be made: The family of approximate distribution $q(\mathbf{Z})$ for latent variables $\mathbf{Z}$ should include the true posterior $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$. KL-divergence will go to infinity, so the minimization has to be considered as a limiting process (in practice by ignoring the term $q \ln q$ that can be considered constant). Also, VB-EM usually has a prior for $\boldsymbol{\theta}$, while EM does not. Let us consider the version of EM with a prior for the parameters.

EM algorithm with a prior for parameters:

$$q(\mathbf{Z}) \leftarrow p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \tag{6}$$

$$\boldsymbol{\theta} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} E_{q(\mathbf{Z})} \{\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})\} \tag{7}$$

VB-EM algorithm:

$$q(\mathbf{Z}) \leftarrow \operatorname*{argmin}_{q(\mathbf{Z})} E_{q(\boldsymbol{\theta})} \{\mathrm{KL}\left(q(\mathbf{Z}) \parallel p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})\right)\} \tag{8}$$

$$q(\boldsymbol{\theta}) \leftarrow \operatorname*{argmin}_{q(\boldsymbol{\theta})} E_{q(\mathbf{Z})} \{\mathrm{KL}\left(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z})\right)\} \tag{9}$$

Kullback-Leibler divergence:

$$\mathrm{KL}\left(q(x) \parallel p(x)\right) = E_{q(x)} \left\{\ln \frac{q(x)}{p(x)}\right\} \tag{10}$$

$$\tag{11}$$