T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 25th of April, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

1. 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:

$$q(\mathbf{Z}) \leftarrow p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \tag{1}$$

$$\boldsymbol{\theta} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, E_{q(\mathbf{Z})} \left\{ \ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \right\} \tag{2}$$

VB-EM algorithm:

$$q(\mathbf{Z}) \leftarrow \underset{q(\mathbf{Z})}{\operatorname{argmin}}\, E_{q(\boldsymbol{\theta})} \left\{ \mathrm{KL}\left(q(\mathbf{Z}) \parallel p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \right) \right\} \tag{3}$$

$$q(\boldsymbol{\theta}) \leftarrow \underset{q(\boldsymbol{\theta})}{\operatorname{argmin}}\, E_{q(\mathbf{Z})} \left\{ \mathrm{KL}\left(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z}) \right) \right\} \tag{4}$$

Kullback-Leibler divergence:

$$\mathrm{KL}\left(q(x) \parallel p(x) \right) = E_{q(x)} \left\{ \ln \frac{q(x)}{p(x)} \right\} \tag{5}$$

$$\tag{6}$$

2. Consider two extensions of probabilistic principal component analysis (PPCA) with mixture models. The model equation $\mathbf{x}_j = \mathbf{A}\mathbf{s}_j + \boldsymbol{\epsilon}_j$ and the noise model $p(\boldsymbol{\epsilon}_j) = N(\boldsymbol{\epsilon}_j \mid \mathbf{0}, v\mathbf{I})$. The parameters are fixed to $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$, $v = 0.01$, and mixture coefficients $\pi = 0.5$ in all cases. The sources $\mathbf{s}_j$ are distributed according to the mixture of Gaussians. The two cases are: (a) The mixing coefficients π_k are shared among the two sources s_{1j} and s_{2j}

$$p(\mathbf{s}_j) = \sum_{k=1}^{2} \pi_k N(\mathbf{s}_j \mid \mu_k, \Sigma_k), \tag{7}$$

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \tag{8}$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{9}$$

(b) The mixture is done individually to the two sources s_{1j} and s_{2j}.

$$p(s_{1j}) = \sum_{k=1}^{2} \pi_{1k} N(s_{1j} \mid \mu_{1k}, \sigma_{1k}^2), \tag{10}$$

$$p(s_{2j}) = \sum_{k=1}^{2} \pi_{2k} N(s_{2j} \mid \mu_{2k}, \sigma_{2k}^2), \tag{11}$$

$$\mu_{11} = 0, \quad \mu_{12} = 0, \quad \mu_{21} = 0, \quad \mu_{22} = 0 \tag{12}$$

$$\sigma_{11} = 1, \quad \sigma_{12} = 0.3, \quad \sigma_{21} = 1, \quad \sigma_{22} = 0.3. \tag{13}$$

Sketch $p(\mathbf{x}_j \mid \mathbf{A}, v)$ in both cases.