T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 14th of March, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

The EM algorithm is useful for latent variable models, where the model defines $P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$, where $\mathbf{X}$ is the data set, $\mathbf{Z}$ are latent variables, and $\boldsymbol{\theta}$ are the model parameters. One would like find the parameters $\boldsymbol{\theta}$ that maximize the likelihood $P(\mathbf{X} \mid \boldsymbol{\theta})$, but the latent variables $\mathbf{Z}$ make the direct treatment of $P(\mathbf{X} \mid \boldsymbol{\theta})$ difficult. For example, in a mixture model, $\mathbf{Z}$ describes to which cluster each data sample belongs to, while $\boldsymbol{\theta}$ describes the general properties of the clusters. EM-algorithm solves the problem by alternating between the following two steps:

$$\text{E-step: } Q(\mathbf{Z}) \leftarrow P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \tag{1}$$

$$\text{M-step: } \boldsymbol{\theta} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} E_{Q(\mathbf{Z})} \left\{ \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \right\}, \tag{2}$$

where E_Q is the expectation over the distribution Q.

1. Given a Naïve Bayes model with three binary variables defined by the tables and data below, run an iteration of the EM algorithm.

P(C)	
C=0	0.7
C=1	0.3

$P(X_1 \mid C)$	C=0	C=1
X_1=0	0.5	0.8
X_1=1	0.5	0.2

$P(X_2 \mid C)$	C=0	C=1
$X_2 = 0$	0.6	0.3
$X_2 = 1$	0.4	0.7

Data:

t	X_{1t}	X_{2t}
1	1	1
2	0	1

(continues on the next page)

Hint: In Problem 2 of the previous exercise session, we already solved:

$$P(C_1 \mid X_{11}, X_{21}) = \begin{pmatrix} 0.769 \\ 0.231 \end{pmatrix} \tag{3}$$

$$P(C_2 \mid X_{12}, X_{22}) = \begin{pmatrix} 0.455 \\ 0.545 \end{pmatrix} \tag{4}$$

2. (a) Run k-means (page 424) until convergence in a one-dimensional problem with five data points (see table below). Use $k = 2$ and initialize with $\mu_1 = 3.5$ and $\mu_2 = 4.8$. (b) Fit a mixture-of-Gaussians (MoG, page 430) to the result by doing an M-step. MoG is a model with a cluster label C and a Gaussian distribution for the observation given the cluster label:

$$p(x \mid C = i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]. \tag{5}$$

You can fit the Gaussians by computing the mean $\mu = E(x)$ and variance $\sigma^2 = E(x^2) - E(x)^2$ of the data in each cluster. (c) Compute $P(C \mid x = 3)$.

Data:

t	x_t
1	1.0
2	2.0
3	4.0
4	5.0
5	6.0

3. Prove Equation (9.70) in the book: For any choise of $Q(\mathbf{Z})$,

$$\ln P(\mathbf{X} \mid \boldsymbol{\theta}) = \mathcal{L}(Q, \boldsymbol{\theta}) + \mathrm{KL}\,(Q \parallel P)\,, \tag{6}$$

where

$$\mathcal{L}(Q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \frac{P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{Q(\mathbf{Z})} \tag{7}$$

$$\mathrm{KL}\,(Q \parallel P) = -\sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \frac{P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})}{Q(\mathbf{Z})}. \tag{8}$$

Note that $\mathcal{L}$ is a functional because one of its arguments, Q, is a function.

4. Show that (a) the E-step (Eq. 1) maximizes $\mathcal{L}(Q, \boldsymbol{\theta})$ w.r.t. Q, and (b) the M-step (Eq. 2) maximizes $\mathcal{L}(Q, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$, and that (c) after convergence, $\mathcal{L}(Q, \boldsymbol{\theta}) = \ln P(\mathbf{X} \mid \boldsymbol{\theta})$. Hint: $\mathrm{KL}\,(Q \parallel P) \geq 0$ for all distributions Q and P.