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:

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

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

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

	t	X_{1t}	X_{2t}
Data:	1	1	1
	2	0	1

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

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

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

Solution:

The E-step was already solved in the previous exercise session: $Q(C_t) = P(C_t \mid X_{1t}, X_{2t})$. For the M-step, let us solve $\theta_1 = P(C = 1)$. The maximum of the expected log likelihood will be found at the zero of the derivative:

$$\frac{\partial}{\partial \theta_1} E_{q(C)} \{\ln L\} = \frac{\partial}{\partial \theta_1} [\ln P(C_1) + \ln P(C_2)] \tag{5}$$

$$= \frac{\partial}{\partial \theta_1} \begin{bmatrix} 0.769 \ln P(C_1 = 0) + 0.231 \ln P(C_1 = 1) \\ +0.455 \ln P(C_2 = 0) + 0.545 \ln P(C_2 = 1) \end{bmatrix} \tag{6}$$

$$= \frac{\partial}{\partial \theta_1} [(0.769 + 0.455) \ln(1 - \theta_1) + (0.231 + 0.545) \ln \theta_1] \tag{7}$$

$$= -\frac{0.769 + 0.455}{1 - \theta_1} + \frac{0.231 + 0.545}{\theta_1} = 0 \tag{8}$$

$$\theta_1 = \frac{0.231 + 0.545}{0.769 + 0.455 + 0.231 + 0.545} = 0.388 \tag{9}$$

We notice that the M-step of Bayesian networks is simply about taking the expected counts (EC) of each case happening and then normalizing them into probabilities.

C	EC(C)	P(C)
C=0	0.769+0.455	0.612
C=1	0.231+0.545	0.388

$X_1 \mid C$	EC($X_1 \mid C = 0$)	EC($X_1 \mid C = 1$)	P($X_1 \mid C = 0$)	P($X_1 \mid C = 1$)
$X_1 = 0$	0.455	0.545	0.372	0.702
$X_1 = 1$	0.769	0.231	0.628	0.298

$X_2 \mid C$	EC($X_1 \mid C = 0$)	EC($X_1 \mid C = 1$)	P($X_1 \mid C = 0$)	P($X_1 \mid C = 1$)
$X_2 = 0$	0	0	0	0
$X_2 = 1$	0.769+0.455	0.231+0.545	1	1

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:

$$p(x \mid C = i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]. \tag{10}$$

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

Solution:

(a) Samples t=1,2,3 are closer to $\mu_1 = 3.5$ and samples t=4,5 are closer to $\mu_2 = 4.8$ so $C_1 = C_2 = C_3 = 1$ and $C_4 = C_5 = 2$ for the first iteration. Next the cluster centers are set to the mean of the samples in the cluster: $\mu_1 = (1.0 + 2.0 + 4.0)/3 = 2.333$ and $\mu_2 = (5.0 + 6.0)/2 = 5.5$. Again, the samples are given to the nearest cluster, and this time $C_1 = C_2 = 1$ and $C_3 = C_4 = C_5 = 2$. Note how sample number 3 at 4.0 switched to cluster 2. The next cluster centers are $\mu_1 = 1.5$ and $\mu_2 = 5.0$. The nearest clusters of each sample do not change anymore so the algorithm has converged.

(b) The class probabilities P(C) are determined by normalizing the number of samples in the cluster into a probability, that is, $P(C = 1) = 2/5 = 0.4$ and $P(C = 2) = 3/5 = 0.6$. The variances are

$$\sigma_1^2 = \frac{1.0^2 + 2.0^2}{2} - 1.5^2 = 0.25 \tag{11}$$

$$\sigma_2^2 = \frac{4.0^2 + 5.0^2 + 6.0^2}{3} - 5.0^2 = 0.667 \tag{12}$$

The curves are shown in Figure 1.

(c) From Bayes theorem

$$P(C = i \mid x = 3.0) = \frac{p(x = 3.0 \mid C = i)P(C = i)}{p(x = 3.0)}, \tag{13}$$

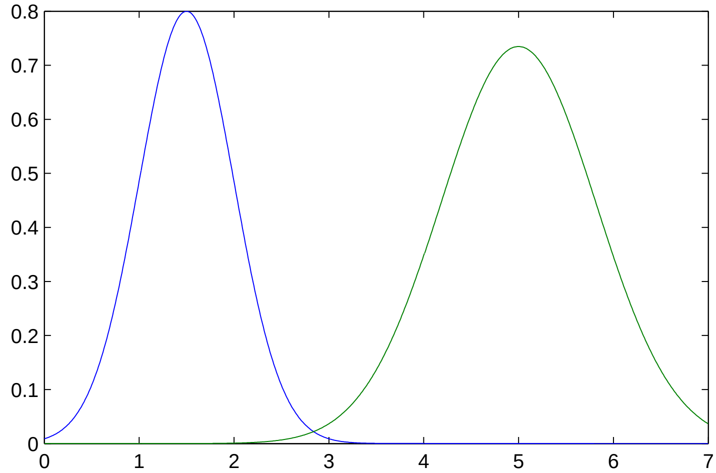

Figure 1: Problem 2. $p(x, C)$ as a function of x. The two curves correspond to $C = 1$ and $C = 2$.

where $p(x = 3.0)$ is constant w.r.t. C and can be ignored. By inserting the definition of the Gaussian density and dropping out constants, we get

$$P(C = 1 \mid x = 3.0) \propto \frac{1}{\sqrt{0.25}} \exp\left[\frac{-(3.0-1.5)^2}{2 \cdot 0.25}\right] \cdot 0.4 = 0.0089 \tag{14}$$

$$P(C = 2 \mid x = 3.0) \propto \frac{1}{\sqrt{0.667}} \exp\left[\frac{-(3.0-5.0)^2}{2 \cdot 0.667}\right] \cdot 0.6 = 0.0366 \tag{15}$$

and after normalization we get $P(C \mid x = 3.0) = \begin{pmatrix} 0.195 \\ 0.805 \end{pmatrix}$ which can be compared to the figure. Even though 3.0 is further away from the cluster center 2, it is more likely to belong to that.

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

$$\ln P(\mathbf{X} \mid \boldsymbol{\theta}) = \mathcal{L}(Q, \boldsymbol{\theta}) + \mathrm{KL}\left(Q \parallel P\right), \tag{16}$$

where

$$\mathcal{L}(Q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \frac{P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{Q(\mathbf{Z})} \tag{17}$$

$$\mathrm{KL}\left(Q \parallel P\right) = -\sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \frac{P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})}{Q(\mathbf{Z})}. \tag{18}$$

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

Solution:

$$\mathcal{L}(Q,\boldsymbol{\theta}) + \mathrm{KL}\left(Q \parallel P\right) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \left[\frac{P(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})}{Q(\mathbf{Z})} \frac{Q(\mathbf{Z})}{P(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})}\right] \tag{19}$$

$$= \sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \left[\frac{P(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})}{P(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})}\right] \tag{20}$$

$$= \sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \left[\frac{P(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})P(\mathbf{X} \mid \boldsymbol{\theta})}{P(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})}\right] \tag{21}$$

$$= \sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln P(\mathbf{X} \mid \boldsymbol{\theta}) \tag{22}$$

$$= \ln P(\mathbf{X} \mid \boldsymbol{\theta}) \sum_{\mathbf{Z}} Q(\mathbf{Z}) \tag{23}$$

$$= \ln P(\mathbf{X} \mid \boldsymbol{\theta}) \tag{24}$$

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: $\mathrm{KL}\left(Q \parallel P\right) \geq 0$ for all distributions Q and P (proof on page 55 of Bishop).

Solution:

(a)

$$\mathcal{L}(Q,\boldsymbol{\theta}) = \ln P(\mathbf{X} \mid \boldsymbol{\theta}) - \mathrm{KL}\left(Q \parallel P\right) \tag{25}$$

$\ln P(\mathbf{X} \mid \boldsymbol{\theta})$ is a constant w.r.t. Q and $\mathrm{KL}\left(Q \parallel P\right) \geq 0$ so $\mathcal{L}(Q,\boldsymbol{\theta})$ is maximized when $\mathrm{KL}\left(Q \parallel P\right) = 0$. This happens when we set $Q(\mathbf{Z}) = P(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})$ as can be seen from

$$\mathrm{KL}\left(Q \parallel Q\right) = -\sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \frac{Q(\mathbf{Z})}{Q(\mathbf{Z})} = -\sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln 1 = 0. \tag{26}$$

(b)

$$\mathcal{L}(Q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln \frac{P(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})}{Q(\mathbf{Z})} \tag{27}$$

$$= \left[\sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln P(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})\right] - \left[\sum_{\mathbf{Z}} Q(\mathbf{Z}) \ln Q(\mathbf{Z})\right] \tag{28}$$

$$= E_{Q(\mathbf{Z})}\left\{\ln P(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})\right\} - \text{const} \tag{29}$$

We got the same form that is used in the M-step.

(c) After convergence, the E and M-steps do not change anything, or:

$$Q(\mathbf{Z}) = P(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta}) \tag{30}$$

$$\boldsymbol{\theta} = \operatorname*{argmax}_{\boldsymbol{\theta}} E_{Q(\mathbf{Z})}\left\{\ln P(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})\right\}. \tag{31}$$

In (a), we already showed that $\mathrm{KL}\,(Q \parallel P) = 0$ when $Q(\mathbf{Z}) = P(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})$, and thus $\mathcal{L}(Q,\boldsymbol{\theta}) = \ln P(\mathbf{X} \mid \boldsymbol{\theta}) - \mathrm{KL}\,(Q \parallel P) = \ln P(\mathbf{X} \mid \boldsymbol{\theta})$.