T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 28th of March, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

1. Given a hidden Markov model (HMM, page 610) and observations $\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}$, show that the predictive distribution of the observations $\mathbf{y}_t$ at time point t follows a mixture distribution.

Solution:

Let us first write the joint distribution of all variables:

$$P(y_1, \ldots, y_t, z_1, \ldots, z_t) = P(z_1)P(y_1 \mid z_1) \prod_{\tau=2}^{t} P(z_\tau \mid z_{\tau-1})P(y_\tau \mid z_\tau). \quad (1)$$

Then we can manipulate the predictive distribution:

$$\begin{aligned} P(y_t \mid y_1, \ldots, y_{t-1}) &= \sum_{z_t} P(y_t, z_t \mid y_1, \ldots y_{t-1}) && (2) \\ &= \sum_{z_t} P(z_t \mid y_1, \ldots y_{t-1}) P(y_t \mid z_t, y_1, \ldots y_{t-1}) && (3) \\ &= \sum_{z_t} P(z_t \mid y_1, \ldots y_{t-1}) P(y_t \mid z_t), && (4) \end{aligned}$$

which is clearly a mixture distribution with the posterior distribution of the latent variable $P(z_t \mid y_1, \ldots y_{t-1})$ as the mixture coefficients and $P(y_t \mid z_t)$ as the component distributions.

2. Show how a second-order Markov chain (page 608) of 3 symbols can be transformed to a hidden Markov model with 9 states and 3 symbols.

Solution:

A second order Markov chain has a model for $P(y_t \mid y_{t-2}, y_{t-1})$.

$P(y_t \mid y_{t-2}, y_{t-1})$	aa	ab	ac	ba	bb	bc	ca	cb	cc
$y_t = a$	.	.	.	.	.	.	.	.	.
$y_t = b$	.	.	.	.	.	.	.	.	.
$y_t = c$	.	.	.	.	.	.	.	.	.

By setting the hidden state z_t to contain both y_{t-1} and y_t as a concatenated symbol, we can emulate the second order Markov chain by a hidden Markov model using the following tables:

$P(y_t \mid z_t)$	aa	ab	ac	ba	bb	bc	ca	cb	cc
$y_t = a$	1	0	0	1	0	0	1	0	0
$y_t = b$	0	1	0	0	1	0	0	1	0
$y_t = c$	0	0	1	0	0	1	0	0	1

$P(z_t \mid z_{t-1})$	aa	ab	ac	ba	bb	bc	ca	cb	cc
$z_t = aa$	$\cdot$	0	0	$\cdot$	0	0	$\cdot$	0	0
$z_t = ab$	$\cdot$	0	0	$\cdot$	0	0	$\cdot$	0	0
$z_t = ac$	$\cdot$	0	0	$\cdot$	0	0	$\cdot$	0	0
$z_t = ba$	0	$\cdot$	0	0	$\cdot$	0	0	$\cdot$	0
$z_t = bb$	0	$\cdot$	0	0	$\cdot$	0	0	$\cdot$	0
$z_t = bc$	0	$\cdot$	0	0	$\cdot$	0	0	$\cdot$	0
$z_t = ca$	0	0	$\cdot$	0	0	$\cdot$	0	0	$\cdot$
$z_t = cb$	0	0	$\cdot$	0	0	$\cdot$	0	0	$\cdot$
$z_t = cc$	0	0	$\cdot$	0	0	$\cdot$	0	0	$\cdot$

where the values $\cdot$ are copied from the table of the second order Markov chain.

This shows that a hidden Markov model is more general than a second order Markov chain (and similarly of a Markov chain of any order).

3. Let us consider a HMM with a discrete hidden variable z with 6 states and a Gaussian observation (emission) probability density function. The dimension of the data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$ is 5 and the covariance function of the Gaussian distribution is diagonal. (a) Quantify the number of parameters in the model, (b) write the joint probability density, (c) and write the Q-function of the EM-algorithm $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ (page 440). Assume that the E-step is done, that is, $\gamma(z_t) = P(z_t \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ and $\xi(z_{t-1}, z_t) = P(z_{t-1}, z_t \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ are given.

Solution:

(a) Parameters $\boldsymbol{\theta}$ include the starting distribution $P(z_1) = \pi = P(z_1 \mid z_0)$ with 6 parameters of which 5 are free, transition matrix $\mathring{A}$ with 36 parameters of which 30 are free, and parameters μ_{ij} and σ_{ij}^2 for the emission distribution (60 parameters, all of them free). That makes altogether 102 parameters of which 95 are free.

(b) A Gaussian distribution with a diagonal covariance can be represented as a product of 1-dimensional Gaussians.

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = \prod_{t=1}^{T} P(z_t \mid z_{t-1}, \boldsymbol{\theta}) p(\mathbf{x}_t \mid z_t, \boldsymbol{\theta}) \tag{5}$$

$$= \prod_{t=1}^{T} a_{z_{t-1}, z_t} \prod_{k=1}^{5} \frac{1}{\sqrt{2\pi\sigma_{z_t,k}^2}} \exp\left[\frac{-(x_{tk} - \mu_{z_t k})^2}{2\sigma_{z_t k}^2}\right] \tag{6}$$

(c)

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \tag{7}$$

$$= \sum_{\mathbf{Z}} P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \left[\ln P(\mathbf{Z} \mid \boldsymbol{\theta}) + \ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})\right] \tag{8}$$

$$= \left[\sum_{t=1}^{T} \sum_{i=1}^{6} \sum_{j=1}^{6} \xi(z_{t-1,i}, z_{tj}) \ln a_i j\right] \tag{9}$$

$$+ \left[\sum_{t=1}^{T} \sum_{i=1}^{6} \sum_{k=1}^{5} \gamma(z_{ti}) \ln\left(\frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left[\frac{-(x_{tk} - \mu_{ik})^2}{2\sigma_{ik}^2}\right]\right)\right] \tag{10}$$

$$= Q_z + \sum_{t=1}^{T} \sum_{i=1}^{6} \sum_{k=1}^{5} \gamma(z_{ti}) \left[-\frac{(x_{tk} - \mu_{ik})^2}{2\sigma_{ik}^2} - \frac{1}{2} \ln(2\pi\sigma_{ik}^2)\right] \tag{11}$$

$$= Q_z + Q_x, \tag{12}$$

where the division into two parts $Q_z + Q_x$ will be useful in Problem 4.

4. In the setting of Problem 3, (a) derive the M-step for the Gaussian means μ_{ik}, where $i = 1 \ldots 6$ denotes the state and $k = 1 \ldots 5$ denotes the data dimension. (b) Derive the M-step for updating the 6×6 transition matrix $\mathbf{A}$.
Solution:

(a) As we maximize the Q-function w.r.t. a particular μ_{ik}, the part Q_z is constant, and from the sums over i and k, all the other terms are constant except the one we are interested in. Therefore we only need:

$$\frac{\partial}{\partial \mu ik} \sum_{t=1}^{T} \gamma(z_{ti}) \frac{-(x_{tk} - \mu ik)^2}{2\sigma_{ik}^2} = 0 \tag{13}$$

$$\sum_{t=1}^{T} \gamma(z_{ti}) \frac{x_{tk} - \mu ik}{\sigma_{ik}^2} = 0 \tag{14}$$

$$\mu_{ik} = \frac{\sum_{t=1}^{T} \gamma(z_{ti}) x_{tk}}{\sum_{t=1}^{T} \gamma(z_{ti})}, \tag{15}$$

that is, μ will be the weighted average of the data points assigned to the cluster (or state) i, the weights being the probabilities γ that this point belongs to this cluster.

(b) Next we should maximize Q w.r.t. an element of the transition matrix $a_i j$. This time Q_x is a constant that can be ignored. If we simply try to find the zero of the gradient, we notice that increasing $a_i j$ will always increase Q so there is no zero of the gradient. We need to take into account the constraint $\sum_{j=1}^{6} a_{ij} = 1 \forall i$. One way to do this is to introduce Lagrange multipliers $\lambda_i > 0$ for each constraint i. We will now maximize

$$Q_z - \lambda_i \left(\sum_{j=1}^{6} a_{ij} - 1 \right) \tag{16}$$

instead. The intuition behind this is to introduce a "counter-force" that balances the ever increasing a_{ij}s. When the force λ_i is just right, it will set the constraint to be true, and the modified cost function in Eq. (16) will be equal to Q_z since $\left(\sum_{j=1}^{6} a_{ij} - 1 \right) = 0$.

Let us try to maximize (16) by finding the zero of the gradient:

$$0 = \frac{\partial}{\partial a_{ij}} \left[\sum_{t=1}^{T} \xi(z_{t-1,i}, z_{tj}) \ln a_{ij} - \lambda_i (\sum_{j'=1}^{6} a_{ij'} - 1) \right] \tag{17}$$

$$= \frac{\sum_{t=1}^{T} \xi(z_{t-1,i}, z_{tj})}{a_{ij}} - \lambda_i \tag{18}$$

$$a_{ij} = \frac{\sum_{t=1}^{T} \xi(z_{t-1,i}, z_{tj})}{\lambda_i}. \tag{19}$$

Thus, λ_i turned out to be a normalization constant, whose value we can compute from

$$\sum_{j=1}^{6} a_{ij} = \sum_{j=1}^{6} \frac{\sum_{t=1}^{T} \xi(z_{t-1,i}, z_{tj})}{\lambda_i} = 1 \tag{20}$$

$$\lambda_i = \sum_{j=1}^{6} \sum_{t=1}^{T} \xi(z_{t-1,i}, z_{tj}). \tag{21}$$