T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Solutions to exercise 10, 30.3.2007

Problem 1.

i) Use the product rule to obtain

$$p(y, L|\theta, \lambda) = p(y|L, \theta, \lambda)p(L|\theta, \lambda) = p(y|L, \theta)p(L|\lambda)$$

Here we dropped λ from $p(y|L, \theta, \lambda)$ because when L is known, λ has no effect on y:s distribution. Also, we dropped θ from $p(L|\theta, \lambda)$ because L is independent of θ.

The first term is

$$p(y|L, \theta) = \prod_i \prod_j [N(y_i|\mu_j, \sigma_j^2)]^{L_{ij}}$$

The exponent L_{ij} is used to force the term to be equal to one when $L_{ij} = 0$. This way, each observation y_i gets exactly one term different from one in the product. This term is the mixture distribution having generated y_i.

The second term is

$$p(L|\lambda) = \prod_i \prod_j \lambda_j^{L_{ij}} \quad (*)$$

Again, the exponent is used to make some terms equal to one. This term can be understood as follows: consider a vector $L_i = (L_{i1}, L_{i2}, \ldots, L_{im})$. Exactly one of the components is one, and others are zero. Since the probability that $L_{ij} = 1$ is λ_j, then the probability that $L_i = (0, 0, \ldots, 1, \ldots, 0)$ is also λ_j. So we have to pick the single λ_j corresponding to $L_{ij} = 1$ for each observation y_i in the product $(*)$.

Finally, multiply the terms together to get

$$p(y, L|\theta, \lambda) = \prod_i \prod_j [N(y_i|\mu_j, \sigma_j^2)]^{L_{ij}} \lambda_j^{L_{ij}} = \prod_i \prod_j [\lambda_j N(y_i|\mu_j, \sigma_j^2)]^{L_{ij}},$$

which is other way of writing the result that was to be shown.

ii) For the Gibbs sampler for θ, we need $p(\theta|y, L)$.

$$p(\theta|y, L) \propto p(y|\theta, L)p(\theta|L) \propto p(y|\theta, L)$$

since θ and L are independent of each other and we assume a constant prior for θ .

Since $p(y_i|\theta, L) = N(y_i|\mu_j, \sigma_j^2),\ L_{ij} = 1$ the posterior factorizes into terms including each μ_j. Each term is a subproblem where the unknown Normal mean μ_j is inferred with a known variance. Old results apply and the posteriors are

$$p(\mu_j|y, L) = N(\mu_j|s_j, \sigma_j^2/n_j)$$

where s_j is the sample average of all y_i:s that are from mixture j, and n_j is the number of such y_i:s. (We assumed σ_j^2 known, so θ consists of μ_j's only.)

Then the Gibbs sampler for L. By the product rule we get

$$p(L|y, \theta) = p(L, y|\theta)/p(y|\theta)$$

Fix i and find the distribution $p(L_{ik} = 1, y_i|\theta)$.

Since $p(L_{ik} = 1, y_i|\theta) = p(y_i|L_{ik} = 1, \theta)p(L_{ik} = 1|\theta) = p(y_i|L_{ik} = 1, \theta)p(L_{ik} = 1)$, we get

$$p(L_{ik} = 1, y_i|\theta) = N(y_i|\mu_k, \sigma_k^2)\lambda_k$$

Since $p(y_i|\theta) = \sum_j p(y_i, L_{ij} = 1|\theta) = \sum_j p(y_i|L_{ij} = 1, \theta)p(L_{ij} = 1|\theta) = \sum_j \lambda_j N(y_i|\mu_j, \sigma_j^2)$, the result is

$$p(L_{ik} = 1|y_i, \theta) = N(y_i|\mu_k, \sigma_k^2)\lambda_k / \sum_j \lambda_j N(y_i|\mu_j, \sigma_j^2)$$

This can be simulated for each i, since it defines a discrete distribution over the values $k = 1, 2, \ldots, m$.

Now the Gibbs sampler is ready and consists of alternating the simulation of μ_j:s and simulation of L_{ij}:s. If desired also $p(L|y, \theta)$ could be written out explicitly but the resulting formula would be cumbersome and is not needed for the simulation.

Problem 2.

We consider the likelihood

$$p(y|\theta, \lambda) = \prod_i p(y_i|\theta, \lambda) = \prod_i \left[\lambda_1 N(y_i|\mu_1, \sigma^2) + \lambda_2 N(y_i|\mu_2, \sigma^2)\right],$$

where θ is the set of parameters. We wish to maximize the likelihood with respect to the parameters μ_1 and μ_2. This is the same as maximizing the log-likelihood. We utilize the Newton-Rhapson update formula

$$\mu_{m,new} = \mu_m - (\log p)'/(\log p)'',$$

where $p = p(y|\theta, \lambda)$.

The derivative of the log-likelihood with respect to μ_m is now (see lectures)

$$(\log p)' = \sum_i p(L_{im} = 1|\theta, y_i)\sigma^{-2}(y_i - \mu_m).$$

The second derivative is, assuming $p(L_{im} = 1|\theta, y_i)$ is constant with respect to μ_m,

$$\begin{aligned}(\log p)'' &\approx \sum_i p(L_{im} = 1|\theta, y_i)(\sigma^{-2}(y_i - \mu_m))' \\ &= \sum_i p(L_{im} = 1|\theta, y_i)(-\sigma^{-2}).\end{aligned}$$

The ratio $(\log p)'/(\log p)''$ is

$$\left[\sum_i p(L_{im}=1|\theta,y_i)(y_i-\mu_m)\sigma^{-2}\right] / \left[\sum_i p(L_{im}=1|\theta,y_i)(-\sigma^{-2})\right].$$

The terms σ^{-2} cancel out and we have

$$-\left[\sum_i p(L_{im}=1|\theta,y_i)(y_i-\mu_m)\right] / \left[\sum_i p(L_{im}=1|\theta,y_i)\right].$$

The mean μ_m does not depend on i so it comes out of the sum: finally,

$$(\log p)'/(\log p)'' = \mu_m - \frac{\sum_i p(L_{im}=1|\theta,y_i)y_i}{\sum_i p(L_{im}=1|\theta,y_i)}.$$

Finally, the Newton-Rhapson step is

$$\mu_{m,new} = \frac{\sum_i p(L_{im}=1|\theta,y_i)y_i}{\sum_i p(L_{im}=1|\theta,y_i)}.$$

EM interpretation: $p(L_{im}=1|\theta,y_i)$ corresponds to $q(L)$. In the E step, we average $\log p(\theta|y)$ over the distribution $q(L)$, and this is what actually happens in $(\log p)'$. In the M step, we assume $q(L)$ is fixed; similarly we did not differentiate $q(L)$ with respect to μ_m in the Newton-Raphson update.

Problem 3.

The Kullback-Leibler divergence is

$$D(q||p) = \int \log\frac{q(x)}{p(x)}q(x)dx = E_q(\log q - \log p).$$

KL divergence gives the average number of bits that are wasted by encoding events from a distribution q with a code based on the distribution p. KL divergence is always nonnegative and zero if $q=p$.

Now

$$-D(q||p(s|a,y)) = E_q(\log p(s|a,y)) - E_q(\log q)$$

and

$$\begin{aligned}
-D(q||p(s|a,y)) + \log p(a|y) &= -D(q||p(s|a,y)) + E_q(\log p(a|y)) \\
&= E_q(\log p(s|a,y)) - E_q(\log q) + E_q(\log p(a|y)) \\
&= E_q(\log(p(s|a,y)p(a|y))\,) - E_q(\log q) \\
&= E_q(\log p(s,a|y)) - E_q(\log q) \\
&= F(q,a).
\end{aligned}$$

The first step, choosing a distribution $q(s)$ that maximizes F, is equivalent to minimizing $D(q||p(s|a,y))$ since $\log p(a|y)$ does not depend on q. So we are looking for a distribution q

as close as possible to $p(s|a_0,y)$ where a_0 is the current value for the parameters. Naturally we may choose $q=p(s|a_0,y)$, making the KL divergence 0.

Next the parameters a are updated to maximize F. Now both terms in F are affected. If a_1 maximizes $F(q,a)$ then the second term $\log p(a|y)$ must increase or remain the same when a_0 is changed to a_1. This is because after the previous step, $F(q,a_0) = 0 + \log p(a_0|y)$. Since

$$F(q,a_1) = -D(p(s|a_0,y)||p(s|a_1,y)) + \log p(a_1|y)$$

where the first term is negative or zero, the last term must be at least as large as $\log p(a_0|y)$ (otherwise $F(q,a_1) < F(q,a_0)$ which is a contradiction as we chose a_1 so that it maximizes F). We get

$$\log p(a_1|y) \geq \log p(a_0|y) \implies p(a_1|y) \geq p(a_0|y).$$

In the Generalized EM (GEM) algorithm, the new parameters a_1 are chosen so that the value of F increases, instead of maximizing F. The above derivations still hold, that is, $\log p(a|y)$ cannot decrease. If a_1 is suitably chosen, the convergence of the GEM algorithm may be faster than the convergence of the original EM.

Problem 4.

In the M-step of the EM algorithm, we wish to maximize

$$F(q,a) = \sum_m \sum_i [\log N(y_i|\mu_m,\Sigma_m) + \log\lambda_m]\tau_{im}$$

with respect to the unknown parameters λ_m, μ_m and Σ_m while regarding τ_{im} as constants. We have written a as a shorthand for the non-latent parameters. Also, we have $\tau_{im} = p(L_{im}=1|a,y_i)$.

We first maximize $F(q,a)$ with respect to λ_m. The first term does not contain λ_m; we may leave it out for now. Include the constraint $\sum_m \lambda_m = 1$ with a Lagrange multiplier β and set the derivative to zero:

$$\frac{\partial}{\partial\lambda_m}\left[\sum_m\sum_i \log\lambda_m\ \tau_{im} + \beta(\sum_m \lambda_m - 1)\right] = \frac{1}{\lambda_m}\sum_i \tau_{im} + \beta = 0$$

which gives

$$\lambda_m = -\frac{1}{\beta}\sum_i \tau_{im}.$$

By using the constraint $\sum_m \lambda_m = 1$ we get $-\frac{1}{\beta}\sum_i\sum_m \tau_{im} = -\frac{1}{\beta}\sum_i 1 = 1 \implies -\beta = N$. Then

$$\lambda_m = \frac{1}{N}\sum_i \tau_{im}.$$

Next, we maximize $F(q,a)$ with respect to μ_m. Only the first term of the expression contains μ_m. We substitute in the probability density function of a Normal distribution

and set the derivative to zero:

$$\begin{aligned}
&\frac{\partial}{\partial \mu_m}\left[\sum_m \sum_i \log N(y_i|\mu_m, \Sigma_m)\tau_{im}\right] \\
=&\frac{\partial}{\partial \mu_m}\left[\sum_m \sum_i (-\frac{1}{2}\log|\Sigma_m| - \frac{1}{2}(y_i-\mu_m)\Sigma_m^{-1}(y_i-\mu_m) + K)\,\tau_{im}\right] \\
=&\sum_i \Sigma_m^{-1}(y_i-\mu_m)\,\tau_{im} = 0.
\end{aligned}$$

Here, K is a constant. Thus the expectation of mixture component m is a weighted sum over observations, the weights τ_{im} telling at which degree each observation y_i comes from the component distribution m:

$$\mu_m = \frac{\sum_i y_i \tau_{im}}{\sum_i \tau_{im}}.$$

Similarly, setting the derivative with respect to Σ_m to zero we would get an update formula for Σ_m. The update formula is not derived here.

Comments: the update formulas are easy to interpret since they are weighted averages over quantities that are clearly related to the parameters. The weights τ_{im} take into consideration the importance of sample y_i in representing the mixture component m.

Some references to the EM algorithm:

Redner and Walker: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, 26(2), 1984.

Bilmes: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models. Tech. Report TR-97-021, UC Berkeley. www.icsi.berkeley.edu/ftp/global/pub/techreports/1997/tr-97-021.pdf