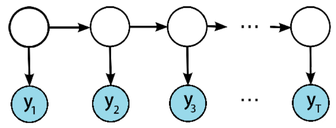
Variational Bayes For Continuous Hidden Markov Models and Its Application

CmpE 547

Bayesian Statistics and Machine Learning

Abdulkadir Çelikkanat

Consider a N-state HMM model $\Phi = \{\pi, A, C, \theta\}$ where π is the initial state probability distribution, $\boldsymbol{A}$ is the transition matrix, $\boldsymbol{C}$ is the mixture coefficient matrix, and θ is the parameter matrix consisting of the Gaussian parameters $\theta_{ik} = \{\mu_{ik}, R_{ik}\}$ for the kth mixture component of the ith state, with mean μ_{ik} and precision R_{ik}.



Let Y=$\{y_1, y_2, ..., y_T\}$ be an observation, X=$\{x_1, x_2, ..., x_T\}$ be the unobserved state sequence, and L=$\{l_1, l_2, ..., l_T\}$ be the indicator sequence, which shows which mixture component generates the observation y_i. Therefore, related complete-data is Z={X,Y,L}, and proability of Z can be written as

$$p(X, Y, L|\Phi) = \pi_{x_1}\left[\prod_{t=1}^{T-1} a_{x_t x_{t+1}}\right]\left[\prod_{t=1}^{T} c_{x_t l_t} f(y_t|\theta_{x_t l_t})\right] \tag{1}$$

the likelihood of the model paremeters θ given the data $\boldsymbol{Y}$ is

$$p(Y|\Phi) = \sum_{X,L} \pi_{x_1}\left[\prod_{t=1}^{T-1} a_{x_t x_{t+1}}\right]\left[\prod_{t=1}^{T} c_{x_t l_t} f(y_t|\theta_{x_t l_t})\right] \tag{2}$$

and the marginal likelihood can be expressed as

$$p(Y) = \frac{p(X, Y, L, \phi)}{p(X, L, \phi|Y)} \tag{3}$$

After taking the logarithm of equation (3), the expectation with respect to the distribition q(X,L,ϕ), then rearranging it, we obtain

$$logp(Y) = F(q) + KL(q||p) \tag{4}$$

where the F(q) known as the negative free energy is

$$F(q) = \int q(X, L, \phi) log \frac{p(Y, X, L, \phi)}{q(X, L, \phi)} dX dL d\phi \tag{5}$$

and the Kullback-Leibler(KL) divergence between approximate and true posterior densities is

$$KL(q||p) = \int q(X, L, \phi) log \frac{q(X, L, \phi)}{p(X, L, \phi|Y)} dX dL d\phi \tag{6}$$

The main aim of the VB is to maximize the lower bound F(q) by tuning distribution q(X,L,ϕ) so that it converges the true posterior distribution $p(X, L, \phi|Y)$. However, here two problems arises:

a) Choosing the suitable form of variational density which is tractable and will make good approximation to the true posterior $p(X, L, \phi|Y)$,

b) Choosing the prior distributions of model parameters $\phi = \{\pi, A, C, \theta\}$

The factorized form

$$q(X, L, \phi) = q(X)q(L)q(\pi)q(A)q(C)q(\theta) \tag{7}$$

which has been successfully applied in many applications of variational methods was preferred.

The prior on the model parameters can be written as

$$p(\phi) = p(\pi)p(A)p(C)p(\theta) \tag{8}$$

where

$$\begin{aligned} p(\pi) &= Dir(\pi_1, ..., \pi_N|u_1^\pi, ..., u_N^\pi), \\ p(A) &= \prod_{i=1}^{N} Dir(a_{i1}, ..., a_{iN}|u_{i1}^A, ..., u_{iN}^A), \\ p(C) &= \prod_{i=1}^{N} Dir(c_{i1}, ..., c_{iN}|u_{i1}^C, ..., u_{iK}^C), \\ p(\theta) &= \prod_{i=1}^{N}\prod_{k=1}^{K} NW(\mu_{ik}, R_{ik}|a_{ik}, b_{ik}, \lambda_{ik}, m_{ik}). \end{aligned} \tag{9}$$

since the Dirichlet distribution is the conjugate prior of the multinomial distribution, and Normal-Wishart distribution is the conjugate prior of a multivariate normal distribution with unknown mean and precision.

M-step

To maximize F(q), the variational posterior $q(X, L, \phi)$ is updated with respect to $q(\phi)$ on fixed hidden variables at q(X,L). If the equation (5) is subsituted with equations (8) and (9), we obtain

$$\begin{aligned} F(q) = \int q(X)q(L)q(\pi)q(A)q(C)q(\theta)\Big[& log\pi_{x_1} \\ & + \sum_{t=1}^{T-1} loga_{x_t x_{t+1}} + \sum_{t=1}^{T} logc_{x_t l_t} + \sum_{t=1}^{T} logf(y_t|\theta_{x_t l_t}) \\ & + logp(\pi) + logp(A) + logp(C) + logp(\theta) - logq(X) \\ & - logq(L) - logq(\pi) - logq(A) - logq(C) \\ & - logq(\theta)\Big] dX dL d\phi \\ & = F(q(A)) + F(q(C)) + F(q(\theta)) + const \end{aligned} \tag{10}$$

Optimization of q(A), $q(\pi)$, q(C), and $q(\theta)$

By collecting all the terms related to q(A) together, we can write

$$\begin{aligned} F(q(A)) = & \int q(A) \sum_X q(X) \sum_{t=1}^{T-1} loga_{x_t x_{t+1}} dA \\ & + \int q(A) logp(A) dA - \int q(A) logq(A) dA \end{aligned} \tag{11}$$

Then, we have

$$F(q(A)) = -\int q(A) log\left[\frac{q(A)}{\prod_{i,j=1}^{N} a_{ij}^{W_{ij}^A - 1}}\right] dA \tag{12}$$

where the hyperparameter $W_{ij}^A = \sum_{t=1}^{T-1} w_{ij}^t + u_{ij}^A$ and $w_{ij}^t = q(x_t = i, x_{t+1} = j)$. Then, F(q(A)) is maximized with respect to q(A) by Gibbs inequality, so

$$q(A) = \prod_{i=1}^{N} Dir(a_{i1}, ..., a_{iN}|W_{i1}^A, ..., W_{iN}^A) \tag{13}$$

Similarly, F(q) can be optimized by using similar operations, and acquire the optimal distributions

$$\begin{aligned} q(\pi) &= Dir(\pi_1, ..., \pi_N|W_1^\pi, ..., W_N^\pi) \\ q(C) &= \prod_{i=1}^{N} Dir(c_{i1}, ..., c_{iK}|W_{i1}^C, ..., W_{iK}^C) \\ q(\theta_{ik}) &= \frac{(\lambda_{ik}/2\pi)^{d/2}}{Z(a_{ik}, b_{ik})(2\pi)^{dw_{ik}/2}} |R_{ik}|^{\frac{a_{ik}+w_{ik}-d}{2}} \\ & \times exp\left[-(\lambda_{ik}^{'}/2)(\mu_{ik} - m_{ik}^{'})^T R_{ik}(\mu_{ik} - m_{ik}^{'})\right] \\ & \times exp\left[-\frac{1}{2} Tr(b_{ik}^{'} R_{ik})\right] \end{aligned} \tag{14}$$

where the parameters of $q(\pi)$, and $q(C)$ are
$W_i^\pi = w_i^\pi + u_i^\pi$, $w_i^\pi = q(x_1 = i)$, $W_{ik}^C = \sum_{t=1}^{T} w_{ik}^t + u_{ik}^C$, $w_{ik}^t = q(x_t = i, l_t = k)$,
and parameters for $q(\theta)$ are
$w_{ik} = \sum_{t=1}^{T} w_{ik}^t$, $x_{ik} = \sum_{t=1}^{T} w_{ik}^t x_t / w_{ik}$, $S_{ik} = \sum_{t=1}^{T} w_{ik}^t (x_t - x_{ik})(x_t - x_{ik})^T$, $a_{ik}^{'} = a_{ik} + w_{ik}$, $b_{ik}^{'} = b_{ik} + S_{ik} + \frac{\lambda_{ik} w_{ik}}{\lambda_{ik} + w_{ik}} (m_{ik} - x_{ik})(m_{ik} - x_{ik})^T$, $\lambda_{ik}^{i} = \lambda_{ik} + w_{ik}$, $m_{ik}^{'} = \frac{\lambda_{ik} m_{ik} + w_{ik} x_{ik}}{\lambda_{ik} + w_{ik}}$

E-step

To maximize F(q), the variational posterior $q(X, L, \phi)$ on hidden variables $q(X, L)$ is updated on fixed model parameters $q(\phi)$. If the equation (5) is subsituted with equations (8) and (9),and then rearranged, we can get the equation

$$F(q) = F(q(X, L)) - KL(q(\psi)||p(\psi)) \tag{15}$$

where

$$\begin{aligned} & F(q(X, L)) \\ & = \sum_X q(X) \int q(\pi) log\pi_{x_1} d\pi \\ & \quad + \sum_X q(X) \int q(A) \sum_{t=1}^{T-1} log(a_{x_t x_{t+1}}) dA \\ & \quad + \sum_{X,L} q(X, L) \int q(C) \sum_{t=1}^{T} log(c_{x_t l_t}) dC \\ & \quad + \sum_{X,L} q(X, L) \int q(\theta) \sum_{t=1}^{T} logf(y_t|\theta_{x_t l_t}) d\theta \\ & \quad - \sum_{X,L} logq(X, L), \end{aligned} \tag{16}$$

Since the second term $KL(q(\psi)||p(\psi))$ in equation (15) is constant, only the first term $F(q(X, L))$ need to be optimized.

Now, defining

$$log\pi_{x_1}^* = \int q(\pi) log\pi_{x_1} d\pi = \psi(W_{x_1}^\pi) - \psi(W_0^\pi) \tag{17}$$

$$loga_{x_t x_{t+1}}^* = \int q(A) loga_{x_t x_{t+1}} dA = \psi(W_{x_t x_{t+1}}^A) - \psi(W_{x_t 0}^A) \tag{18}$$

$$logc_{x_t l_t}^* = \int q(C) logc_{x_t l_t} dC = \psi(W_{x_t l_t}^C) - \psi(W_{x_t 0}^C) \tag{19}$$

$$\begin{aligned} logf^*(y_t|\theta_{x_t|l_t}) = & \int q(\theta) logf(y_t|\theta_{x_t l_t}) d\theta \\ = & \frac{-d}{2} log 2\pi - \frac{1}{2} log\left|\frac{b_{x_t l_t}}{2}\right| \\ & + \frac{1}{2}\sum_{i=1}^{d} \psi\left(\frac{a_{x_t l_t + 1 - i}}{2}\right) \\ & - \frac{1}{2} a_{x_t l_t} (x_t - m_{x_t l_t})^T b_{x_t l_t}^{-1} (x_t - m_{y_t l_t}) \\ & - \frac{d}{2\lambda_{y_t l_t}} \end{aligned} \tag{20}$$

where $\psi(.)$ is the digamma function and the parameters $W_0^\pi = \sum_{i=1}^{i=N} W_i^\pi$, $W_{x_t 0}^A = \sum_{i=1}^{i=N} W_{x_t i}^A$, and $W_{x_t 0}^C = \sum_{i=1}^{i=N} W_{x_t i}^C$. Then, after subsituting (17)-(20) into (16), we obtain

$$F(q(X, L)) = \sum_{X,L} q(X, L) log \frac{\pi_{x_1}^* \prod_{t=1}^{T-1} a_{x_t} a_{x_t + 1} \prod_{t=1}^{T} c_{x_t l_t}^* f^*(y_t|\theta_{x_t l_t})}{q(X, L)} \tag{21}$$

so the optimized

$$q(X, L) = \frac{1}{Z}\left[\pi_{x_1}^* \prod_{t=1}^{T} a_{x_t x_{t+1}}^* \prod_{t=1}^{T} c_{x_t l_t}^* f^*(y_t|\theta_{x_t l_t})\right] \tag{22}$$

It can be noticed that $Z = q(X|\phi^*)$, after comparing it with 2.

Convergence

Each iteration increases the negative free energy $F(q)$, or it remains as unchanged, so it is significant to approximate the marginal likelihood. This quantity can be computed by using the equations (15) and (21)-(22) in the following way

$$\begin{aligned} F(q) &= F(q(X, L)) - KL(q(\phi)||p(\phi)) \\ &= logq(Y||\phi^*) - KL_{Dir}(q(\pi)||p(\pi)) \\ &\quad - KL_{Dir}(q(A)||p(A)) - KL_{Dir}(q(C)||p(C)) \\ &\quad - KL_{NW}(q(\theta)||p(\theta)) \end{aligned} \tag{23}$$

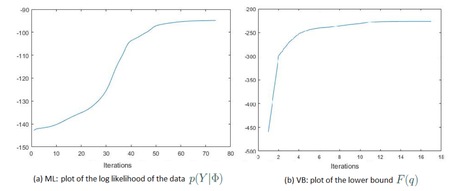
Hence, the algorithm can be terminated when the change in $F(q)$ is negligibly small.

Application

For an experiment, a synthetic data consisting of four different letter $\{a, b, c, d\}$ sequences has been generated. The data set contain 19 letter sequences and the longest sequence contains 426 letters, smallest contains 14 letters. Some instances of the generated data set as follows:

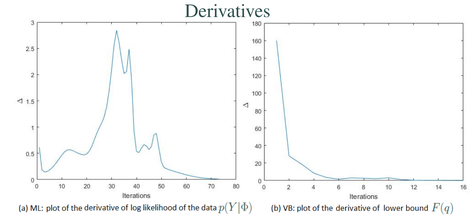
$$\begin{aligned} y_{1:T_1} &= (bbcccbdcdbbbcc...acdbcacddabcbbbbcadab) \\ y_{1:T_2} &= (cddadcdbaacabd...aacadbbadacbdccbbddac) \\ &... \\ &... \\ y_{1:T_{18}} &= (ccdabccbbdabdb...cbcaccaddccccbdbbdbbbcb) \\ y_{1:T_{19}} &= (bdaadbbbbacabadbdbcb...cadadaaddaaccd) \end{aligned}$$

In the figures below, the performances of ML and VB are compared over this generated data set. It is also assumed that the number of hidden states of HMM is 15.



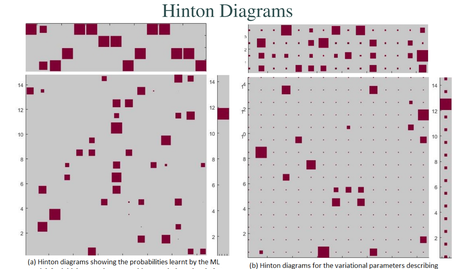
(a) ML: plot of the log likelihood of the data $p(Y|\Phi)$ (b) VB: plot of the lower bound $F(q)$

As shown in the figures above, it takes ML about 70 iterations to converge, and takes rougly 12 iterations for VB.



(a) ML: plot of the derivative of log likelihood of the data $p(Y|\Phi)$ (b) VB: plot of the derivative of lower bound $F(q)$

The derivatives of the functions given in the previous figures for ML and VB algorithms.



(a) Hinton diagrams showing the probabilities learnt by the ML model, for initial state prior **π**, transition matrix **A**, and emission matrix **C**

(b) Hinton diagrams for the variational parameters describing the posterior distributions over the parameters q(π), q(A), and q(C) respectively