STAT 534 Lecture 8

May 16, 2013

Hidden Markov Models Summary

©Marina Meilă

mmp@stat.washington.edu

Reading: Article by Lawrence Rabiner "A tutorial on hidden markov models and selected applications in speech recognition"; [Optional CRLS Problem 15–5, page 367]

Notation (consistent with the Rabiner article)

$q_t \in \{1, \ldots, N\}$	the state at time t
$O_t \in \{1, \ldots, M\}$	the output at time t
$A = [a_{ij}]_{i,j=1}^{N}$	the transition matrix
$B = [b_i(k)]_{i=1}^{N}{}_{k=1}^{M}$	the emission matrix
$\pi = [\pi_i]_{i=1}^{M}$	the initial state distribution

1 The Forward-Backward Algorithm

$$\begin{aligned} \alpha_t(i) &= P[\, O_{1:t},\ q_t = i \,] \\ \beta_t(i) &= P[\, O_{t+1:T} \,|\, q_t = i \,] \end{aligned}$$

The Forward pass

$$\begin{aligned} \alpha_1(i) &= \pi_i b_i(O_1) \\ \alpha_t(i) &= \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} b_i(O_t) \end{aligned}$$

The Backward pass

$$\begin{aligned}\beta_T(i) &= 1 \\ \beta_t(i) &= \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)\end{aligned}$$

The likelihood of the outputs

$$P[\, O_{1:T} \,] \;=\; \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \;=\; \sum_{i=1}^{N} \alpha_T(i)$$

The probability of a state or transition given the data

$$\begin{aligned}\gamma_t(i) &= P[\, q_t = i \,|\, O_{1:T} \,] \;=\; \frac{\alpha_t(i)\beta_t(i)}{P[\, O_{1:T} \,]} \\ \xi_t(i,j) &= P[\, q_t = i,\; q_{t+1} = j \,|\, O_{1:T} \,] \;=\; \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P[\, O_{1:T} \,]}\end{aligned}$$

2 The Baum-Welch Algorithm

The BW algorithm is an EM algorithm that maximizes the likelihood of the observations given the model (A, B, π).

Log-likelihood of a sequence of length T

$$l(A, B, \pi) \;=\; \ln P(O_{1:T} | A, B, \pi). \tag{1}$$

The **Complete Log-Likelihood** of a sequence of length T is defined as

$$l_c(A, B, \pi) \;=\; \ln P(O_{1:T}, q_{1:T} | A, B, \pi). \tag{2}$$

Now, as you know from the May 7 Lecture,

$$\ln P(O_{1:T}, q_{1:T} | A, B, \pi) \;=\; \ln \pi_0(q_1) + \sum_{t=1}^{T} \ln b_{q_t O_t} + \sum_{t=1}^{T-1} \ln a_{q_t q_{t+1}}. \tag{3}$$

Define now the **indicator variables** $z_{it} = 1$ if $q_t = i$ and 0 otherwise. By using the z_{it} indicator variables, the above can be rewritten as a sum.

$$l_c(A, B, \pi) = \sum_{i=1}^{N} z_{i1} \ln \pi_0(i) + \sum_{t=1}^{T} \sum_{i=1}^{n} z_{it} \ln b_{iO_t} + \sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{j=1}^{n} z_{it} z_{j,t+1} \ln a_{ij} \quad (4)$$

We cannot evaluate l_c from observed outputs only, because we do not know the variables z_{it}. The idea of the Baum-Welch algorithm is to compute the expectation of l_c w.r.t. the current value of the parameters A, B, π. The function $Q(\lambda = (A, B, \pi), \bar{\lambda} = \bar{A}, \bar{B}, \bar{\pi}))$ (see equation (41) in the Rabiner tutorial) is the expectation

$$E_{P(q_{1:T}^{(1:n)} | O_{1:T}^{(1:n)}, A, B, \pi)} \left[l_c(\bar{A}, \bar{B}, \bar{\pi}) \right] = E_{P(q_{1:T}^{(1:n)} | O_{1:T}^{(1:n)}, A, B, \pi)} \left[\ln P(O_{1:T}, q_{1:T} | \bar{A}, \bar{B}, \bar{\pi}) \right]. \quad (5)$$

In $Q(,)$, A, B, π represent the current values of the parameters, while $(\bar{A}, \bar{B}, \bar{\pi})$ represent the new values we want to update to. It is easy to see, too, that

$$E_{P(q_{1:T}^{(1:n)} | O_{1:T}^{(1:n)}, A, B, \pi)}[z_{it}] = \gamma_i(t) \quad E_{P(q_{1:T}^{(1:n)} | O_{1:T}^{(1:n)}, A, B, \pi)}[z_{it} z_{j,t+1}] = \xi_{ij}(t) \quad (6)$$

Hence, (2) becomes

$$Q((A, B, \pi), (\bar{A}, \bar{B}, \bar{\pi})) = \sum_{i=1}^{N} \gamma_1(i) \ln \pi_0(i) + \sum_{t=1}^{T} \sum_{i=1}^{N} \gamma_t(i) \ln b_{iO_t} + \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_t(i, j) \ln a_{ij}. \quad (7)$$

The Expectation step of BW computes the expectation of l_c in (2), which will depend on the γ, ξ quantities; while the Maximization step maximizes the expected l_c w.r.t. $\bar{A}, \bar{B}, \bar{\pi}$ in equation (7).

Expectation step: Compute α, β, γ, ξ from the current parameter estimates A, B, π

Maximization step: Reestimate parameters by

$$\begin{aligned} \pi_i &= \gamma_1(i) \\ a_{ij} &= \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\ b_i(k) &= \frac{\sum_{t:\, O_t = k} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \end{aligned}$$

For multiple sequences sampled independently, the (complete) log-likelihood is the sum of the log-likelihoods of individual sequences, hence,

$$l(A, B, \pi) = \sum_{k=1}^{n} \ln P(O_{1:T}^{(k)} | A, B, \pi) \tag{8}$$

$$l_c(A, B, \pi) = \sum_{k=1}^{n} \ln P(O_{1:T}^{(k)}, q_{1:T}^{(k)} | A, B, \pi). \tag{9}$$

Hence,

$$Q((A, B, \pi), (\bar{A}, \bar{B}, \bar{\pi})) = \sum_{k=1}^{n} \left[\sum_{i=1}^{N} \gamma_1^{(k)}(i) \ln \pi_0(i) + \right. \tag{10}$$

$$\left. \sum_{t=1}^{T} \sum_{i=1}^{N} \gamma_t^{(k)}(i) \ln b_{iO_t^{(k)}} + \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_t^{(k)}(i, j) \ln a_{ij} \right] \tag{11}$$

3 The Viterbi Algorithm

Notation	
$\delta_t(i)$	the optimum cost of a state sequence $[q_{1:t-1},\, q_t = i]$
	$= \max_{q_{1:t-1}} P[q_{1:t-1},\, q_t = i,\, O_{1:t}]$
$\psi_t(i)$	pointer for backtracking $= q_{t-1}$ in the optimal state sequence

Initialization

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(O_1) \\ \psi_1(i) &= \text{undefined} \end{aligned}$$

Recursion

$$\begin{aligned} \delta_t(i) &= \max_{j=1,\ldots N} [\delta_{t-1}(j) a_{ji}] b_i(O_t) \\ \psi_t(i) &= \operatorname*{argmax}_{j=1,\ldots N} [\delta_{t-1}(j) a_{ji}] \end{aligned}$$