Derivation of EM algorithm for the Hidden Markov Model

Yong See Foo

April 30, 2020

1 Model

Suppose we observe $\mathbf{X} = (x_1, \dots x_T)$, which correspond to unobserved states $\mathbf{Z} = (z_1, \dots z_T)$, where $z_t \in \{1, \dots S\}$ for $t = 1, \dots, T$. The model is specified with

- initial state distribution $p(z_1 = j) = \pi_j$,
- transition probabilities $p(z_t = j \mid z_{t-1} = i) = A_{i,j}$,
- and emission probabilities $p(x_t \mid z_t) = f(x_t; \phi_{z_t})$, i.e. ϕ_j are the parameters governing the density $p(x_t \mid z_t = j)$.

2 EM algorithm

Our aim is to find the parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi})$ to maximise the likelihood $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$. We can achieve this by iteratively updating

$$\boldsymbol{\theta}^{(n+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underbrace{\mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(n)}} [\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})]}_{:= Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n)}\right)}$$

until convergence.

2.1 M-step

One difficulty is that it is intractable to calculate $p\left(\mathbf{Z} \,\middle|\, \mathbf{X}, \boldsymbol{\theta}^{(n)}\right)$ for every possible sequence $\mathbf{Z}$. We first look at the M-step and anticipate what quantities related to $\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(n)}$ need to be obtained during the E-step.

The log-likelihood factorises into

$$\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = \log \pi_{z_1} + \sum_{t=2}^{T} \log A_{z_{t-1}, z_t} + \sum_{t=1}^{T} \log f(x_t; \phi_{z_t}).$$

2.1.1 Updating $\boldsymbol{\pi}$

We seek to maximise

$$\sum_{Z \in \{1, \dots, S\}^T} p\left(\mathbf{Z} \,\middle|\, \mathbf{X}, \boldsymbol{\theta}^{(n)}\right) \log \pi_{z_1} = \sum_{j=1}^{s} p\left(z_1 = j \,\middle|\, \mathbf{X}, \boldsymbol{\theta}^{(n)}\right) \log \pi_j.$$

Applying Lagrange multipliers then gives

$$\pi_j^{(n+1)} = \frac{p\left(z_1 = j \,\middle|\, \mathbf{X}, \boldsymbol{\theta}^{(n)}\right)}{\sum_{j'=1}^{S} p\left(z_1 = j' \,\middle|\, \mathbf{X}, \boldsymbol{\theta}^{(n)}\right)}.$$

2.1.2 Updating ϕ_j

We seek to maximise

$$\sum_{Z\in\{1,\ldots,S\}^T}\left[p\Big(\mathbf{Z}\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)\sum_{t=1}^{T}\log f(x_t;\phi_{z_t})\right]=\sum_{t=1}^{T}\sum_{j=1}^{S}p\Big(z_t=j\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)\log f(x_t;\phi_j).$$

This is equivalent to the M-step for mixture models, i.e. if EM can be applied to the mixture model where the observations are distributed according to f, then this step can be dealt with similarly. So far, we will need to obtain

$$\gamma_j(t)\coloneqq p\Big(z_t=j\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)$$

during the E-step.

2.1.3 Updating $A_{i,\cdot}$

We seek to maximise

$$\sum_{Z\in\{1,\ldots,S\}^T}\left[p\Big(\mathbf{Z}\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)\sum_{t=2}^{T}\log A_{z_{t-1},z_t}\right]=\sum_{t=2}^{T}\sum_{i=1}^{S}\sum_{j=1}^{S}p\Big(z_{t-1}=i,z_t=j\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)\log A_{i,j}.$$

Applying Lagrange multipliers then gives

$$A_{i,j}^{(n+1)}=\frac{\sum_{t=2}^{T}p\Big(z_{t-1}=i,z_t=j\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)}{\sum_{t=2}^{T}p\Big(z_{t-1}=i\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)}=\frac{\sum_{t=2}^{T}p\Big(z_{t-1}=i,z_t=j\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)}{\sum_{t=1}^{T-1}\gamma_i(t)}.$$

So, in the E-step we also need to obtain

$$\xi_{i,j}(t)=p\Big(z_{t-1}=i,z_t=j\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big).$$

2.2 E-step

2.2.1 Obtaining $\gamma_i(t)$

We have

$$\begin{aligned}\gamma_j(t)&=p\Big(z_t=j\;\Big|\;\mathbf{X},\boldsymbol{\theta}^{(n)}\Big)\\&\propto p\Big(z_t=j,\mathbf{X}\;\Big|\;\boldsymbol{\theta}^{(n)}\Big)\\&=p\Big(z_t=j\;\Big|\;\boldsymbol{\theta}^{(n)}\Big)p\Big(\mathbf{X}_{1:t}\;\Big|\;z_t=j,\boldsymbol{\theta}^{(n)}\Big)p\Big(\mathbf{X}_{t+1:T}\;\Big|\;z_t=j,\boldsymbol{\theta}^{(n)}\Big)\\&=\underbrace{p\Big(\mathbf{X}_{1:t},z_t=j\;\Big|\;\boldsymbol{\theta}^{(n)}\Big)}_{\alpha_j(t)}\underbrace{p\Big(\mathbf{X}_{t+1:T}\;\Big|\;z_t=j,\boldsymbol{\theta}^{(n)}\Big)}_{\coloneqq\beta_j(t)},\end{aligned}$$

where in the third line we utilise the conditional independence of $\mathbf{X}_{1:t}$ and $\mathbf{X}_{t+1:T}$ given z_t.

The quantities $\alpha_j(t)$ and $\beta_j(t)$ can be found by using dynamic programming (i.e. recurrence relations):

$$\begin{aligned}\alpha_j(1)&=\pi_j^{(n)}f(x_1;\phi_j)\\\alpha_j(t)&=\sum_{i=1}^{S}p\Big(\mathbf{X}_{1:t},z_{t-1}=i,z_t=j\;\Big|\;\boldsymbol{\theta}^{(n)}\Big)\\&=\sum_{i=1}^{S}p\Big(x_t\;\Big|\;z_t=j,\boldsymbol{\theta}^{(n)}\Big)p\Big(z_t=j\;\Big|\;z_{t-1}=i,\boldsymbol{\theta}^{(n)}\Big)p\Big(\mathbf{X}_{1:t-1},z_{t-1}=i\;\Big|\;\boldsymbol{\theta}^{(n)}\Big)&&\text{(Markov property)}\\&=\sum_{i=1}^{S}f\Big(x_t;\phi_j^{(n)}\Big)A_{i,j}^{(n)}\alpha_i(t-1)&&\text{for } t>1\end{aligned}$$

$$
\begin{aligned}
\beta_j(T) &= 1 \\
\beta_j(t) &= \sum_{k=1}^{S} p\Big(\mathbf{X}_{t+1:T}, z_{t+1} = k \;\Big|\; z_t = j, \boldsymbol{\theta}^{(n)}\Big) \\
&= \sum_{k=1}^{S} p\Big(x_{t+1} \;\Big|\; z_{t+1} = k, \boldsymbol{\theta}^{(n)}\Big) p\Big(\mathbf{X}_{t+2:T} \;\Big|\; z_{t+1} = k, \boldsymbol{\theta}^{(n)}\Big) p\Big(z_{t+1} = k \;\Big|\; z_t = j, \boldsymbol{\theta}^{(n)}\Big) && \text{(Markov property)} \\
&= \sum_{k=1}^{S} f\Big(x_{t+1}; \phi_k^{(n)}\Big) \beta_k(t+1) A_{j,k}^{(n)} && \text{for } t < T.
\end{aligned}
$$

This allows us to calculate

$$
\gamma_j(t) = \frac{\alpha_j(t)\beta_j(t)}{\sum_{j'=1}^{S} \alpha_{j'}(t)\beta_{j'}(t)}.
$$

2.2.2 Obtaining $\xi_i(t)$

We have

$$
\begin{aligned}
\xi_{i,j}(t) &= p\Big(z_{t-1} = i, z_t = j \;\Big|\; \mathbf{X}, \boldsymbol{\theta}^{(n)}\Big) \\
&= p\Big(z_{t-1} = i, z_t = j \;\Big|\; \boldsymbol{\theta}^{(n)}\Big) p\Big(\mathbf{X}_{1:t-1} \;\Big|\; z_{t-1} = i, \boldsymbol{\theta}^{(n)}\Big) p\Big(\mathbf{X}_{t:T} \;\Big|\; z_t = j, \boldsymbol{\theta}^{(n)}\Big) \\
&\propto p\Big(z_{t-1} = i, z_t = j, \mathbf{X} \;\Big|\; \boldsymbol{\theta}^{(n)}\Big) \\
&= p\Big(z_t = j \;\Big|\; z_{t-1} = i, \boldsymbol{\theta}^{(n)}\Big) p\Big(\mathbf{X}_{1:t-1}, z_{t-1} = i \;\Big|\; \boldsymbol{\theta}^{(n)}\Big) p\Big(\mathbf{X}_{t+1:T} \;\Big|\; z_t = j, \boldsymbol{\theta}^{(n)}\Big) \\
&= A_{ij}^{(n)} \alpha_i(t-1) f\Big(x_t; \phi_j^{(n)}\Big)\, \beta_j(t),
\end{aligned}
$$

where we utilise the conditional independence of $\mathbf{X}_{1:t-1}$ and $\mathbf{X}_{t:T}$ given z_{t-1} and z_t, and also use the Markov property. This implies that

$$
\xi_{i,j}(t) = \frac{A_{ij}^{(n)} \alpha_i(t-1) f\Big(x_t; \phi_j^{(n)}\Big)\, \beta_j(t)}{\sum_{i'=1}^{S} \sum_{j'=1}^{S} A_{i'j'}^{(n)} \alpha_{i'}(t-1) f\Big(x_t; \phi_{j'}^{(n)}\Big)\, \beta_{j'}(t)}.
$$