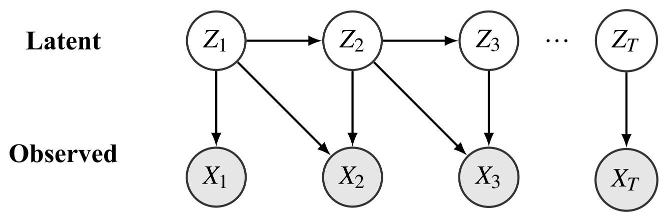
Notes

Oct 9 2019

We tend to find the Forward-backward algorithm for the new model.

Suppose we have T observations and T hidden states. Let X_i denote the ith observation where $X_i \in \{v_1, v_2, \cdots, v_k\}$ (or just 1, $\cdots$ k for convinience). Let Z_i denote the ith hidden states where $Z_i \in \{o_1, o_2, \cdots, o_m\}$ (or just 1, $\cdots$ m for convinience). The graphical model is below.



1 Expectation Part

Generally we need to know $P(X_{1:T})$ for future computation. From property of HMM we know

$$P(X_{1:T}, Z_{1:T}) = P(Z_1)\,P(X_1|Z_1)\prod_{t=2}^{T} P(Z_t|Z_{t-1})\,P(X_t|Z_{t-1}, Z_t)$$

Thus we repeat same operation,

$$\begin{aligned}
P(X_{1:T}) &= \sum_{Z_{1:T}} P(X_{1:T}, Z_{1:T}) \\
&= \sum_{Z_{1:T}} \underbrace{P(Z_1)P(X_1|Z_1)}_{S_1(Z_1)} \prod_{t=2}^{T} P(Z_t|Z_{t-1})\,P(X_t|Z_{t-1}, Z_t) \\
&= \sum_{Z_{2:T}} \underbrace{\left(\sum_{Z_1} S_1 P(Z_2|Z_1)P(X_2|Z_1, Z_2)\right)}_{S_2(Z_2)} \prod_{i=3}^{T} P(Z_t|Z_{t-1})\,P(X_t|Z_{t-1}, Z_t) \\
&\cdots \\
&= \sum_{Z_{j+1:T}} \underbrace{\left(\sum_{Z_j} S_j P(Z_{j+1}|Z_j)P(X_{j+1}|Z_j, Z_{j+1})\right)}_{S_{j+1}(Z_{j+1})} \prod_{t=j+2}^{T} P(Z_t|Z_{t-1})\,P(X_t|Z_{t-1}, Z_t)
\end{aligned}$$

By same kind of trick, we also have

$$
\begin{aligned}
P(X_{1:T}) &= \sum_{Z_{1:T}} P(X_{1:T}, Z_{1:T}) \\
&= \sum_{Z_{1:T-1}} \underbrace{\sum_{Z_T} P(Z_T|Z_{T-1})P(X_T|Z_{T-1},Z_T)}_{R_{T-1}(Z_{T-1})} P(Z_1)P(X_1|Z_1)\prod_{t=2}^{T-1} P(Z_t|Z_{t-1})P(X_t|Z_{t-1},Z_t) \\
&= \sum_{Z_{1:T-1}} \underbrace{\Big(\sum_{Z_{T-1}} R_{T-1}P(Z_{T-1}|Z_{T-2})P(X_{T-1}|Z_{T-2},Z_{T-1})\Big)}_{R_{T-2}(Z_{T-2})} P(Z_1)P(X_1|Z_1)\prod_{i=2}^{T-2} P(Z_t|Z_{t-1})P(X_t|Z_{t-1},Z_t) \\
&\cdots \\
&= \sum_{Z_{1:j}} \underbrace{\Big(\sum_{Z_j} R_j P(Z_j|Z_{j-1})P(X_j|Z_{j-1},Z_j)\Big)}_{R_{j-1}(Z_{j-1})} P(Z_1)P(X_1|Z_1)\prod_{t=2}^{j-1} P(Z_t|Z_{t-1})P(X_t|Z_{t-1},Z_t)
\end{aligned}
$$

Thus we decomposed $P(X_{1:T})$ as function of $P(Z_{j+1}|Z_j)$, $P(X_{j+1}|Z_j,Z_{j+1})$(or $P(X_1|Z_1)$) and $P(Z_1)$ in two ways. Then we set the parameter vector θ as

- $\pi = (\pi_1, \cdots, \pi_m)$, where $\pi_i = P(Z_1 = i)$
- $\phi_0 = \Big(b_{in}\Big)_{m\times k}$, where $b_{in} = P(X_1 = n|Z_1 = i)$
- $\phi = \big(b_{ijn}\big)_{m\times m\times k}$, where $b_{ijn} = P(X_t = n|Z_{t-1} = i, Z_t = j)$
- $T = \Big(t_{ij}\Big)_{m\times m}$, where $t_{ij} = P(Z_{t+1} = j|Z_t = i)$
- $\theta = (\pi, \phi_0, \phi, T)$

And the two notations, S and R, have specific meanings and we introduce α and β here. (If having priors θ, we may condition on it).

$$
\begin{aligned}
\alpha_t(Z_t) &:= S_t(Z_t) = P(X_1, \cdots, X_t, Z_t|\theta) \\
\beta_t(Z_t) &:= R_t(Z_t) = P(X_T, \cdots, X_{t+1}|Z_t, \theta)
\end{aligned}
$$

They are easy to see by checking the first term and induction relations:

$$
\begin{aligned}
\alpha_t(Z_t) &= \sum_{Z_{t-1}} \alpha_{t-1}P(Z_t|Z_{t-1})P(X_t|Z_{t-1},Z_t) \\
\beta_t(Z_t) &= \sum_{Z_{t+1}} \beta_{t+1}P(Z_{t+1}|Z_t)P(X_{t+1}|Z_t,Z_{t+1})
\end{aligned}
$$

We also tend to find ξ and γ that

$$\xi_t(Z_t, Z_{t+1}) = P(Z_t, Z_{t+1}|X_{1:T}, \theta)$$

$$\gamma_t(Z_t) = P(Z_t|X_{1:T}, \theta)$$

Clearly, γ can be acquired by summing over Z_{t+1} of ξ and

$$\begin{aligned}
\xi_t(Z_t, Z_{t+1}) =& P(Z_t, Z_{t+1}|X_{1:T}, \theta) \\
=& \frac{P(Z_t, Z_{t+1}, X_{1:T}|\theta)}{P(X_{1:T}|\theta)} \\
=& \frac{P(X_{1:t}, Z_t|\theta)P(Z_{t+1}|Z_t, \theta)P(X_{t+2:T}|Z_{t+1}, \theta)P(X_{t+1}|Z_t, Z_{t+1})}{P(X_{1:T}|\theta)} \\
=& \frac{\alpha_t(Z_t)T_{Z_t,Z_{t+1}}\beta_{t+1}(Z_{t+1})\phi(Z_t, Z_{t+1}, X_{t+1})}{\sum_{Z_t,Z_{t+1}} \alpha_t(Z_t)T_{Z_t,Z_{t+1}}\beta_{t+1}(Z_{t+1})\phi(Z_t, Z_{t+1}, X_{t+1})}
\end{aligned}$$

2 Maximization Part

Then we will use these terms for Baum-Welch algorithm. Recall that

$$Q(\theta, \theta_k) = \mathbb{E}_{\theta_k}(\log p(X, Z|\theta)|X)$$

thus followed by

$$\begin{aligned}
\log P(X, Z|\theta) =& \log p(Z_1|\theta) + \sum_{t=1}^{T-1} \log p(Z_{t+1}|Z_t, \theta) + \log p(X_1|Z_1, \theta) + \sum_{t=1}^{T-1} \log p(X_{t+1}|Z_{t+1}, Z_t, \theta) \\
=& \sum_{i=1}^{m} 1(Z_1 = i) \log \pi_i + \sum_{t=1}^{T-1} \sum_{i=1}^{m} \sum_{j=1}^{m} 1(Z_t = i, Z_{t+1} = j) \log T_{ij} \\
&+ \sum_{i=1}^{m} 1(Z_1 = i) \log \phi_0(i, X_1) + \sum_{t=1}^{T-1} \sum_{i=1}^{m} \sum_{j=1}^{m} 1(Z_t = i, Z_{t+1} = j) \log \phi_{ij}(X_{t+1})
\end{aligned}$$

Expectation of indicator is just probability of condition in it, thus

$$\begin{aligned}
Q(\theta, \theta_k) =& \sum_{i=1}^{m} P_{\theta_k}(Z_1 = i|X) \log \pi_i + \sum_{t=1}^{T-1} \sum_{i=1}^{m} \sum_{j=1}^{m} P_{\theta_k}(Z_t = i, Z_{t+1} = j|X) \log T_{ij} \\
&+ \sum_{i=1}^{m} P_{\theta_k}(Z_1 = i|X) \log \phi_0(i, X_1) + \sum_{t=1}^{T-1} \sum_{i=1}^{m} \sum_{j=1}^{m} P_{\theta_k}(Z_t = i, Z_{t+1} = j|X) \log \phi_{ij}(X_{t+1})
\end{aligned}$$

And by out definition, this is

$$Q(\theta, \theta_k) = \sum_{i=1}^{m} \gamma_{1i} \log \pi_i \phi_0(i, X_1) + \sum_{t=1}^{T-1} \sum_{i,j=1}^{m} \xi_{tij} \log T_{ij} + \sum_{t=1}^{T-1} \sum_{i,j=1}^{m} \xi_{tij} \log \phi_{ij}(X_{t+1})$$

By method of Lagrangian multipliers, we have

$$\begin{aligned}
\hat{\pi}_i =& \frac{\gamma_{1i}}{\sum_{j=1}^{m} \gamma_{1j}} = \gamma_{1i} \\
\hat{T}_{ij} =& \frac{\sum_{t=1}^{T-1} \xi_{tij}}{\sum_{t=1}^{T-1} \sum_{j=1}^{m} \xi_{tij}} = \frac{\sum_{t=1}^{T-1} \xi_{tij}}{\sum_{t=1}^{T-1} \gamma_{ti}} \\
\hat{\phi_{0in}} =& \begin{cases} 1 & n = X_1 \\ 0 & otherwise \end{cases} \\
\hat{\phi_{ijn}} =& \frac{\sum_{X_{t+1}=n} \xi_{ijt}}{\sum_{t=1}^{T-1} \xi_{ijt}}
\end{aligned}$$

Notice that optimization of ϕ_0 is rather simple, we can solve this by using multiple input.

3 Multiple Observation

When we have multiple observations, we tend to average them in some sense to get better approximation. We consider different way of optimizing following

$$
\begin{aligned}
\pi_i &= P\left(Z_1 = i\right) \\
b_{0in} &= P\left(X_1 = n | Z_1 = i\right) \\
b_{ijn} &= P\left(X_t = n | Z_{t-1} = i, Z_t = j\right) \\
t_{ij} &= P\left(Z_{t+1} = j | Z_t = i\right)
\end{aligned}
$$

For multiple input, we have

$$
\begin{aligned}
P(X|\theta) &= \prod_{k=1}^{K} P\left(X^k|\theta\right) \\
&= \prod_{k=1}^{K} P_k
\end{aligned}
$$

Since observations are independent, we have following result and by maximizing every single P_k, we get our multiple input approximation.

$$
\begin{aligned}
t_{ij} &= P\left(Z_{t+1} = j | Z_t = i, (X^1, \cdots, X^K)\right) \\
&= \frac{P\left(Z_{t+1} = j, Z_t = i | (X^1, \cdots, X^K)\right)}{P(Z_t = i | (X^1, \cdots, X^K))} \\
&= \frac{\sum_j P\left(Z_{t+1} = j, Z_t = i | X^j\right) P(X_j | (X^1, \cdots, X^K))}{\sum_j P(Z_t = i | X_j) P(X^j | (X^1, \cdots, X^K))} \\
&= \frac{\frac{1}{K} \sum_j P(Z_{t+1} = j, Z_t = i | X^j)}{\frac{1}{K} \sum_j P(Z_t = i | X^j)} \\
&= \frac{\sum_j P(Z_{t+1} = j, Z_t = i, X^j)/P(X^j)}{\sum_j P(Z_t = i, X^j)/P(X^j)}
\end{aligned}
$$

Thus

$$
\bar{t}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) t_{ij} b_{ij}\left(Z_{t+1}^{(k)}\right) \beta_j^k(Z_{t+1})}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) \beta_t^k(i)}
$$

and by same way

$$
\begin{aligned}
\overline{b_{ijn}} &= \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{X_{t+1} = n} \alpha_t^k(i) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) \beta_{t+1}^k(j)} \\
b_{0in} &= \frac{\sum_{X^k = n} 1}{K} \\
\pi_i &= \frac{\sum_k \pi_i^k}{K}
\end{aligned}
$$