Forward/Backward Learning of Hidden Markov Models with discrete multinomial observed nucleotides

I. The Model

We assume a K-state model for the system with a discrete emission alphabet Σ, where $M =\mid \Sigma \mid$ is the size of the alphabet. We define a $K \times K$ transition matrix A, such that $P(z_{t+1} = j|z_t = i) = A_{ji}$, where $\{z_1, ..., z_T\}$ is the hidden state sequence, and T is the length of the sequence. We also define a $K \times M$ emission matrix E, such that $P(X_t = m|z_t = i) = E_{im}$, where X_t is the emitted 'letter' at time point t.

Given the observed nucleotide sequence $\boldsymbol{X}$, our goal is to estimate the HMM parameters $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{\pi}\}$ using a maximum likelihood approach.

II. Problem Setup

In this section we will follow the Expectation-Maximization steps suggested by Bishop [1]. Let us maximize the probability $P(\boldsymbol{X}|\boldsymbol{\theta})$ of observing the nucleotide sequence $\boldsymbol{X}$:

$$P(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta}). \tag{1}$$

Defining the likelihood function as $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) = \log P(\boldsymbol{X}|\boldsymbol{\theta})$, we obtain:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) &= \log \sum_{\boldsymbol{z}} P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta}) \\ &= \log \sum_{\boldsymbol{z}} Q(\boldsymbol{z}) \frac{P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})}{Q(\boldsymbol{z})} \\ &= \log E_Q \left(\frac{P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})}{Q(\boldsymbol{z})} \right). \end{aligned} \tag{2}$$

Here $Q(\boldsymbol{z})$ is an arbitrary pdf on $\boldsymbol{z}$. Since there are exponentially many (K^T) possible paths $\boldsymbol{z}$, maximizing $P(\boldsymbol{X}|\boldsymbol{\theta})$ is not computationally pragmatic. We therefore define a new likelihood function $\tilde{\mathcal{L}}$ and use Jensen's inequality to obtain:

$$\mathcal{L} = \log E_Q(P/Q) \leq E_Q \log(P/Q) \equiv \tilde{\mathcal{L}} \tag{3}$$

We simplify the problem and find the model parameter $\hat{\boldsymbol{\theta}}$ that maximizes $\tilde{\mathcal{L}}$ instead:

$$\begin{aligned}\tilde{\mathcal{L}}(\boldsymbol{\theta}|\boldsymbol{X}) &= E_Q(\log P - \log Q) \\ &= \sum_{\boldsymbol{z}} Q(\boldsymbol{z})(\log P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta}) - \log Q(\boldsymbol{z}))\end{aligned} \tag{4}$$

If at the k^{th} step of the EM interations, the inferred model parameter is $\hat{\boldsymbol{\theta}}_k$, the maximization over $Q(\boldsymbol{z})$ implies

$$\hat{Q}_{k+1}(\boldsymbol{z}) = P(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}_k). \tag{5}$$

Therefore, the optimal $\boldsymbol{\theta}$ at the $(k+1)^{\text{th}}$ step will be:

$$\hat{\boldsymbol{\theta}}_{k+1} = \underset{\boldsymbol{\theta} \in \Omega}{\operatorname{argmax}} \sum_{\boldsymbol{z}} P(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}_k) \log P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta}). \tag{6}$$

Let's factorize $\log P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})$ using the Markov property of the system:

$$\begin{aligned}\log P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta}) &= \log \left(P(z_1|\boldsymbol{\pi}) \prod_{t=2}^{T} P(z_t|z_{t-1}, \boldsymbol{A}) \prod_{t=1}^{T} P(X_t|z_t, \boldsymbol{E}) \right) \\ &= \log \pi_{z_1} + \sum_{t=2}^{T} \log A_{z_t, z_{t-1}} + \sum_{t=1}^{T} \log E_{z_t, X_t}.\end{aligned} \tag{7}$$

We now introduce two notations:

$$\langle z_t^i \rangle := \sum_{\boldsymbol{z}} z_t^i P(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}), \tag{8}$$

$$\langle z_t^i, z_{t-1}^j \rangle := \sum_{\boldsymbol{z}} z_t^i z_{t-1}^j P(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}), \tag{9}$$

where $z_t^i := \delta_{z_t,i}$, $i \in \{1, ..., K\}$.

Applying these notations and the factorization of $\log P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})$, we obtain from Eq. (6):

$$\hat{\boldsymbol{\theta}}_{k+1} = \underset{\boldsymbol{\theta}\in\Omega}{\operatorname{argmax}} \left(\sum_{i=1}^{K} \langle z_1^i \rangle \log \pi_{z_1} + \sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} \langle z_t^i, z_{t-1}^j \rangle \log A_{z_t, z_{t-1}} + \sum_{t=1}^{T} \sum_{i=1}^{K} \langle z_t^i \rangle \log E_{z_t, X_t} \right). \tag{10}$$

III. Maximization

Maximization over the model parameters $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{\pi}\}$ yields:

$$\hat{\pi}_n = \frac{\langle z_1^n \rangle}{\sum_{i=1}^{K} \langle z_1^i \rangle}, \tag{11}$$

$$\hat{A}_{nk} = \frac{\sum_{t=2}^{T} \langle z_t^n, z_{t-1}^k \rangle}{\sum_{i=1}^{K} \sum_{t=2}^{T} \langle z_t^i, z_{t-1}^k \rangle}, \tag{12}$$

$$\hat{E}_{nm} = \frac{\sum_{t=1}^{T} \langle z_t^n \rangle \delta_{X_t, m}}{\sum_{t=1}^{T} \langle z_t^n \rangle}. \tag{13}$$

Here we used the following probability constraints:

$$\sum_{i=1}^{K} \pi_i = 1, \tag{14}$$

$$\sum_{i=1}^{K} A_{i,n} = 1, \quad (\text{for } \forall n), \tag{15}$$

$$\sum_{m=1}^{M} E_{n,m} = 1, \quad (\text{for } \forall n). \tag{16}$$

IV. Forward-Backward

Let's recall the definitions of $\langle z_t^i \rangle$ and $\langle z_t^i, z_{t-1}^j \rangle$, and use the rules of probability to obtain:

$$\langle z_t^i \rangle := \sum_{\boldsymbol{z}} z_t^i P(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) \equiv \sum_{z_t} z_t^i P(z_t|\boldsymbol{X}, \hat{\boldsymbol{\theta}}), \tag{17}$$

$$\langle z_t^i, z_{t-1}^j \rangle := \sum_{\boldsymbol{z}} z_t^i z_{t-1}^j P(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) \equiv \sum_{z_{t-1}, z_t} P(z_{t-1}, z_t|\boldsymbol{X}, \hat{\boldsymbol{\theta}}). \tag{18}$$

Thus, we need $P(z_t|\boldsymbol{X}, \hat{\boldsymbol{\theta}})$ and $P(z_{t-1}, z_t|\boldsymbol{X}, \hat{\boldsymbol{\theta}})$ in order to obtain $\langle z_t^i \rangle$ and $\langle z_t^i, z_{t-1}^j \rangle$. Using the sum and product rules of probability we can see that

$$P(z_t|\boldsymbol{X}) = \frac{P(X_{1...t}, z_t) P(X_{t+1...T}|z_t)}{P(\boldsymbol{X})}, \tag{19}$$

$$P(z_{t-1}, z_t|\boldsymbol{X}) = \frac{P(X_{1...t-1}, z_{t-1}) P(X_t|z_t) P(z_t|z_{t-1}) P(X_{t+1...T}|z_t))}{P(\boldsymbol{X})}. \tag{20}$$

Let's introduce forward/backward coefficients:

$$\alpha(z_t) := P(X_{1...t}, z_t), \tag{21}$$

$$\beta(z_t) := P(X_{t+1...T}|z_t). \tag{22}$$

Using these coefficients, we obtain:

$$P(z_t|\boldsymbol{X}) = \frac{\alpha(z_t)\beta(z_t)}{P(\boldsymbol{X})}, \tag{23}$$

$$P(z_{t-1}, z_t|\boldsymbol{X}) = \frac{\alpha(z_{t-1}) \hat{E}_{z_t, X_t} \hat{A}_{z_t, z_{t-1}} \beta(z_{t+1})}{P(\boldsymbol{X})}. \tag{24}$$

Applying the rules of probability, we find a recursive relation for the α and β coefficients [1]:

$$\alpha(z_t) = P(X_t|z_t) \sum_{z_{t-1}} \alpha(z_{t-1}) P(z_t|z_{t-1}) = E_{z_t,X_t} \sum_{z_{t-1}} \hat{A}_{z_t,z_{t-1}} \alpha(z_{t-1}), \tag{25}$$

$$\beta(z_t) = \sum_{z_{t+1}} \beta(z_{t+1}) P(X_{t+1}|z_{t+1}) P(z_{t+1}|z_t) = \sum_{z_{t+1}} \hat{E}_{z_{t+1},X_{t+1}} \hat{A}_{z_{t+1},z_t} \beta(z_{t+1}). \tag{26}$$

The corresponding boundary conditions are:

$$\alpha(z_1) = P(X_1, z_1) = P(X_1|z_1)P(z_1) = \hat{E}_{z_1,X_1} \pi_{z_1}, \tag{27}$$

$$\beta(z_T) = 1. \tag{28}$$

Finally, we use $\alpha(z_T)$ to calculate $P(\boldsymbol{X})$:

$$1 = \sum_{z_T} P(z_T|\boldsymbol{X}) = \sum_{z_T} \frac{\alpha(z_T)\beta(z_T)}{P(\boldsymbol{X})} \equiv \frac{\sum_{z_T} \alpha(z_T)}{P(\boldsymbol{X})} \tag{29}$$

$$\Rightarrow P(\boldsymbol{X}) = \sum_{z_T} \alpha(z_T). \tag{30}$$

REFERENCES

[1] Christopher Bishop, *Pattern Recognition and Machine Learning* (Springer, 2007).