1. HMM Model

Given by a group of states $\mathbf{x} = \{1, ..., K\}$ and a group of observances or continuous observance $\mathbf{y}$, the $\mathbf{x}$ has Markov property and $\mathbf{y}$ is observed under some probability distribution that conditioned on state $\mathbf{x}$.

Denoting some symbols as followed:

N : the length of samples, or, the length of observe sequence.

n : time index, which is positive integer.

x_n , y_n : the state and observance at n th sampling time, specifically, x_1 and y_1 indicate the initial state and the first observance respectively.

$a_{lk} = p(x_n = k \mid x_{n-1} = l)$: the transfer probability from state $x_{n-1} = l$ to $x_n = k$, $0 \le l, k \le K$, and because of the Markov property, a_{lk} has no relation to the state before time $n-1$, i.e. the probability distribution of state x_n is only based on x_{n-1} .

$A = \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \cdots & a_{KK} \end{bmatrix}$: state transfer matrix.

x_1^N, y_1^N : the whole sequence of states the observances.

x_a^b, y_a^b : the states and observances from time a to b , where $a < b$, $1 \le a \le N$, $1 \le b \le N$.

2. **Forward & Backward Procedure**

First, we should initialize some parameters. We can use some simple classifier (like Kmeans) to separate the samples into some groups, i.e. the number of states K , then we could have the initial state probability and transfer matrix, defined followed:

$$\pi(k) = \frac{number\ of\ \ state\ k}{N} \quad 1 \le k \le K$$

$$a_{lk} = \frac{number\ of\ \ state\ \ from\ l\ to\ k}{number\ of\ \ state\ l}$$

Then we set:

$$
\begin{aligned}
S_n &= p(y_n | \mathrm{y}_1^{n-1}) \\
&= \sum_{i=1}^{K} \left[p\left(y_n | x_n = i\right) p\left(x_n = i | y_1^{n-1}\right) \right] \\
&= \sum_{i=1}^{K} \left[p\left(y_n | x_n = i\right) \sum_{k=1}^{K} \left[p\left(x_{n-1} = k | y_1^{n-1}\right) a_{ki} \right] \right]
\end{aligned}
$$

Where $n > 1$.

2.1 Forward Procedure

Define

$$
\begin{aligned}
\alpha_n(k) &= p(x_n = k | y_1^n) \\
&= \frac{p(x_n = k, y_n | y_1^{n-1})}{S_n}
\end{aligned}
$$

- Initialization

$$
\alpha_1(k) = \frac{p(\mathrm{x}_1 = k) p(y_1 | \mathrm{x}_1 = k)}{\sum_{k=1}^{K} p(\mathrm{x}_1 = k) p(y_1 | \mathrm{x}_1 = k)} = \frac{\pi(k) p(y_1 | \mathrm{x}_1 = k)}{\sum_{k=1}^{K} \pi(k) p(y_1 | \mathrm{x}_1 = k)}
$$

- Propagate

$$
\alpha_{n+1}(k) = \frac{p(y_{n+1} | x_{n+1} = k) \sum_{l=1}^{K} a_{lk} \alpha_n(l)}{S_{n+1}}
$$

2.2 Backward Procedure

Define

$$
\begin{aligned}
\beta_n(k) &= \frac{p(y_{n+1}^N | x_n = k)}{p(y_{n+1}^N | y_1^n)} \\
&= \frac{1}{S_{n+1}} \bullet \frac{p(y_{n+1}^N | x_n = k)}{p(y_{n+2}^N | y_1^{n+1})}
\end{aligned}
$$

- Initialization

$$
\beta_N(k) = 1
$$

- Propagate

$$\beta_n(k) = \frac{\sum_{l=1}^{K} a_{kl}\, p(y_{n+1} | \mathrm{x}_{n+1} = l)\beta_{n+1}(l)}{S_{n+1}}$$

2.3 Posterior Update

$$\gamma_n(k) = p(x_n = k | y_1^N) = \alpha_n(k)\beta_n(k)$$

$$\xi_n(l,k) = p(x_n = l, x_{n+1} = k | y_1^N) = \frac{\alpha_n(l)\beta_{n+1}(k)a_{lk}\, p(y_{n+1} | x_n = l)}{S_{n+1}}$$

$$a_{lk}^* = \frac{\sum_{n=1}^{N-1} \xi_n(l,k)}{\sum_{n=1}^{N-1} \gamma_n(l)}$$

$$\pi^*(k) = \frac{\gamma_1^{(l)}(k)}{\sum_{k=1}^{K} \gamma_1^{(l)}(k)}$$

Then let $a_{lk} = a_{lk}^*$ and $\pi(k) = \pi^*(k)$.

3. **EM Algorithm Used in HMM Situation**

3.1 Expectation Step

$$Q\left(\Theta | \Theta^{(l)}\right) = E\left[\ln p\left(y_1^N, x_1^N \middle| \Theta\right) \middle| y_1^N, \Theta^{(l)}\right]$$

Where

$$\begin{aligned}\ln p\left(y_1^N, x_1^N \middle| \Theta\right) &= \ln\left[\pi(x_1) f_{x_1}(y_1) \prod_{n=2}^{N} a_{x_{n-1}x_n} f_{x_n}(y_n)\right] \\ &= \ln\left(\pi(x_1)\right) + \sum_{n=1}^{N} \ln\left(f_{x_n}(y_n)\right) + \sum_{n=2}^{N} \ln\left(a_{x_{n-1}x_n}\right)\end{aligned}$$

So $Q\left(\Theta | \Theta^{(l)}\right)$ can be derived:

$$Q\left(\Theta | \Theta^{(l)}\right) = \sum_{k=1}^{K} \gamma_1^{(l)}(k) \ln\left(\pi(k)\right) + \sum_{k=1}^{K}\sum_{n=1}^{N} \gamma_n^{(l)}(k) \ln\left(f_k(y_n)\right) + \sum_{k=1}^{K}\sum_{i=1}^{K}\sum_{n=1}^{N-1} \xi_n^{(l)}(k,i) \ln\left(a_{ki}\right)$$

3.2 Maximization Step

To get the appropriate $\Theta^{(l+1)}$ that maximize $Q\left(\Theta \mid \Theta^{(l)}\right)$, which can be written as:

$$\Theta^{(l+1)} = \arg\max_{\Theta} Q\left(\Theta \mid \Theta^{(l)}\right)$$

We can use partial differential to find the extreme point.

- Mean value $\mu_k^{(l+1)}$

$$\frac{\partial Q\left(\Theta \mid \Theta^{(l)}\right)}{\partial \mu_k^{(l+1)}} = 0$$

$$\mu_k^{(l+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(l)}(k) y_n}{\sum_{n=1}^{N} \gamma_n^{(l)}(k)}$$

or

$$\boldsymbol{\mu}^{(l+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(l)}(k) \mathbf{y}_n}{\sum_{n=1}^{N} \gamma_n^{(l)}(k)}$$

- Variance $\sigma_k^{2\,(l+1)}$

$$\frac{\partial Q\left(\Theta \mid \Theta^{(l)}\right)}{\partial \sigma_k^{2\,(l+1)}} = 0$$

$$\sigma_k^{2\,(l+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(l)}(k) \left(y_n - \mu_k^{(l+1)}\right)^2}{\sum_{n=1}^{N} \gamma_n^{(l)}(k)}$$

or

$$\frac{\partial Q\left(\Theta \mid \Theta^{(l)}\right)}{\partial \Gamma^{2\,(l+1)}} = 0$$

$$\Sigma^{(l+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(l)}(k) \left(\mathbf{y}_n - \boldsymbol{\mu}^{(l+1)}\right)\left(\mathbf{y}_n - \boldsymbol{\mu}^{(l+1)}\right)^T}{\sum_{n=1}^{N} \gamma_n^{(l)}(k)}$$

- Initial Probability $\pi^{(l+1)}(k)$

Because there is a constraint for the Initial probability:

$$\sum_{k=1}^{K} \pi^{(l+1)}(k) = 1$$

So we introduce a Lagrange multiplier λ into the expression to cope with this situation:

$$L\left(\Theta, \lambda \middle| \Theta^{(l)}\right) = Q\left(\Theta \middle| \Theta^{(l)}\right) + \left(1 - \sum_{k=1}^{K} \pi^{(l+1)}(k)\right) \bullet \lambda$$

Then there is a new relation for $L\left(\Theta, \lambda \middle| \Theta^{(l)}\right)$,

$$\frac{\partial L\left(\Theta, \lambda \middle| \Theta^{(l)}\right)}{\partial \pi^{(l+1)}(k)} = 0$$

$$\pi^{(l+1)}(k) = \frac{1}{\lambda} \gamma_1^{(l)}(k)$$

$$\because \quad \sum_{k=1}^{K} \pi^{(l+1)}(k) = 1$$

$$\therefore \quad \lambda = \sum_{k=1}^{K} \gamma_1^{(l)}(k)$$

At last, we could derive:

$$\pi^{(l+1)}(k) = \frac{\gamma_1^{(l)}(k)}{\sum_{k=1}^{K} \gamma_1^{(l)}(k)}$$

- State transfer probability $a_{ki}{}^{(l+1)}$

Because there is a constraint for the Initial probability:

$$\forall k, \qquad \sum_{i=1}^{K} a_{ki}{}^{(l+1)} = 1$$

So we introduce a Lagrange multiplier λ into the expression to cope with this situation:

$$L\left(\Theta, \lambda \middle| \Theta^{(l)}\right) = Q\left(\Theta \middle| \Theta^{(l)}\right) + \left(1 - \sum_{i=1}^{K} a_{ki}{}^{(l+1)}\right) \bullet \lambda$$

Then we have

$$\frac{\partial L\left(\Theta, \lambda \middle| \Theta^{(l)}\right)}{\partial a_{ki}} = 0$$

$$a_{ki} = \frac{1}{\lambda} \sum_{n=1}^{N-1} \xi_n^{(l)}(k, i)$$

$$\because \quad \sum_{i=1}^{K} a_{ki} = 1$$

$$\therefore \quad \lambda = \sum_{n=1}^{N-1} \sum_{i=1}^{K} \xi_n^{(l)}(k,i) = \sum_{n=1}^{N-1} \gamma_n^{(l)}(k)$$

At last, we could derive:

$$a_{ki} = \frac{\sum_{n=1}^{N-1} \xi_n^{(l)}(k,i)}{\sum_{n=1}^{N-1} \gamma_n^{(l)}(k)}$$