1. HMM Model

Given by a group of states $\mathbf{x} = \{0, 1, ..., k-1\}$ 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 nth sampling time, specifically, x_1 and y_1 indicate the initial state and the first observance respectively.

$P_{ij} = P\{x_n = j \mid x_{n-1} = i\}$: the transfer probability from state $x_{n-1} = i$ to $x_n = j$, $0 \le i, j \le k-1$, and because of the Markov property, P_{ij} 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}.

$\mathbf{P} = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}$: state transfer matrix.

$\mathbf{X}_1^N, \mathbf{Y}_1^N$: the whole sequence of states the observances.

$\mathbf{X}_a^b, \mathbf{Y}_a^b$: the states and observances from time a to b, where $a < b$, $1 \le a \le N$, $1 \le b \le N$.

If the observance probability distribution is discrete, define matrix:

$$\mathbf{B}_{n \times m} = \{b_{ij}\}, i = 0, ..., k-1, \ j = 0, ..., m-1$$

where m is the number of distinct observation symbols per state, i.e., the discrete alphabet size. The components $b_{ij} = P\{y_n = j \mid x_n = i\}$ are the probability for each observe y_n given by the state x_n.

If the observance probability distribution is continuous, probability density function (pdf) should be introduced, the most general way is using Gaussian Mixture Model (GMM). Then the pdf of each observance condition on specific state can be defined as:

$$P\{y_n \mid x_n = i\} = \sum_{j=1}^{m} c_{ij} f(y, \mu_{ij}, \sigma_{ij}) \qquad i = 0, ..., k-1$$

Where $f(\bullet)$ is a Gaussian pdf, μ_{ij} and σ_{ij} represent the mean and variance for the kth mixture component in state i, m is the number of mixture components for each state. The $f(\bullet)$ can be multi-dimensional, thus, μ_{ij} will be a vector and σ_{ij} a matrix.

2. Baum Welch Algorithm

Baum–Welch algorithm is used to find the unknown parameters of a hidden Markov model (HMM). It makes use of the forward-backward algorithm and is named for Leonard E. Baum and Lloyd R. Welch. (Wikipedia, https://en.wikipedia.org/wiki/Baum-Welch_algorithm)

2.1 Useful Denotes

We introduce two important terms into the algorithm, $\gamma_n(i)$ the probability of state equals to i at time n given by the observance sequence Y_1^N, and $\xi_n(i,j)$ the probability of state equals to i at time n and state equals to j at time $n+1$ given by observe sequence Y_1^N. These two terms can be depicted as follows:

$$\gamma_n(i) = P\left\{x_n = i \mid \mathrm{Y}_1^N\right\}$$

$$\xi_n(i,j) = P\left\{x_n = i, x_{n+1} = j \mid \mathrm{Y}_1^N\right\}$$

Also we define some symbols for a simplicity of representation.

$\boldsymbol{\pi} = \left(P\{\mathrm{x}_1 = 0\}, P\{\mathrm{x}_1 = 1\}, \ldots, P\{\mathrm{x}_1 = k-1\}\right)$: initial probability distribution of states.

$\alpha_n(i) = P\{\mathrm{x}_n \mid \mathrm{Y}_1^n\}$: forward result at each sample time n for each state i.

2.2 Baum Welch Algorithm

2.2.1 Forward Procedure

Forward procedure propagates to get from α_1 to α_N, the steps as followed:

1. Initialization

Set $n=1$, calculate each $\alpha_n(i) = \alpha_1(i)$ by the formula:

$$\alpha_1(i) = P\{\mathrm{x}_n \mid \mathrm{Y}_1^1\} = \frac{P\{y_1 \mid \mathrm{x}_1 = i\}\boldsymbol{\pi}(i)}{\sum_{i=0}^{k-1} P\{y_1 \mid \mathrm{x}_1 = i\}\boldsymbol{\pi}(i)} \quad (1)$$

2. Propagate

Set $n=2$,calculate each $\alpha_n(i)$ by the formulas:

$$P\{\mathrm{x}_n = j \mid \mathrm{Y}_1^{n-1}\} = \sum_{i=0}^{k-1} P\{\mathrm{x}_n = j \mid x_{\mathrm{n}-1} = i\} \cdot P\{\mathrm{x}_{n-1} \mid \mathrm{Y}_1^{n-1}\} \\ = \sum_{i=0}^{k-1} P_{ij} \cdot \alpha_{n-1}(i) \tag{2}$$

$$\alpha_n(i) = \frac{P\{\mathrm{y}_n \mid \mathrm{x}_n = i\} \cdot P\{\mathrm{x}_n = i \mid \mathrm{Y}_1^{n-1}\}}{\sum_{i=0}^{k-1} P\{\mathrm{y}_n \mid \mathrm{x}_n = i\} \cdot P\{\mathrm{x}_n = i \mid \mathrm{Y}_1^{n-1}\}} \tag{3}$$

Where $0 \le i, j \le k-1$. Then making $n = n+1$, redo the equations (2) and (3), and stop when $n > N$.

2.2.2 Backward Procedure

By preceding backward procedure, we can get $\gamma_N(i)$ to $\gamma_1(i)$, and $\xi_{N-1}(i, j)$ to $\xi_1(i, j)$. In order to calculate the $\gamma_n(i)$, we could have the equations as followed:

$$\gamma_n(i) = P\{\mathrm{x}_n = i \mid Y_1^N\} \\ = \sum_{j=0}^{k-1} P\{\mathrm{x}_n = i \mid x_{n+1} = j, \mathrm{Y}_1^N\} P\{\mathrm{x}_{n+1} = j \mid \mathrm{Y}_1^N\} \tag{4}$$

$$P\left\{\mathrm{x}_n = i \mid x_{n+1} = j, Y_1^N\right\} = P\left\{\mathrm{x}_n = i \mid x_{n+1} = j, Y_1^n\right\} \\ = \sum_{j=1}^{k-1} \frac{P\left\{x_n = i \mid Y_1^n\right\} P\left\{x_{n+1} = j \mid x_n = i\right\}}{P\left\{x_{n+1} = j \mid Y_1^n\right\}} \\ = \sum_{j=1}^{k-1} \frac{\alpha_n(i) P_{ij}}{P\left\{x_{n+1} = j \mid Y_1^n\right\}} \tag{5}$$

Substitute (2), (3) and (5) in (4), we could have:

$$\gamma_n(i) = \sum_{j=1}^{k-1} \frac{\alpha_n(i) \cdot P_{ij} \cdot \gamma_{n+1}(i)}{P\{\mathrm{x}_{n+1} = j \mid Y_1^n\}} \tag{6}$$

And also, we could get the term $\xi_n(i, j)$

$$\xi_n(i, j) = \frac{\alpha_n(i) \cdot P_{ij} \cdot \gamma_{n+1}(i)}{P\{\mathrm{x}_{n+1} = j \mid \mathrm{Y}_1^n\}} \tag{7}$$

1. Initialization

Set each $\gamma_N(i)$ equals to $\alpha_N(i)$,

$$\gamma_N(i) = \alpha_N(i), \quad i = 0, \ldots, \mathrm{k}-1 \tag{8}$$

2. Propagate

Set $n = N-1$, calculate each $\gamma_n(i)$ and $\xi_n(i, j)$ for $0 \le i, j \le k-1$ by equations (6) and (7). And then make $n = n-1$, propagate until $n = 1$.

3. Reference

1. L.R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". Proceedings of the IEEE, Vol.77, Issue.2, pp.257-286, 1989
2.