Hidden Markov Model - I

Markov Chain

Stochastic Processes

A *stochastic process*, say S_t, is a collection of related random variables indexed by t. S_t and t can be discrete or continuous. Moreover, they can be vector-valued. In our discussion, we assume S_t and t are discrete unless otherwise specified.

Markov Chain

The *Markov chain* makes the assumption that

$$Pr(S_t|S_{t-1}, S_{t-2}, \ldots, S_1) = Pr(S_t|S_{t-1}) \tag{1}$$

Under this assumption, the probability of a state sequence is

$$Pr(S_1, \ldots, S_T) = \prod_{t=1}^{T} Pr(S_t|S_{1:t-1}) = Pr(S_1)\prod_{t=2}^{T} Pr(S_t|S_{t-1}) \tag{2}$$

State Space and State Transition Probability

We can define the *state space*,

$$\Omega = \{1, 2, \ldots, N\} \tag{3}$$

for the values for S_t. The probability

$$Pr(S_t = j|S_{t-1} = i) = a_{ij}(t), \quad i, j = 1, \ldots, N \tag{4}$$

is called the *state transition probability*, and

$$Pr(S_1 = i) = \pi_i, \quad i = 1, \ldots, N \tag{5}$$

is called the *initial probability*.

Time-Invariant Markov Chain

A Markov chain is *time-invariant* if the transition probabilities do not vary with time, i.e.,

$$a_{ij}(t) = a_{ij} \tag{6}$$

An Example of Markov Chain

Consider the stock market, DJI. Let the state space be

$$\mathcal{X} = \{1 = \text{up}, 2 = \text{down}, 3 = \text{flat}\}. \tag{7}$$

Assume that the transition probability and the initial probability are

$$A = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}, \quad \pi = \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix} \tag{8}$$

What is the probability that DJI is up for the first 5 days?

Hidden Markov Models

A sequence of observations may not be modeled well by a Markov chain. In some cases, we can assume that there is an *unobserved* Markov chain, and the observed sequence depends probabilistically on the hidden state sequence. Such a model is called a hidden Markov model (HMM).

State Emitting Probability

In addition to the A and π as needed in a Markov chain, the characterization of an HMM also includes the distribution of observation given the Markov state, which is called the *state emitting probability*.

Parameter Set

Let O be an observation and S be the hidden state. Let the *state-emitting probability* be denoted by

$$b_i(\mathbf{k}) = p(O = \mathbf{k}|S = i). \tag{9}$$

The parameters in the probability models of an HMM can be represented by

$$\lambda = (\pi, A, B). \tag{10}$$

Example: Coin Toss

Suppose there are a number of coins $1, \ldots, N$, each with its own bias. A coin is randomly selected is flipped, and the outcome (but not the coin) is recorded. This is repeated for T times. What is the probability of recording $(o_1, \ldots, o_T)$?

$$\begin{aligned} Pr(o_1, \ldots, o_T) &= \sum_{c_1, \ldots, c_T} Pr(o_1, \ldots, o_T, c_1, \ldots, c_T) \\ &= \sum_{c_1, \ldots, c_T} Pr(o_1, \ldots, o_T|c_1, \ldots, c_T)Pr(c_1, \ldots, c_T). \end{aligned} \tag{11}$$

Suppose the (un-observed) coin sequence is a Markov chain, then we have

$$Pr(o_1, \ldots, o_T) = \sum_{c_1, \ldots, c_T} \left(\prod_t Pr(o_t|c_t)\right)\left(Pr(c_1)\prod_t Pr(c_t|c_{t-1})\right) \tag{12}$$

Another Example: Urns and Balls

Suppose there are N urns. At turn t, one of the urns, urn U_t, is randomly selected and a ball B_t is randomly selected from U_t, and the color C_t is recorded. Suppose the selection of U_t is dependent on U_{t-1}. Then the observed color sequence is characterized by a HMM.

Fundamental Problems in HMM

Evaluation Problem

Given the observations $\mathbf{o}$ and the model set λ of an HMM $\mathcal{H}$, evaluate

$$p(\mathbf{o}|\mathcal{H}) \tag{14}$$

A brute-force summation over all possible hidden state sequences give

$$\begin{aligned} p(\mathbf{o}|\lambda) &= \sum_{\mathbf{s}} p(\mathbf{o}, \mathbf{S} = \mathbf{s}|\lambda) \\ &= \sum_{\mathbf{s}} p(\mathbf{o}|\mathbf{S} = \mathbf{s}, \lambda)p(\mathbf{S} = \mathbf{s}|\lambda) \\ &= \sum_{s_1, \ldots, s_T} p(s_1)p(o_1|s_1)\prod_{t=2}^{T} p(o_t|s_t)a_{s_{t-1}s_t} \end{aligned} \tag{15}$$

with a time complexity of $O(TN^T)$. With a better algorithm, the complexity can be reduced to $O(TN^2)$

Decoding Problem

Given $\mathbf{o}$ and λ, determine the optimal state sequence $\mathbf{s}^*$

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} p(\mathbf{S} = \mathbf{s}|\mathbf{o}, \mathcal{H}) = \arg\max_{\mathbf{s}} p(\mathbf{o}, \mathbf{S} = \mathbf{s}|\mathcal{H}). \tag{16}$$

Let the partial state sequence be denoted by $\mathbf{s}_t = s_1, \ldots, s_t$, and define

$$\delta_t(t) \triangleq \max_{\mathbf{s}_{t-1}} p(\mathbf{s}_{t-1}, S_t = i, \mathbf{o}_t) \tag{17}$$

We have

$$\begin{aligned} \delta_j(t+1) &= \max_{\mathbf{s}_t} p(\mathbf{s}_t, S_{t+1} = j, \mathbf{o}_t, o_{t+1}) \\ &= \max_{\mathbf{s}_t} p(\mathbf{s}_t, \mathbf{o}_t)\, p(o_{t+1}, S_{t+1} = j|\mathbf{s}_t, \mathbf{o}_t) \\ &= \max_{\mathbf{s}_{t-1}} \max_{s_t} p(\mathbf{s}_{t-1}, S_t = i, \mathbf{o}_t)\, a_{ij}\, p(o_{t+1}|S_{t+1} = j) \\ &= \max_i\, \delta_i(t)\, a_{ij}\, b_j(o_{t+1}) \end{aligned} \tag{18}$$

Estimation Problem

Given $\mathbf{o}$, estimate the parameter set λ^*

$$\lambda^* = \arg\max_{\lambda} f(\mathbf{o}, \lambda). \tag{19}$$

Forward-Backward Algorithm

Forward Probability

The forward probability is defined as

$$\alpha_i(t) \triangleq p(o_1, \ldots, o_t, S_t = i) \tag{20}$$

We artificially introduce an initial non-emitting state 1 for time 1^-, and a final non-emitting state N for time T^+. Let

$$\pi_i = a_{1i}. \tag{21}$$

The following relation holds for $\alpha_i(t)$:

$$\begin{aligned} \alpha_j(1) &= p(o_1, S_1 = j, S_1 = 1) \\ &= a_{1j}b_j(o_1), \quad j = 2, \ldots, N-1 \\ \alpha_j(t) &= p(o_1, \ldots, o_t, S_t = j) \\ &= \sum_{i=2}^{N-1} p(o_1, \ldots, o_t, S_{t-1} = i, S_t = j) \\ &= \sum_{i=2}^{N-1} p(o_1, \ldots, o_{t-1}, S_{t-1} = i)p(o_t, S_t = j|S_{t-1} = i) \\ &= \sum_{i=2}^{N-1} \alpha_i(t-1)a_{ij}b_j(o_t), \quad j = 2, \ldots, N-1 \\ \alpha_N(T) &= p(o_1, \ldots, o_T, S_{T^+} = N) \\ &= \sum_{i=2}^{N-1} p(o_1, \ldots, o_T, S_T = i, S_{T^+} = N) \\ &= \sum_{i=2}^{N-1} \alpha_i(T)a_{iN} \end{aligned} \tag{22}$$

Since $p(\mathbf{o}) = \alpha_N(T)$, one can compute $p(\mathbf{o})$ with $O(TN^2)$ time complexity.

Backward Probability

The backward probability is defined as

$$\beta_i(t) \triangleq p(o_{t+1}, \ldots, o_T|S_t = i) \tag{23}$$

The following relation holds for $\beta_i(t)$:

$$\begin{aligned} \beta_i(T) &= p(S_T = i|S_T = i) \\ &= a_{iN}, \quad i = 2, \ldots, N-1 \\ \beta_i(t) &= p(o_{t+1}, \ldots, o_T|S_t = i) \\ &= \sum_{j=2}^{N-1} p(o_{t+1}, \ldots, o_T, S_{t+1} = j|S_t = i) \\ &= \sum_{j=2}^{N-1} p(o_{t+1}, \ldots, o_T|S_t = i, S_{t+1} = j)p(S_{t+1} = j|S_t = i) \\ &= \sum_{j=2}^{N-1} p(o_{t+2}, \ldots, o_T|S_{t+1} = j)p(o_{t+1}|S_{t+1} = j)p(S_{t+1} = j|S_t = i) \\ &= \sum_{j=2}^{N-1} a_{ij}b_j(o_{t+1})\beta_j(t+1); \\ \beta_1(1^-) &= p(o_1, \ldots, o_T|S_1 = 1) \\ &= \sum_{j=2}^{N-1} p(o_1, \ldots, o_T, S_1 = j|S_1 = 1) \\ &= \sum_{j=2}^{N-1} p(S_1 = j|S_1 = 1)p(o_1, \ldots, o_T|S_1 = j) \\ &= \sum_{j=2}^{N-1} p(S_1 = j|S_1 = 1)p(o_1|S_1 = j)p(o_2, \ldots, o_T|S_1 = j) \\ &= \sum_{j=2}^{N-1} a_{1j}b_j(o_1)\beta_j(1). \end{aligned} \tag{24}$$

Observation Data Likelihood

We have for any time t and state i

$$\begin{aligned} p(S_t = i, \mathbf{o}) &= p(o_1, \ldots, o_t, o_{t+1}, \ldots, o_T, S_t = i) \\ &= p(o_{1:t}, S_t = i)p(o_{t+1:T}|S_t = i, o_{1:t}) \\ &= p(o_{1:t}, S_t = i)p(o_{t+1:T}|S_t = i) \\ &= \alpha_i(t)\beta_i(t) \end{aligned} \tag{25}$$

It follows that the data likelihood can be computed by

$$p(\mathbf{o}) = \sum_i p(S_t = i, \mathbf{o}) = \sum_i \alpha_i(t)\beta_i(t) \tag{26}$$

States at Consecutive Times

The joint probability for the observation sequence and two states of consecutive times is given by

$$\begin{aligned} &p(S_t = i, S_{t+1} = j, \mathbf{o}) \\ &= p(o_1, \ldots, o_t, o_{t+1}, \ldots, o_T, S_t = i, S_{t+1} = j) \\ &= p(o_{1:t}, S_t = i)p(o_{t+1:T}, S_{t+1} = j|S_t = i, o_{1:t}) \\ &= p(o_{1:t}, S_t = i)p(S_{t+1} = j|S_t = i, o_{1:t})p(o_{t+1:T}|S_{t+1} = j, S_t = i, o_{1:t}) \\ &= p(o_{1:t}, S_t = i)p(S_{t+1} = j|S_t = i)p(o_{t+1:T}|S_{t+1} = j)p(o_{t+2:T}|S_{t+1} = j, o_{t+1}) \\ &= p(o_{1:t}, S_t = i)p(S_{t+1} = j|S_t = i)p(o_{t+1}|S_{t+1} = j)p(o_{t+2:T}|S_{t+1} = j) \\ &= \alpha_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1) \end{aligned} \tag{27}$$