Chapter 4
Hidden Markov Models (HMMs)

Qing Zhou

UCLA Department of Statistics

Stats 201C Advanced Modeling and Inference
Lecture Notes

Outline

1. Elements of an HMM
2. MLE via EM
3. The Viterbi algorithm
4. Extensions

Elements of an HMM

Example (coin toss): two coins.
Coin 1: unbiased coin, $P_H = P_T = 0.5$.
Coin 2: biased coin, $P_H = 0.9$ and $P_T = 0.1$.
Switch between 1 and 2 via a *hidden* Markov chain with transition matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

Observed	Y:	H	T	T	T	H	H	H	T	H	H	$\cdots$
Hidden	Z:	1	1	1	2	2	2	2	1	1	2	$\cdots$

Elements of an HMM

Elements

- Hidden states $\{1, \ldots, N\}$: state space for Z_t.
- Observed symbols $\{1, \ldots, M\}$: space for Y_t.
- State transition matrix $A = (a_{ij})_{N \times N}$,

$$a_{ij} = \mathbb{P}(Z_{t+1} = j \mid Z_t = i).$$

- Emission probabilities $B = (b_j(k))$,

$$b_j(k) = \mathbb{P}(Y_t = k \mid Z_t = j).$$

- Initial state distribution $\pi = (\pi_1, \ldots, \pi_N)$: $\mathbb{P}(Z_1 = j) = \pi_j$.

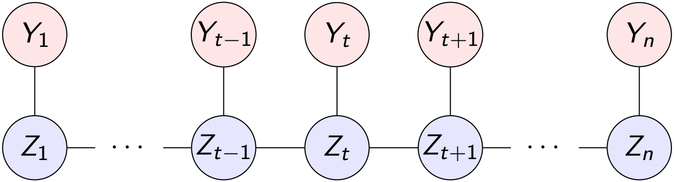
Elements of an HMM

Joint probability:

$$\begin{aligned}
\mathbb{P}(Y, Z) &= \mathbb{P}(Z_1)\mathbb{P}(Y_1 \mid Z_1)\mathbb{P}(Z_2 \mid Z_1)\mathbb{P}(Y_2 \mid Z_2) \\
&\qquad \cdots \mathbb{P}(Z_n \mid Z_{n-1})\mathbb{P}(Y_n \mid Z_n) \\
&= \mathbb{P}(Z_1)\mathbb{P}(Y_1 \mid Z_1) \prod_{t=2}^{n} \mathbb{P}(Z_t \mid Z_{t-1})\mathbb{P}(Y_t \mid Z_t) \\
&:= f_1(Z_1, Y_1) \prod_{t=2}^{n} g_t(Z_{t-1}, Z_t) f_t(Z_t, Y_t) \qquad (1)
\end{aligned}$$

Elements of an HMM

Graphical model for HMM: $\{(Z_t, Y_t) : t = 1, \ldots, n\}$.



Conditional independence:

- Undirected graph with each node representing a random variable V_i. If node j separates nodes i and k then

$$V_i \perp V_k \mid V_j.$$

- For HMM, V_{t-i}, Y_t and V_{t+j} are mutually independent conditional on Z_t, here V_k can be either Y_k or Z_k.

Elements of an HMM

Two basic problems to solve:

- Given Y, how to estimate model parameters $\theta = (A, B)$?
- Given Y and model parameter θ (or $\hat{\theta}$), how to predict hidden states Z?

MLE via EM

Problem setup:

- Observed data Y, missing data Z.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathbb{P}(Y; \theta) = \underset{\theta}{\operatorname{argmax}} \sum_{Z_1} \cdots \sum_{Z_n} \mathbb{P}(Y, Z_1, \ldots, Z_n; \theta)$$

- $(Z_1, \ldots, Z_n)$ given Y is a Markov chain. Regarding Y_t as constants in (1)

$$\begin{aligned}\mathbb{P}(Z \mid Y) &\propto \mathbb{P}(Z_1)\mathbb{P}(Y_1 \mid Z_1) \prod_{t=2}^{n} \mathbb{P}(Z_t \mid Z_{t-1})\mathbb{P}(Y_t \mid Z_t) \\ &:= g_1(Z_1) \prod_{t=2}^{n} g_t(Z_{t-1}, Z_t).\end{aligned}$$

MLE via EM

Complete data log-likelihood:

- Assume initial distribution π is known.
- Use indicators: $Z_{tj} = I(Z_t = j)$.

$$\mathbb{P}(Y, Z; \theta) \propto \prod_{j=1}^{N} \prod_{k=1}^{M} \prod_{t: Y_t = k} \{b_j(k)\}^{Z_{tj}} \times \prod_{i=1}^{N} \prod_{j=1}^{N} \prod_{t=2}^{n} (a_{ij})^{Z_{(t-1)i} Z_{tj}}.$$

$$\log \mathbb{P}(Y, Z; \theta) = \sum_{j,k} \underbrace{\sum_{t: Y_t = k} Z_{tj}}_{D_{jk}} \log b_j(k) + \sum_{i,j} \underbrace{\sum_{t=2}^{n} Z_{(t-1)i} Z_{tj}}_{C_{ij}} \log a_{ij}$$

$$= \sum_{j} \left[\sum_{k=1}^{M} D_{jk} \log b_j(k) \right] + \sum_{i} \left[\sum_{j=1}^{N} C_{ij} \log a_{ij} \right]$$

MLE via EM

- Sufficient statistic:
  $D_{jk} = \sum_{t:Y_t=k} Z_{tj}$: # of emissions of symbol k from state j.
  $C_{ij} = \sum_{t=2}^{n} Z_{(t-1)i} Z_{tj}$: # of state transitions from i to j.
- Normalization constraints: $\sum_k b_j(k) = 1$ for each j and $\sum_j a_{ij} = 1$ for each i.
- Complete data MLE (Z are given):

  Let $D_{j\bullet} = \sum_k D_{jk}$ and $C_{i\bullet} = \sum_j C_{ij}$,

$$\hat{b}_j(k) = \frac{D_{jk}}{D_{j\bullet}}, \quad j = 1, \ldots, N, \quad k = 1, \ldots, M,$$

$$\hat{a}_{ij} = \frac{C_{ij}}{C_{i\bullet}}, \quad i, j = 1, \ldots, N.$$

MLE via EM

But Z unobserved, use EM:

- E-step:

$$\begin{aligned}&\mathbb{E}\{\log \mathbb{P}(Y, Z; \theta) \mid Y; \theta^{(m)}\} \\ &= \sum_{j,k} \underbrace{\mathbb{E}(D_{jk} \mid Y; \theta^{(m)})}_{D_{jk}^{(m)}} \log b_j(k) + \sum_{i,j} \underbrace{\mathbb{E}(C_{ij} \mid Y; \theta^{(m)})}_{C_{ij}^{(m)}} \log a_{ij}.\end{aligned}$$

- M-step:

$$b_j(k)^{(m+1)} = \frac{D_{jk}^{(m)}}{D_{j\bullet}^{(m)}}, \quad a_{ij}^{(m+1)} = \frac{C_{ij}^{(m)}}{C_{i\bullet}^{(m)}}.$$

MLE via EM

How to calculate $D_{jk}^{(m)}$ and $C_{ij}^{(m)}$?

- $D_{jk}^{(m)} = \mathbb{E}(D_{jk} \mid Y; \theta^{(m)}) = \sum_{t:Y_t=k} \mathbb{E}(Z_{tj} \mid Y; \theta^{(m)})$.
- $C_{ij}^{(m)} = \mathbb{E}(C_{ij} \mid Y; \theta^{(m)}) = \sum_{t=2}^{n} \mathbb{E}(Z_{(t-1)i} Z_{tj} \mid Y; \theta^{(m)})$.
- Thus, given model parameter $\theta = \theta^{(m)}$, we need to calculate:

$$\mathbb{P}(Z_t = j \mid Y)$$
$$\mathbb{P}(Z_{t-1} = i, Z_t = j \mid Y)$$

  for each t and all i, j.

MLE via EM

Use conditional independence:

$$\begin{aligned}\mathbb{P}(Z_t = j \mid Y) &\propto \mathbb{P}(Y, Z_t = j) \\ &= \underbrace{\mathbb{P}(Y_{1:t}, Z_t = j)}_{\alpha_t(j)} \cdot \underbrace{\mathbb{P}(Y_{(t+1):n} \mid Z_t = j)}_{\beta_t(j)} \\ &= \alpha_t(j)\beta_t(j).\end{aligned}$$

$$\Rightarrow \mathbb{P}(Z_t = j \mid Y) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^N \alpha_t(i)\beta_t(i)} := u_t(j), \quad j = 1, \ldots, N \quad (2)$$

by normalization.

MLE via EM

$$
\begin{aligned}
&\mathbb{P}(Z_{t-1} = i, Z_t = j \mid Y) \propto \mathbb{P}(Y, Z_{t-1} = i, Z_t = j) \\
=&\underbrace{\mathbb{P}(Y_{1:(t-1)}, Z_{t-1} = i)}_{\alpha_{t-1}(i)} \cdot \underbrace{\mathbb{P}(Z_t = j \mid Z_{t-1} = i)}_{a_{ij}} \\
&\times \underbrace{\mathbb{P}(Y_t \mid Z_t = j)}_{b_j(Y_t)} \cdot \underbrace{\mathbb{P}(Y_{(t+1):n} \mid Z_t = j)}_{\beta_t(j)} \\
=&a_{ij} b_j(Y_t) \alpha_{t-1}(i) \beta_t(j).
\end{aligned}
$$

By normalization, for all i, j,

$$
\begin{aligned}
\mathbb{P}(Z_{t-1} = i, Z_t = j \mid Y) &= \frac{a_{ij} b_j(Y_t) \alpha_{t-1}(i) \beta_t(j)}{\sum_k \sum_\ell a_{k\ell} b_\ell(Y_t) \alpha_{t-1}(k) \beta_t(\ell)} \\
&:= w_t(i, j). \qquad (3)
\end{aligned}
$$

MLE via EM

Recall $\alpha_t(i) = \mathbb{P}(Y_{1:t}, Z_t = i)$. We have

$$\begin{aligned}
\alpha_{t+1}(j) &= \mathbb{P}(Y_{1:(t+1)}, Z_{t+1} = j) \\
&= \sum_{i=1}^{N} \mathbb{P}(Y_{1:(t+1)}, Z_t = i, Z_{t+1} = j) \\
&= b_j(Y_{t+1}) \sum_{i=1}^{N} a_{ij}\alpha_t(i).
\end{aligned}$$

Forward summation to calculate $\alpha_t(j)$ for all j and t:

1. Initialization: $\alpha_1(i) = \pi_i b_i(Y_1)$ for $i = 1, \ldots, N$.
2. Recursion: For $t = 1, \ldots, n-1$,

$$\alpha_{t+1}(j) = b_j(Y_{t+1}) \sum_{i=1}^{N} a_{ij}\alpha_t(i), \quad j = 1, \ldots, N.$$

MLE via EM

Similarly, *backward summation* to calculate $\beta_t(i)$ for all i and t:

1. Initialization: $\beta_n(i) = 1$ for $i = 1, \ldots, N$.
2. Recursion: For $t = n - 1, \ldots, 1$,

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(Y_{t+1}) \beta_{t+1}(j), \quad i = 1, \ldots, N.$$

Note that (i) both $\alpha_t(i)$ and $\beta_t(i)$ are calculated given $\theta^{(m)}$;
(ii) recursions make use of $Z \mid Y$ is a Markov chain.

MLE via EM

EM algorithm for HMMs:

- E-step: Given $\theta^{(m)}$,
  1. forward and backward summations to calculate $\alpha_t(i)$ and $\beta_t(i)$;
  2. calculate $u_t(j)$ and $w_t(i,j)$ by (2) and (3);
  3. $D_{jk}^{(m)} = \sum_{t:Y_t=k} u_t(j)$ and $C_{ij}^{(m)} = \sum_{t=2}^{n} w_t(i,j)$.
- M-step:

$$b_j(k)^{(m+1)} = \frac{D_{jk}^{(m)}}{D_{j\bullet}^{(m)}}, \quad a_{ij}^{(m+1)} = \frac{C_{ij}^{(m)}}{C_{i\bullet}^{(m)}}.$$

Iterate between the two steps until convergence. Monitor observed data likelihood (should be non-decreasing)

$$\mathbb{P}(Y \mid \theta^{(m)}) = \sum_i \alpha_n(i).$$

The Viterbi algorithm

Predict hidden states Z given model parameter $\theta = \hat{\theta}$.

- MAP (maximum a posteriori):
  $\hat{z} = \text{argmax}_z \, \mathbb{P}(Z = z \mid Y) = \text{argmax}_z \, \mathbb{P}(Y, Z = z)$.
- Derive recursion to maximize $\mathbb{P}(Y, z_1, \ldots, z_n)$, using the Markovian structure of $Z \mid Y$.

$$
\begin{aligned}
\delta_{t+1}(j) &:= \max_{z_1,\ldots,z_t} \mathbb{P}(Y_{1:(t+1)}, z_{1:t}, Z_{t+1} = j) \\
&= \max_{1 \le i \le N} \underbrace{\max_{z_1,\ldots,z_{t-1}} \mathbb{P}(Y_{1:t}, z_{1:(t-1)}, Z_t = i)}_{\delta_t(i)} a_{ij} b_j(Y_{t+1}) \\
&= \max_{1 \le i \le N} \{\delta_t(i) a_{ij}\} \, b_j(Y_{t+1}).
\end{aligned}
$$

- By definition,

$$
\max_{z_1,\ldots,z_n} \mathbb{P}(Y, z_1, \ldots, z_n) = \max_{1 \le i \le N} \delta_n(i).
$$

The Viterbi algorithm (dynamic programming):

- Initialization: $\delta_1(i) = \pi_i b_i(Y_1)$ for $i = 1, \dots, N$.
- Forward maximization: For $t = 1, \dots, n-1$

$$\delta_{t+1}(j) = \max_{1 \le i \le N} \{\delta_t(i) a_{ij}\} \, b_j(Y_{t+1}), \quad j = 1, \dots, N,$$
$$\gamma_{t+1}(j) = \operatorname*{argmax}_{1 \le i \le N} \{\delta_t(i) a_{ij}\}.$$

- Backward tracking to find $\hat{z}$: Put $\hat{z}_n = \operatorname{argmax}_i \delta_n(i)$; for $t = n-1, \dots, 1$, $\hat{z}_t = \gamma_{t+1}(\hat{z}_{t+1})$.

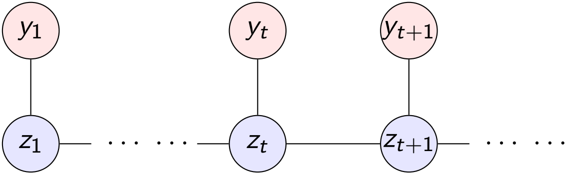
Extensions

Continuous observations: $Y_t = y_t \in \mathbb{R}$.

- Emission density: $Y_t \mid Z_t = j \sim f(y_t; \gamma_j)$.
- Forward summation: $\alpha_{t+1}(j) = f(y_{t+1}; \gamma_j) \sum_{i=1}^{N} \alpha_t(i) a_{ij}$; similarly, replace $b_j(Y_{t+1})$ by $f(y_{t+1}; \gamma_j)$ in backward summation.
- M-step, estimate of γ_j depends on the parametric family f.

Extensions

Kalman filtering:



Continuous observations y_t and continuous states z_t.

- Model:

$$z_{t+1} = az_t + \epsilon_t, \quad \epsilon_t \sim_{iid} \mathcal{N}(0, \tau^2)$$
$$y_t = z_t + \eta_t, \quad \eta_t \sim_{iid} \mathcal{N}(0, \xi^2).$$

- Goal: Online prediction $p(z_t \mid y_1, \ldots, y_t)$.

Extensions

Two lemmas about normal distributions:

Lemma 1

If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \mid X \sim \mathcal{N}(aX, \sigma_2^2)$, then

$$Y \sim \mathcal{N}(a\mu_1, a^2\sigma_1^2 + \sigma_2^2).$$

Lemma 2

If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \mid X \sim \mathcal{N}(X, \sigma_2^2)$, then $X \mid Y \sim \mathcal{N}(\mu, \sigma^2)$, where

$$\mu = \frac{\sigma_1^2 Y + \sigma_2^2 \mu_1}{\sigma_1^2 + \sigma_2^2},$$
$$1/\sigma^2 = 1/\sigma_1^2 + 1/\sigma_2^2.$$

Extensions

Induction:

1. For $t = 1$, $z_1 \mid y_1 \sim \mathcal{N}(y_1, \xi^2) := \mathcal{N}(\mu_1, \sigma_1^2)$.
2. Assume
$$z_t \mid y_1, \ldots, y_t \sim \mathcal{N}(\mu_t, \sigma_t^2), \tag{4}$$
find $[z_{t+1} \mid y_1, \ldots, y_{t+1}]$.

   Since $z_{t+1} \mid z_t \sim \mathcal{N}(az_t, \tau^2)$ by transition model, with (4),
$$z_{t+1} \mid y_1, \ldots, y_t \sim \mathcal{N}(a\mu_t, \tau^2 + a^2\sigma_t^2),$$
by Lemma 1.

   From emission model, $y_{t+1} \mid z_{t+1} \sim \mathcal{N}(z_{t+1}, \xi^2)$.

Extensions

Thus,

$$
\begin{aligned}
p(z_{t+1} \mid y_1, \ldots, y_t, y_{t+1}) &\propto p(z_{t+1} \mid y_1, \ldots, y_t) p(y_{t+1} \mid z_{t+1}) \\
&= \phi(z_{t+1}; a\mu_t, \tau^2 + a^2\sigma_t^2)\phi(y_{t+1}; z_{t+1}, \xi^2) \\
&= \phi(z_{t+1}; a\mu_t, \tau^2 + a^2\sigma_t^2)\phi(z_{t+1}; y_{t+1}, \xi^2)
\end{aligned}
$$

Applying Lemma 2,

$$
\begin{aligned}
\therefore \quad & z_{t+1} \mid y_1, \ldots, y_t, y_{t+1} \sim \mathcal{N}(\mu_{t+1}, \sigma_{t+1}^2), \\
& \mu_{t+1} = \frac{w_1^{(t)} a\mu_t + w_2 y_{t+1}}{w_1^{(t)} + w_2}, \\
& 1/\sigma_{t+1}^2 = w_1^{(t)} + w_2, \\
& w_1^{(t)} = \left(\tau^2 + a^2\sigma_t^2\right)^{-1}, \quad w_2 = 1/\xi^2.
\end{aligned}
$$

References

- Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77: 257-286.
- Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994) Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235: 1501-1531.
- Liu, J.S. (2001) *Monte Carlo Strategies in Scientific Computing*, §2.4.