Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods **Lecture: 21**

Matrix based representations for HMMs

Instructor: Sham Kakade

1 HMMs

The HMM defines a probability distribution over sequences of hidden states (h_t) and observations (x_t). We write the set of hidden states as $\{1, \ldots, k\}$ and set of observations as $\emptyset = \{1, \ldots, d\}$, where $k \leq d$.

Let $T \in \mathbb{R}^{k \times k}$ be the state transition probability matrix with $T_{i,j} = \Pr[h_{t+1} = i | h_t = j]$, so

$$\mathbb{E}[h_{t+1}|h_t] = T h_t \,.$$

Let $O \in \mathbb{R}^{d \times k}$ be the observation probability matrix with $O_{i,j} = \Pr[x_t = i | h_t = j]$, so

$$\mathbb{E}[x_t|h_t] = O h_t \,.$$

Let $\pi \in \mathbb{R}^k$ be the initial state distribution with $\pi_i = \Pr[h_1 = i] = [\mathbb{E}h_1]_i$. The conditional independence properties that an HMM satisfies are that: 1) conditioned on the previous hidden state, the next hidden state is sampled independently of all other events in the history and 2) conditioned on the current hidden state, the current observation is sampled independently from all other events in the history. These conditional independence properties of the HMM imply that T and O fully characterize the probability distribution of any sequence of states and observations.

Note that $\mathbb{E}[h_t | x < t]$ is just the posterior distribution $\Pr(h_t | x_{x<t})$, i.e.

$$(\mathbb{E}[h_t | x < t])_i = \Pr(h_t = i | x_{<t})$$

since h_t is a binary vector.

A useful way of computing the probability of sequences is in terms of 'observation operators', an idea which dates back to the literature on multiplicity automata. The following definitions are useful.

For $x = 1, \ldots, d$, define

$$A_x = T \,\mathrm{Diag}(O_{x,1}, \ldots, O_{x,k}).$$

Also, given $x_1, \ldots x_t$, define the vector $\Pr(h_{t+1} = \cdot, x_1, \ldots, x_t)$ as follows:

$$(\Pr(h_{t+1} = \cdot, x_1, \ldots, x_t))_j = \Pr(h_{t+1} = j, x_1, \ldots, x_t)$$

Lemma 1.1. *For any t:*

$$\Pr(h_t = \cdot, x_1, \ldots, x_t) = A_{x_t} \ldots A_{x_1} \pi.$$

and

$$\Pr(x_1, \ldots, x_t) \;=\; 1_m^\top A_{x_t} \ldots A_{x_1} \pi.$$

Proof. We have:

$$\begin{aligned}
\Pr(h_{t+1} = j, x_1, x_2 \ldots, x_t) &= \sum_{h_1, \ldots h_t} \Pr[x_1, h_1, x_2, h_2 \ldots, x_t, h_t, h_{t+1} = j] \\
&= \sum_{h_1, \ldots h_t} \Pr(h_{t+1} = j | h_t) \Pr(x_t | h_t) \ldots \Pr(h_3 | h_2) \Pr(x_2 | h_2) \Pr(h_2 | h_1) \Pr(x_1 | h_1) \Pr(h_1) \\
&= \sum_{h_1, \ldots h_t} T_{j,h_t} O_{x_t,h_t} \ldots T_{h_3,h_2} O_{x_2,h_2} T_{h_2,h_1} O_{x_1,h_1} \pi_{h_1} \\
&= \sum_{h_1, \ldots h_t} [A_{x_t}]_{j,h_t} \ldots [A_{x_2}]_{h_3,h_2} [A_{x_1}]_{h_2,h_1} \pi_{h_1} \\
&= [A_{x_t} \ldots A_{x_1} \pi]_j
\end{aligned}$$

which proves the claim the first claim. Since

$$1_m^\top \Pr(h_t = \cdot, x_1, \ldots, x_t) = \sum_j \Pr(h_{t+1} = j, x_1, x_2 \ldots, x_t) = \Pr(x_1, x_2 \ldots, x_t)$$

the second claim follows. □

The point of this lemma is to show that HMMs have a certain natural linear algebra structure. In fact, we have a concise (matrix based) method for keeping tracking of our belief state, which sufficient for making predictions. Note:

$$\begin{aligned} \Pr[h_{t+1} = j | x_1, \ldots, x_t] &= \frac{\Pr[h_{t+1} = j, x_1, \ldots, x_t]}{\Pr[x_1, \ldots, x_t]} \\ &= \frac{(A_{x_t} \ldots A_{x_1} \pi)_j}{1_m^\top A_{x_t} \ldots A_{x_1} \pi} \end{aligned}$$

Define our belief state as:

$$g_t = \mathbb{E}[h_t | x_{<t}] = \Pr[h_t = \cdot | x_1, \ldots, x_{t-1}] = \frac{(A_{x_{t-1}} \ldots A_{x_1} \pi)_j}{1_m^\top A_{x_{t-1}} \ldots A_{x_1} \pi}$$

Hence,

$$g_{t+1} = \frac{A_{x_t} \ldots A_{x_1} \pi}{1_m^\top A_{x_t} \ldots A_{x_1} \pi} = \frac{A_{x_t} g_t}{1_m^\top A_{x_t} g_t}$$

Furthermore:

$$\begin{aligned} \Pr[x_{t+1} | x_1, \ldots, x_t] &= \frac{\Pr[x_1, \ldots, x_{t+1}]}{\Pr[x_1, \ldots, x_t]} \\ &= \frac{1_m^\top A_{x_{t+1}} \ldots A_{x_1} \pi}{1_m^\top A_{x_t} \ldots A_{x_1} \pi} \\ &= 1_m^\top A_{x_{t+1}} g_{t+1} \end{aligned}$$

from Bayes rule. Also note:

$$\Pr[x_{t+1} | x_1, \ldots, x_t] = (O g_{t+1})_{x_{t+1}} = O_{x_t} \cdot g_{t+1}$$

One can also explicitly verify this

$$1_m^\top A_{x_{t+1}} = O_{x_{t+1}}$$

using the definition of $A_\cdot$.

Lemma 1.2. *We have the following concise matrix method for posterior probability updates and prediction:*

$$\begin{aligned} g_1 &= \pi \\ g_{t+1} &= \frac{A_{x_t} g_t}{1_m^\top A_{x_t} g_t} \\ \Pr[x_{t+1} | x_1, \ldots, x_t] &= O g_{t+1} \\ 1_m^\top A_{x_{t+1}} g_{t+1} & \end{aligned}$$