LECTURE 16
Hidden Markov Models

The idea of a hidden Markov model (HMM) is an extension of a Markov chain. The basic formalism is that we have two variables $X_1, ..., X_T$ which are observed and $Z_1, ..., Z_T$ which are hidden states and they have the following conditional dependence structure

$$\begin{aligned} x_{t+1} &= f(x_t; \theta_1) \\ z_{t+1} &= g(x_{t+1}; \theta_2), \end{aligned}$$

where we think of t as time and $f(\cdot)$ and $g(\cdot)$ are conditional distributions. In this case we think of time as discrete. Typically in HMMs we consider the hidden states to be discrete, there are more general state space models where both the hidden variables and the observables are continuous. The parameters of the conditional distribution $g(\cdot)$ is often called the transition probabilities and the parameters for observed distribution $g(x_{t+1}; \theta_2)$ are often called the emission probabilities. We will often use the notation $x_{1:t} \equiv x_1, ..., x_t$.

The questions normally asked using a HMM include:

- Filtering: Given the observations $x_1, ..., x_t$ we want to know the hidden states $z_1, ..., z_t$ so we want to infer – $p(z_{1:t} \mid x_{1:t})$.
- Smoothing: Given the observations $x_1, ..., x_T$ we want to know the hidden states $z_1, ..., z_t$ where $t < T$. Here we are using past and future observation to infer hidden states – $p(z_{1:t})$
- Posterior sampling: $z_{1:T} \sim p(z_{1:T} \mid x_{1:T})$

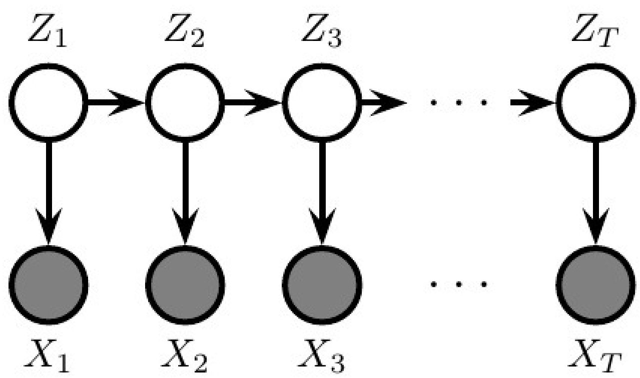
The hidden variables in an HMM are what make inference challenging. We start by writing down the joint (complete) likelihood

$$\text{Lik}(x_1, ..., x_T, z_1, ..., z_T; \theta_1, \theta_2) = \pi(z_1) \prod_{t=2}^{T} f(z_{t+1} \mid z_t, \theta_1) \prod_{t=1}^{T} g(x_t \mid z_t, \theta_2),$$

here $\pi(\cdot)$ is the probability of the initial state. One can obtain the likelihood of the observed data by marginalization

$$\text{Lik}(x_1, ..., x_T; \theta_1, \theta_2) = \sum_{z_1, \ldots, x_T} \left(\pi(z_1) \prod_{t=2}^{T} f(z_{t+1} \mid z_t, \theta_1) \prod_{t=1}^{T} g(x_t \mid z_t, \theta_2) \right).$$

Naively the above sum is brutal since it consists of all possible hidden trajectories. If we assume N hidden states then we would have N^T possible trajectories. We will see that the Markov structure will buy us a great deal in terms of reducing computations.



16.1. EM algorithm

We start with the complete log likelihood

$$\begin{aligned}
\ell_c(z, x; \theta) &= \log[\text{Lik}(z, x \mid \theta)] \\
&= \log\left\{ p(z_1) \left[\prod_{t=1}^{T} p(z_t \mid z_{t-1}) \right] \left[\prod_{t=1}^{T} p(x_t \mid z_t) \right] \right\} \\
&= \log \pi(z_1) + \sum_{t=1}^{T-1} \log a_{z_t, z_{t+1}} + \sum_{t=1}^{T} \log p(x_t \mid z_t, \theta_2).
\end{aligned}$$

We then write the expected complete log likelihood

$$\begin{aligned}
\mathbb{E}\ell_c(z, x; \theta) &= \mathbb{E} \log[\text{Lik}(z, x \mid \theta)] \\
&= \mathbb{E} \log\left\{ p(z_1) \left[\prod_{t=1}^{T} p(z_t \mid z_{t-1}) \right] \left[\prod_{t=1}^{T} p(x_t \mid z_t) \right] \right\} \\
&= \sum_{k=1}^{N} \mathbb{E}[z_1^k] \log \pi_k + \log \pi(z_1) + \sum_{t=1}^{T-1} \sum_{j,k=1}^{K} \mathbb{E}[z_t^j z_{t+1}^k] \log a_{jk} \\
&\quad + \sum_{t=1}^{T} \mathbb{E}[\log p(X_t \mid Z_t, \theta_2)],
\end{aligned}$$

where z_t^k indicates that at time t one is in the k-th state.

For the E step of theEM algorithm we will need to compute

$$\mathbb{E}[Z_1^k] = \mathbb{E}[Z_1^k \mid X_{1:T}, \theta] = p(Z_1^k = 1 \mid X_{1:T}, \theta)$$

This is what we expect since Z_1 follows a Multinomial distribution, so its expectation is simply the vector of posterior probabilities. We will also need to compute

$$\mathbb{E}[Z_t^j, Z_{t+1}^k] = \mathbb{E}[Z_t^j, Z_{t+1}^k \mid X_{1:T}, \theta] = \sum_{t=1}^{T-1} p(Z_t^j Z_{t+1}^k \mid X_{1:T}, \theta)$$

Note that intuitively, $\mathbb{E}[Z_t^j, Z_{t+1}^k]$ counts how often we see transition pairs.

We now state the forward-backward algorithm which is an efficient way of computing the expectations above. We would like to compute $p(z_1 \mid x_{1:T})$ so we start by writing

$$\begin{aligned} p(z_t \mid x_{1:T}) &= \frac{p(z_t, x_{1:T})}{p(y_{1:T})} \\ p(z_t, x_{1:T}) &= p(x_{1:T} \mid z_t)p(z_t) \\ &= p(x_{1:t}, z_t)p(x_{t+1:T} \mid z_t) \\ &= \alpha(z_t)\beta(z_t), \end{aligned}$$

where $\alpha(z_t)$ looks back and $\beta(z_t)$ looks forward. Both can be computed recursively.

For α:

$$\begin{aligned} \alpha(z_t) &= p(x_{1:t}, z_t) \\ &= \sum_{z_{t-1}} p(x_{1:t}, z_t, z_{t-1}) \\ &= \sum_{z_{t-1}} p(x_{1:t-1}, z_{t-1})p(x_t, z_t \mid x_{1:t-1}, z_{t-1}) \\ &= \sum_{z_{t-1}} p(x_{1:t-1}, z_{t-1})p(z_t \mid z_{t-1})p(x_t \mid z_t) \\ &= \sum_{z_{t-1}} \alpha(z_{t-1})p(z_t \mid z_{t-1})p(x_t \mid z_t), \end{aligned}$$

note that given parameter models the above is easy to compute since $p(x_t \mid z_t)$ is the emission probability and $p(z_t \mid z_{t-1})$ is the state transition probability. Note that we can initialize α as $\alpha(z_1) = p(x_1, z_1) = p(z_1)p(x_1 \mid z_1)$.

For β:

$$\begin{aligned} \beta(z_t) &= p(x_{t+1:T} \mid z_t) \\ &= \sum_{z_{t+1}} p(x_{t+1:T}, z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} p(x_{t+1:T} \mid z_{t+1}, z_t)p(z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} p(x_{t+2:T} \mid z_{t+1})p(x_{t+1} \mid y_{t+1})p(z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} \beta(z_{t+1})p(x_{t+1} \mid z_{t+1})p(z_{t+1} \mid z_t) \end{aligned}$$

note that given parameter models the above is easy to compute since $p(x_{t+1} \mid z_{t+1})$ is the emission probability and $p(z_{t+1} \mid z_t)$ is the state transition probability. Note that we can initialize β as $\beta(z_{T-1}) = p(x_T \mid z_{T-1}) = \sum_{z_T} p(x_T \mid z_T)p(z_T \mid z_{T-1})$.

This results in an algorithm with two phases
forward phase: $\alpha(z_t) = p(x_t \mid z_t) \sum_{z_{t-1}} p(z_t \mid z_{t-1})\alpha(z_{t-1})$
backward phase: $\beta(z_t) = \sum_{z_{t+1}} p(x_{t+1} \mid z_{t+1})p(z_{t+1} \mid z_t)\beta(z_{t+1})$.
Also we observe

$$p(z_t \mid x_{1:T}) = \frac{p(z_1 \mid x_{1:T})}{p(x_{1:T})} \propto \alpha(z_t)\beta(z_t).$$

Recall in the E step we need to compute

$$\mathbb{E}[Z_1^k] = p(z_1^k \mid x_{1:T}) \propto \alpha(z_1)\beta(z_1),$$

and

$$\begin{aligned}
\mathbb{E}[Z_t^j Z_{t+1}^k] &= p(z_t^j z_{t+1}^k \mid x_{1:T}) \\
&\propto p(z_t^j z_{t+1}^k, x_{t+1:T}) \\
&\propto p(x_{t+2:T} \mid z_{t+1}^k) p(x_{t+1} \mid z_{t+1}^k) p(z_{t+1}^k \mid z_t^j) p(z_t^j \mid x_{1:t})) \\
&= \beta(z_{t+1}^k)\, p(x_{t+1} \mid z_{t+1}^k)\, p(z_{t+1}^k \mid z_t^j)\, \alpha(z_t^j).
\end{aligned}$$

The above equations provide our estimates of $\mathbb{E}[Z_1^k]$ and $\mathbb{E}[Z_t^j Z_{t+1}^k]$ give current model parameters and the α and β computations.

We now specify the M step. For notation, we set the parameters of the transition probabilities are denoted as $a_{jk} = p(z_t^j \mid z_{t+1}^k)$, the initial probabilities as π_i, the parameters of the emission probabilities which is again a multinomial as $\eta_{jk} = p(x_t^j \mid z_t^k)$. The complete log likelihood with the parameters can be stated as

$$\sum_{i=1}^N E[Z_1^i] \log \pi_i + \sum_{t=1}^T \sum_{i,j=1}^N E[Z_t^i Z_t^j] \log a_{ij} + \sum_{t=1}^T \sum_{i,j=1}^{N,O} \mathbb{E}[Z_t^i X_t^j] \log \eta_{ij},$$

we have assumed N hidden states and O observable states. For ease of notation we define the following terms $\hat{z}_t^i = E[Z_t^i]$, $\quad \widehat{z}_t^{ij} = E[Z_t^i Z_t^j]$. We now write down the sufficient statistics

$$z_1^i, \quad m_{ij} = \sum_{t=1}^T \widehat{z}_t^{ij}, \quad n_{ij} = \sum_{t=1}^T \hat{z}_t^i x_t^j.$$

Given the sufficient statistics and the parameters we minimize the complete log likelihood subject to the constraints

$$\sum_i \pi_i = 1, \quad \sum_{j=1}^N a_{ij} = 1, \quad \sum_{i=1}^O n_{ij} = 1.$$

Using Lagrange multipliers we obtain

$$\begin{aligned}
\hat{\pi}_i &= z_1^i \\
\hat{a}_{ij} &= \frac{m_{ij}}{\sum_{k=1}^N m_{ik}} \\
\hat{\eta}_{ij} &= \frac{n_{ij}}{\sum_{k=1}^O n_{ik}}.
\end{aligned}$$