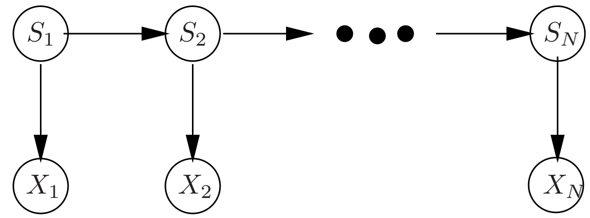
Hidden Markov Models (HMMs)

Seungjin Choi

Department of Computer Science
POSTECH, Korea
seungjin@postech.ac.kr

HMMs?



- HMMs are a ubiquitous tool for modeling time series data.
- Assume that the observation X_t was generated by some process whose state S_t is hidden from the observer.
- Discrete hidden state satisfies the Markov property.
- HMMs can be viewed as a particular instance of Bayesian network.

HMM vs LDS

- What are common in both HMM and LDS (a.k.a. Kalman filter and smoother)?
  - Both have the same independence diagram and consequently the learning and inference algorithms for both have the same structure.
  - Both assume that a hidden state variable evolves with Markovian dynamics.
- What are different?
  - HMM uses a discrete state variable with arbitrary dynamics and arbitrary measurements.
  - LDS uses a continuous state variable with linear Gaussian dynamics and measurements.

Three Basic Tasks in HMM

Classification Compute the probability that a measurement sequence $\{x_1, \ldots, x_N\}$ came from this model, i.e. $p(x_1, \ldots, x_N|\theta)$.

Inference Compute the probability that the system was in state ξ at time t, i.e., $p(s_t = \xi|x_1, \ldots, x_N)$.

Learning Determine the parameter settings that maximize the probability of the measurement sequences.

Learning HMMs will be done by EM!

Parameterization of HMM

The joint distribution of a sequence of states and observations is given by

$$p(s_{1:N}, x_{1:N}) = p(s_1)p(x_1|s_1)\prod_{t=2}^{N}\left[p(s_t|s_{t-1})p(x_t|s_t)\right].$$

The following parameterization is required to define a probability distribution over sequences of observations:

- Initial state: $\pi = p(s_1)$.
- State transition probability: $A_{ij} = p(s_{t,i}|s_{t,j})$.
- Emission probability: $E_{ij} = p(x_{t,i}|s_{t,j})$.

Details on Parameterization

The log probability of the hidden variables and observations is written as

$$\log p(s_{1:N}, x_{1:N}) = \log p(s_1) + \sum_{t=1}^{N}\log p(x_t|s_t) + \sum_{t=2}^{N}\log p(s_t|s_{t-1}).$$

- Transition probability

$$\begin{aligned} p(s_t|s_{t-1}) &= \prod_{i=1}^{K}\prod_{j=1}^{K}(A_{ij})^{s_{t,i}s_{t-1,j}}, \\ \log p(s_t|s_{t-1}) &= \sum_{i=1}^{K}\sum_{j=1}^{K} s_{t,i}s_{t-1,j}\log A_{ij} = s_t^\top(\log A)\, s_{t-1}. \end{aligned}$$

- Initial state probability: $\log p(s_1) = s_1^\top \log \pi$.
- Emission probability: $\log p(x_t|s_t) = x_t^\top(\log E)\, s_t$. ($E \in \mathbb{R}^{D\times K}$)

The parameter set is $\theta = \{A, \pi, E\}$.

Learning HMM

The log probability of the hidden variables and observations is written as

$$\begin{aligned} \log p(s_{1:N}, x_{1:N}) &= \log p(s_1) + \sum_{t=1}^{N}\log p(x_t|s_t) + \sum_{t=2}^{N}\log p(s_t|s_{t-1}) \\ &= s_1^\top \log\pi + \sum_{t=1}^{N} x_t^\top(\log E)\, s_t + \sum_{t=2}^{N} s_t^\top(\log A)\, s_{t-1}. \end{aligned}$$

EM for HMM

E-step Evaluate $\langle \log p(s_{1:N}, x_{1:N})\rangle_{p(s|x,\theta)} \Rightarrow$ Need to compute $\langle s_t\rangle$ and $\left\langle s_t s_{t-1}^\top\right\rangle$.
M-step Re-estimate θ which maximizes the complete-data log-likelihood.

Expected Complete-Data Log-Likelihood

We have to consider the following constraints:

$$\sum_{i=1}^{K} A_{ij} = 1, \quad \sum_{i=1}^{D} E_{ij} = 1, \quad \sum_{i=1}^{K}\pi_i = 1.$$

To this end, we consider the following Lagrangian:

$$\left\langle \widetilde{\mathcal{L}} \right\rangle = \langle \mathcal{L}\rangle + \sum_{j=1}^{K}\lambda_j\left(1 - \sum_{i=1}^{K} A_{ij}\right) + \sum_{j=1}^{K}\rho_j\left(1 - \sum_{i=1}^{D} E_{ij}\right) + \eta\left(1 - \sum_{i=1}^{K}\pi_i\right),$$

where the expected complete-data log-likelihood is given by

$$\begin{aligned} \langle \mathcal{L}\rangle &= \sum_{i=1}^{K}\langle s_{1,i}\rangle \log \pi_i + \sum_{t=1}^{N}\sum_i\sum_j x_{t,i}\log E_{ij}\,\langle s_{t,j}\rangle \\ &\quad + \sum_{t=2}^{N}\sum_i\sum_j \log A_{ij}\,\langle s_{t,i}s_{t-1,j}\rangle. \end{aligned}$$

M-Step

Solving

$$\frac{\partial\left\langle\widetilde{\mathcal{L}}\right\rangle}{\partial A_{ij}}=0,\quad \frac{\partial\left\langle\widetilde{\mathcal{L}}\right\rangle}{\partial E_{ij}}=0,\quad \frac{\partial\left\langle\widetilde{\mathcal{L}}\right\rangle}{\partial \pi_i}=0,$$

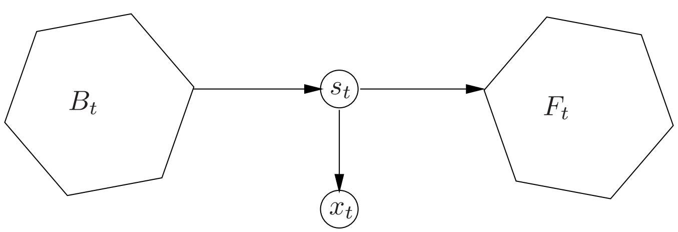
leads to the following updating rules:

$$\begin{aligned} A_{ij} &= \frac{\sum_{t=2}^{N}\langle s_{t,i}s_{t-1,j}\rangle}{\sum_{t=2}^{N}\langle s_{t-1,j}\rangle}, \\ E_{ij} &= \frac{\sum_{t=1}^{N}x_{t,i}\langle s_{t,j}\rangle}{\sum_{t=1}^{N}\langle s_{t,j}\rangle}, \\ \pi_i &= \langle s_{1,i}\rangle . \end{aligned}$$

Inference for E-Step

- Due to the restrictive assumption of a Markov chain, we are able to get an exact inference algorithm.
- E-step is relatively complicated, however, there exists a well-known algorithm, forward-backward recursion.
- In order to compute the expected complete-data log-likelihood, we need to calculate the posterior distribution over latent variables, i.e., $p(s_t|x_{1:N})$.
- Inference involves filtering as well as smoothing.
  - Filtering: $p(s_t|x_1,\ldots,x_t)$.
  - Prediction: $p(s_t|x_1,\ldots,x_\tau)$ for $\tau<t$.
  - Smoothing: $p(s_t|x_1,\ldots,x_\tau)$ for $\tau>t$.

Generic Forward-Backward Propagation (1)



Each state variable separates the graph into three independent parts:

$$p(B_t,s_t,x_t,F_t)=p(B_t,s_t)p(x_t|s_t)p(F_t|s_t),$$

where

$$\begin{aligned} B_t &= \{s_1,\ldots,s_{t-1},x_1,\ldots,x_{t-1}\}, \\ F_t &= \{s_{t+1},\ldots,s_N,x_{t+1},\ldots,x_N\}. \end{aligned}$$

Generic Forward-Backward Propagation (2)

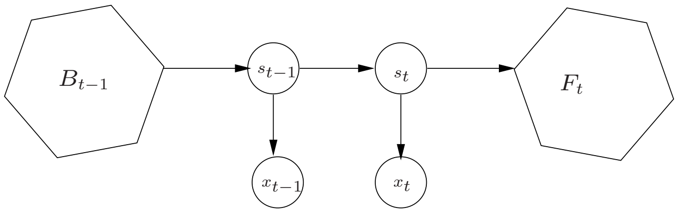
We are interested in computing $p(s_t,x_{1:N})$:

$$\begin{aligned} p(s_t,x_{1:N}) &= \sum_{s_1,\ldots,s_{t-1}}\sum_{s_{t+1},\ldots,s_N}p(B_t,s_t,x_t,F_t) \\ &= \left[\sum_{s_1,\ldots,s_{t-1}}p(B_t,s_t)\right]p(x_t|s_t)\left[\sum_{s_{t+1},\ldots,s_N}p(F_t|s_t)\right] \\ &= p(B_t^x,s_t)p(x_t|s_t)p(F_t^x|s_t), \end{aligned}$$

where

$$\begin{aligned} B_t^x &= \{x_1,\ldots,x_{t-1}\}, \\ F_t^x &= \{x_{t+1},\ldots,x_N\}. \end{aligned}$$

Generic Forward-Backward Propagation (3)



The main idea is to compute $p(B_t^x, s_t)$ and $p(F_t^x|s_t)$ recursively on the left and right subgraphs.

We define

$$\begin{aligned}
\alpha_t(s_t) &= p(B_t^x, s_t)p(x_t|s_t) \\
&= p(s_t, B_t^x, x_t), \\
\beta_t(s_t) &= p(F_t^x|s_t).
\end{aligned}$$

With these definitions, we have $\boxed{p(s_t, x_{1:N}) = \alpha_t(s_t)\beta_t(s_t)}$.

Forward Recursion

It follows from the independence diagram that we have

$$\begin{aligned}
\alpha_t(s_t) &= p(x_t|s_t)p(B_t^x, s_t) \\
&= p(x_t|s_t)\sum_{s_{t-1}} p(B_{t-1}^x, x_{t-1}, s_{t-1}, s_t) \\
&= p(x_t|s_t)\sum_{s_{t-1}} \left[p(B_{t-1}^x, s_{t-1})p(x_{t-1}|s_{t-1})p(s_t|s_{t-1})\right] \\
&= p(x_t|s_t)\sum_{s_{t-1}} \left[\alpha_{t-1}(s_{t-1})p(s_t|s_{t-1})\right].
\end{aligned}$$

Initialization

$$\alpha_1(s_1) = p(s_1)p(x_1|s_1).$$

Backward Recursion

It follows from the independence diagram that we have

$$\begin{aligned}
\beta_{t-1}(s_{t-1}) &= p(F_{t-1}^x|s_{t-1}) \\
&= \sum_{s_t} p(F_x^x, x_t, s_t|s_{t-1}) \\
&= \sum_{s_t} \left[p(s_t|s_{t-1})p(x_t|s_t)p(F_t^x|s_t)\right] \\
&= \sum_{s_t} \left[p(s_t|s_{t-1})p(x_t|s_t)\beta_t(s_t)\right].
\end{aligned}$$

Initialization

$$\beta_N(s_N) = 1.$$

E-Step

We need to compute $\langle s_{t,i}\rangle$, $\langle s_{t,i}\, s_{t-1,j}\rangle$.

$$\begin{aligned}
\langle s_{t,i}\rangle = \gamma_{t,i} &= p(s_{t,i}=1|x_{1:N})\cdot 1 + p(s_{t,i}=0|x_{1:N})\cdot 0 \\
&= p(s_{t,i}=1|x_{1:N}) \\
&= \frac{\alpha_{t,i}\beta_{t,i}}{\sum_j \alpha_{t,j}\beta_{t,j}}, \\
\langle s_{t,i}\, s_{t-1,j}\rangle = \xi_{t,ij} &= p(s_{t,i}, s_{t-1,j}=1|x_{1:N}) \\
&= \frac{\alpha_{t-1,j}A_{ij}p(x_t|s_{t,i}=1)\beta_{t,i}}{\sum_{k,l}\alpha_{t-1,l}A_{kl}p(x_t|s_{t,k}=1)\beta_{t,k}},
\end{aligned}$$

where we used

$$\begin{aligned}
p(s_{t-1}, s_t, x_{1:N}) &= p(B_{t-1}^x, s_{t-1})p(x_{t-1}|s_{t-1})p(s_t|s_{t-1})p(x_t|s_t)p(F_t^x|s_t) \\
&= \alpha_{t-1}(s_{t-1})p(s_t|s_{t-1})p(x_t|s_t)\beta_t(s_t).
\end{aligned}$$

Algorithm Outline: HMM

E-step: Forward-backward recursion

$$\alpha_t(s_t) = p(x_t|s_t)\sum_{s_{t-1}}\left[\alpha_{t-1}(s_{t-1})p(s_t|s_{t-1})\right],$$
$$\beta_{t-1}(s_{t-1}) = \sum_{s_t}\left[p(s_t|s_{t-1})p(x_t|s_t)\beta_t(s_t)\right].$$

Compute $\langle s_{t,i}\rangle$, $\langle s_{t,i}\, s_{t-1,j}\rangle$:

$$\langle s_{t,i}\rangle = \frac{\alpha_{t,i}\beta_{t,i}}{\sum_j \alpha_{t,j}\beta_{t,j}},$$
$$\langle s_{t,i}\, s_{t-1,j}\rangle = \frac{\alpha_{t-1,j}A_{ij}p(x_t|s_{t,i}=1)\beta_{t,i}}{\sum_{k,l}\alpha_{t-1,l}A_{kl}p(x_t|s_{t,k}=1)\beta_{t,k}}$$

M-step: Update parameters:

$$A_{ij} = \frac{\sum_{t=2}^N \langle s_{t,i}s_{t-1,j}\rangle}{\sum_{t=2}^N \langle s_{t-1,j}\rangle}, \quad E_{ij} = \frac{\sum_{t=1}^N x_{t,i}\langle s_{t,j}\rangle}{\sum_{t=1}^N \langle s_{t,j}\rangle}, \quad \pi_i = \langle s_{1,i}\rangle.$$

Scaling

We reformulate the forward-backward recursion in terms of scaled α's and β's. The rescaling is also useful for avoiding numerical underflow.

Define $c_t = p(x_t|x_1,\ldots,x_{t-1})$.

We factor c_t out of the original definition of $\alpha_t(s_t)$:

$$\begin{aligned}\alpha_t(s_t) &= p(s_t,x_1,\ldots,x_t)\\ &= p(x_1,\ldots,x_t)p(s_t|x_1,\ldots,x_t)\\ &= \left(\prod_{\tau=1}^t c_\tau\right)\widehat{\alpha}_t(s_t).\end{aligned}$$

Similarly, we define

$$\begin{aligned}\beta_t(s_t) &= p(x_{t+1},\ldots,x_N|s_t)\\ &= \left(\prod_{\tau=t+1}^N c_\tau\right)\widehat{\beta}_t(s_t).\end{aligned}$$

Recursion for $\widehat{\alpha}_t(s_t)$ and $\widehat{\beta}_t(s_t)$

- Recursion for $\widehat{\alpha}_t(s_t)$:

$$\alpha_t(s_t) = p(x_t|s_t)\sum_{s_{t-1}}\left[\alpha_{t-1}(s_{t-1})p(s_t|s_{t-1})\right],$$
$$\left(\prod_{\tau=1}^t c_\tau\right)\widehat{\alpha}_t(s_t) = p(x_t|s_t)\sum_{s_{t-1}}\left[\left(\prod_{\tau=1}^{t-1} c_\tau\right)\widehat{\alpha}_{t-1}(s_{t-1})p(s_t|s_{t-1})\right].$$

$$\widehat{\alpha}_t(s_t) = \frac{1}{c_t}\sum_{s_{t-1}}\left[\widehat{\alpha}_{t-1}(s_{t-1})p(s_t|s_{t-1})\right].$$

- Recursion for $\widehat{\beta}_t(s_t)$:

$$\widehat{\beta}_{t-1}(s_{t-1}) = \frac{1}{c_t}\sum_{s_t}\left[p(s_t|s_{t-1})p(x_t|s_t)\widehat{\beta}_t(s_t)\right].$$

Marginal Distribution

The marginal distribution become exact in terms of the scaled α's and β's (the distribution do not require normalization):

$$p(s_t|x_{1:N}) = \widehat{\alpha}_t(s_t)\widehat{\beta}_t(s_t),$$
$$p(s_{t-1},s_t|x_{1:N}) = \frac{1}{c_t}\widehat{\alpha}_{t-1}(s_{t-1})p(s_t|s_{t-1})p(x_t|s_t)\widehat{\beta}_t(s_t).$$