732A96 Advanced Machine Learning

Graphical Models and Hidden Markov Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 6: Autoregressive and Explicit-Duration Hidden Markov Models

Contents

Literature

- Main source
  - Chiappa, S. Explicit-Duration Markov Switching Models. *Foundations and Trends in Machine Learning* 7, 803-886, 2014. Sections 1-3.3.
- Additional source
  - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapter 13.1-13.2.

Autoregressive Hidden Markov Models: Definition

- To overcome the poor modeling of long range correlations in HMMs, by allowing $pa_G(X^t) \supset \{Z^t\}$.

$$\begin{tikzcd}
Z^0 \arrow[r] \arrow[d] & Z^1 \arrow[r] \arrow[d] & Z^2 \arrow[r] \arrow[d] & Z^3 \arrow[r] \arrow[d] & Z^4 \arrow[d] \\
X^0 \arrow[r] \arrow[rr, bend right] & X^1 \arrow[r] \arrow[rr, bend right] & X^2 \arrow[r] \arrow[rr, bend right] & X^3 \arrow[r] & X^4
\end{tikzcd}$$

- Hereinafter, we focus on the simplest AR-HMM, i.e. $pa_G(X^t) = \{Z^t, X^{t-1}\}$.

Autoregressive Hidden Markov Models: Learning

- Almost the same as for HMMs.
- Recall that maximizing the log likelihood function over $x^{0:T}$ is inefficient (no closed form solution) and ineffective (multimodal).
- Consider instead maximizing its expectation

$$\mathrm{E}_{Z^{0:T}}[\log p(Z^{0:T}, x^{0:T})] = \sum_{z^{0:T}} p(z^{0:T}|x^{0:T}) \log p(z^{0:T}, x^{0:T})$$

$$= \sum_{z^{0:T}} p(z^{0:T}|x^{0:T})[\log \theta_{z^0} + \sum_{t=1}^{T} \log \theta_{z^{t+1}|z^t} + \sum_{t=1}^{T} \log \theta_{x^t|z^t,x^{t-1}}]$$

$$= \sum_{z^0} p(z^0|x^{0:T}) \log \theta_{z^0} + \sum_{t=1}^{T} \sum_{z^t} \sum_{z^{t+1}} p(z^t, z^{t+1}|x^{0:T}) \log \theta_{z^{t+1}|z^t} + \sum_{t=1}^{T} \sum_{z_t} p(z^t|x^{0:T}) \log \theta_{x^t|z^t,x^{t-1}}$$

- Then
  - $\theta_{z^0}^{ML} = \frac{p(z^0|x^{0:T})}{\sum_{z^0} p(z^0|x^{0:T})}$
  - $\theta_{z^{t+1}|z^t}^{ML} = \frac{\sum_{t=1}^{T} p(z^t,z^{t+1}|x^{0:T})}{\sum_{t=1}^{T} \sum_{z^{t+1}} p(z^t,z^{t+1}|x^{0:T})}$
  - $\theta_{x^t|z^t,x^{t-1}}^{ML} = \frac{\sum_{t=1}^{T} p(z^t|x^{0:T}) 1_{\{x^t,x^{t-1} \in x^{0:T}\}}}{\sum_{t=1}^{T} p(z^t|x^{0:T}) 1_{\{x^{t-1} \in x^{0:T}\}}}$
- Note that computing $p(Z^0|x^{0:T})$, $p(Z^t, Z^{t+1}|x^{0:T})$ and $p(Z^t|x^{0:T})$ requires inference: Forward-backward algorithm.

Autoregressive Hidden Markov Models: Forward-Backward Algorithm

$$\begin{aligned} p(Z^t|x^{0:T}) &= \frac{p(x^{0:t-1}, x^{t+1:T}|Z^t, x^t)p(Z^t, x^t)}{p(x^{0:T})} \\ &= \frac{p(x^{0:t-1}|Z^t, x^t)p(Z^t, x^t)p(x^{t+1:T}|Z^t, x^t)}{p(x^{0:T})} \text{ by } X^{0:t-1} \perp_p X^{t+1:T}|Z^t \cup X^t \\ &= \frac{p(x^{0:t}, Z^t)p(x^{t+1:T}|Z^t, x^t)}{p(x^{0:T})} = \frac{\alpha(Z^t)\beta(Z^t)}{\sum_{z^t} \alpha(z^t)\beta(z^t)} \end{aligned}$$

$$p(Z^t, Z^{t+1}|x^{0:T}) = \frac{p(x^{0:t-1}, x^{t+2:T}|Z^t, Z^{t+1}, x^t, x^{t+1})p(Z^t, Z^{t+1}, x^t, x^{t+1})}{p(x^{0:T})}$$

$$= \frac{p(x^{0:t-1}|Z^t, x^t)p(x^{t+2:T}|Z^{t+1}, x^{t+1})p(x^{t+1}|Z^{t+1}, x^t)p(Z^{t+1}|Z^t)p(Z^t, x^t)}{p(x^{0:T})}$$

by $X^{0:t-1} \perp_p X^{t+2:T}|Z^t \cup Z^{t+1} \cup X^t \cup X^{t+1}$

$X^{0:t-1} \perp_p Z^{t+1} \cup X^{t+1}|Z^t \cup X^t$

$X^{t+2:T} \perp_p Z^t \cup X^t|Z^{t+1} \cup X^{t+1}$

$X^{t+1} \perp_p Z^t|Z^{t+1} \cup X^t$

$Z^{t+1} \perp_p X^t|Z^t$

$$= \frac{\alpha(Z^t)\beta(Z^{t+1})p(x^{t+1}|Z^{t+1}, x^t)p(Z^{t+1}|Z^t)}{\sum_{z^t} \sum_{z^{t+1}} \alpha(z^t)\beta(z^{t+1})p(x^{t+1}|z^{t+1}, x^t)p(z^{t+1}|z^t)}$$

Autoregressive Hidden Markov Models: Forward-Backward Algorithm

$$
\begin{aligned}
\boldsymbol{\alpha(Z^t)} &= p(x^t|Z^t, x^{t-1})p(Z^t, x^{t-1})p(x^{0:t-2}|Z^t, x^{t-1}) \text{ by } X^{0:t-2} \perp_p X^t | Z^t \cup X^{t-1} \\
&= p(x^t|Z^t, x^{t-1})p(x^{0:t-1}, Z^t) = p(x^t|Z^t, x^{t-1}) \sum_{z^{t-1}} p(x^{0:t-1}, Z^t|z^{t-1})p(z^{t-1}) \\
&= p(x^t|Z^t, x^{t-1}) \sum_{z^{t-1}} p(x^{0:t-1}|z^{t-1})p(Z^t|z^{t-1})p(z^{t-1}) \text{ by } X^{0:t-1} \perp_p Z^t | Z^{t-1} \\
&= p(x^t|Z^t, x^{t-1}) \sum_{z^{t-1}} p(x^{0:t-1}, z^{t-1})p(Z^t|z^{t-1}) = p(x^t|Z^t, x^{t-1}) \sum_{z^{t-1}} \boldsymbol{\alpha(z^{t-1})} p(Z^t|z^{t-1})
\end{aligned}
$$

$$\alpha(Z^0) = p(x^0|Z^0)p(Z^0)$$

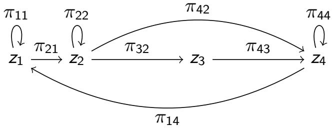
$$
\begin{aligned}
\boldsymbol{\beta(Z^t)} &= \sum_{z^{t+1}} p(x^{t+1:T}, z^{t+1}|Z^t, x^t) = \sum_{z^{t+1}} p(x^{t+1:T}|z^{t+1}, Z^t, x^t)p(z^{t+1}|Z^t, x^t) \\
&= \sum_{z^{t+1}} p(x^{t+1:T}|z^{t+1}, x^t)p(z^{t+1}|Z^t) \text{ by } X^{t+1:T} \perp_p Z^t | Z^{t+1} \cup X^t \text{ and } X^t \perp_p Z^{t+1} | Z^t \\
&= \sum_{z^{t+1}} p(x^{t+2:T}|z^{t+1}, x^{t+1})p(x^{t+1}|z^{t+1}, x^t)p(z^{t+1}|Z^t) \text{ by } X^{t+2:T} \perp_p X^t | Z^{t+1} \cup X^{t+1} \\
&= \sum_{z^{t+1}} \boldsymbol{\beta(z^{t+1})} p(x^{t+1}|z^{t+1}, x^t)p(z^{t+1}|Z^t)
\end{aligned}
$$

$$\beta(Z^T) = 1 \text{ by } p(Z^T|x^{0:T}) = \frac{\alpha(Z^T)\beta(Z^T)}{p(x^{0:T})} = p(Z^T|x^{0:T})\beta(Z^T)$$

- **Exercise**. Derive the Viterbi algorithm for AR-HMMs.

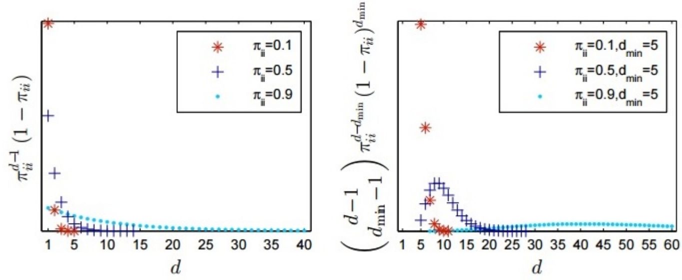
Explicit-Duration Hidden Markov Models: Definition

- To control the distribution of the number of time steps for which the HMM remains in a given state.
- Consider the following Markov chain over the states:



- The probability of remaining in state z_i for exactly $d-1$ time steps is

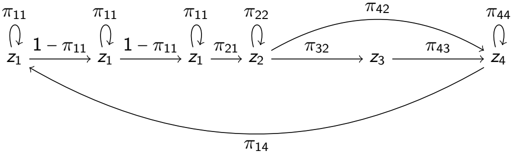
$$p(Z^{t+1:t+d-1} = z_i, Z^{t+d} \neq z_i | Z^t = z_i) = \pi_{ii}^{d-1}(1-\pi_{ii}) = Geometric(d; 1-\pi_{ii})$$



Source: Chiappa (2014).

Explicit-Duration Hidden Markov Models: Definition

- We can obtain a negative binomial distribution by imposing a minimum duration on the time spent in a state, and duplicating the corresponding observation model. For instance, when $d_{min=3}$



- The probability of remaining in state z_i for exactly $d-1$ time steps is

$$\begin{aligned} p(Z^{t+1:t+d-1} &= z_i, Z^{t+d} \neq z_i | Z^t = z_i) = \binom{d-1}{d_{min}-1} \pi_{ii}^{d-d_{min}} (1-\pi_{ii})^{d_{min}} \\ &= NegativeBinomial(d-d_{min}; d_{min}, \pi_{ii}) \end{aligned}$$

Explicit-Duration Hidden Markov Models: Decreasing

- The geometric and negative binomial cases define a segment duration distribution implicitly. We now define the segment duration distribution explicitly, by introducing auxiliary unobserved random variables.
- Specifically, let us use the random variables $C^{1:T}$ which take **decreasing** values within a segment, starting with the duration of the segment and ending with 1. Then

$$p(z_i^t|z_j^{t-1}, C^{t-1}) = \begin{cases} \pi_{ij} & \text{if } C^{t-1} = 1 \\ 1_{i=j} & \text{if } C^{t-1} > 1 \end{cases}$$

$$p(C^t|Z^t, C^{t-1}) = \begin{cases} \rho(Z^t, C^t) & \text{if } C^{t-1} = 1 \\ 1_{C^t = C^{t-1} - 1} & \text{if } C^{t-1} > 1 \end{cases}$$

where $\rho(Z^t, C^t)$ is the segment duration distribution.

- In graphical terms, we have

$$\begin{tikzcd}
C^{t-1} \arrow[r] \arrow[dr] & C^t \arrow[r] \arrow[dr] & C^{t+1} \\
Z^{t-1} \arrow[u] \arrow[r] \arrow[d] & Z^t \arrow[u] \arrow[r] \arrow[d] & Z^{t+1} \arrow[u] \arrow[d] \\
X^{t-1} & X^t & X^{t+1}
\end{tikzcd}$$

Explicit-Duration Hidden Markov Models: Increasing

- Alternatively, we can use the random variables $C^{1:T}$ which take **increasing** values within a segment, starting 1 and ending with the duration of the segment. Then

$$p(z_i^t | z_j^{t-1}, C^t) = \begin{cases} \pi_{ij} & \text{if } C^t = 1 \\ 1_{i=j} & \text{if } C^t > 1 \end{cases}$$

$$p(C^t | Z^{t-1}, C^{t-1}) = \begin{cases} 1 - \lambda(Z^{t-1}, C^{t-1}) & \text{if } C^t = 1 \\ \lambda(Z^{t-1}, C^{t-1}) & \text{if } C^t = C^{t-1} + 1 \end{cases}$$

  where $\lambda(Z^{t-1}, C^{t-1})$ is the probability of the segment continuing.

- In graphical terms, we have

$$\begin{tikzcd}
C^{t-1} \arrow[r] \arrow[d] & C^t \arrow[r] \arrow[d] & C^{t+1} \arrow[d] \\
Z^{t-1} \arrow[r] \arrow[d] \arrow[ur] & Z^t \arrow[r] \arrow[d] \arrow[ur] & Z^{t+1} \arrow[d] \\
X^{t-1} & X^t & X^{t+1}
\end{tikzcd}$$

Explicit-Duration Hidden Markov Models: Decreasing-Duration

- Alternatively, we can use the random variables $C^{1:T}$ which take **decreasing** values within a segment, starting with the duration of the segment and ending with 1. And the random variables $D^{1:T}$ which indicate the **duration** of the current segment. Then

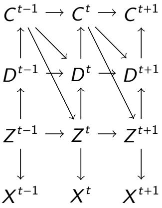
$$p(z_i^t | z_j^{t-1}, C^{t-1}) = \begin{cases} \pi_{ij} & \text{if } C^{t-1} = 1 \\ 1_{i=j} & \text{if } C^{t-1} > 1 \end{cases}$$

$$p(C^t | C^{t-1}, D^t) = \begin{cases} 1_{C^t = D^t} & \text{if } C^{t-1} = 1 \\ 1_{C^t = C^{t-1} - 1} & \text{if } C^{t-1} > 1 \end{cases}$$

$$p(D^t | Z^t, C^{t-1}, D^{t-1}) = \begin{cases} \rho(Z^t, D^t) & \text{if } C^{t-1} = 1 \\ 1_{D^t = D^{t-1}} & \text{if } C^{t-1} > 1 \end{cases}$$

  where $\rho(Z^t, D^t)$ is the segment duration distribution.

- In graphical terms, we have



Explicit-Duration Hidden Markov Models: Increasing-Duration

- **Exercise**. Derive the transition model for the increasing-duration case.
- Relatively easy to adapt the FB and Viterbi algorithms to explicit-duration HMMs.