732A96/TDDE15 Advanced Machine Learning
Hidden Markov Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 6: Hidden Markov Models

Contents

Literature

- Main source
  - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapter 13.1-13.2.
- Additional source
  - Ghahramani, Z. An Introduction to Hidden Markov Models and Bayesian Networks. *International Journal of Pattern Recognition and Artificial Intelligence* 15, 9-42, 2001.

Dynamic Bayesian Networks: Definition

- To model **sequential data**, e.g. time series data.
- **Simplification**: Time is discretized in equal width intervals, i.e. $t = 0, 1, \ldots$
- Consider a finite set of discrete random variables $X^t = \{X_1^t, \ldots, X_n^t\}$ representing the state at time t of a system described by $X = \{X_1, \ldots, X_n\}$.
- A **dynamic Bayesian network** (DBN) is a BN over $X^{0:T} = \{X^0, \ldots, X^T\}$. Thus, it defines $p(x^{0:T})$.

$$
\begin{tikzcd}
X_1^0 \arrow[r] \arrow[d] \arrow[rrrr, bend left] & X_1^1 \arrow[dr] \arrow[rrr, bend left] & X_1^2 \arrow[d] \arrow[rr, bend left] & X_1^3 \arrow[r] & X_1^4 \arrow[d] \arrow[dd, bend left] \\
X_2^0 \arrow[r] \arrow[d] & X_2^1 \arrow[u] \arrow[dr] & X_2^2 \arrow[r] & X_2^3 & X_2^4 \arrow[d] \\
X_3^0 \arrow[r] & X_3^1 \arrow[u] & X_3^2 \arrow[u] \arrow[r] & X_3^3 & X_3^4
\end{tikzcd}
$$

- **Assumption**: The system is Markovian, i.e. $X^{t+1} \perp_p X^{0:t-1} | X^t$.

$$
\begin{tikzcd}
X_1^0 \arrow[r] \arrow[d] & X_1^1 \arrow[dr] & X_1^2 \arrow[d] & X_1^3 \arrow[r] & X_1^4 \arrow[d] \arrow[dd, bend left] \\
X_2^0 \arrow[r] \arrow[d] & X_2^1 \arrow[u] \arrow[dr] & X_2^2 \arrow[r] & X_2^3 & X_2^4 \arrow[d] \\
X_3^0 \arrow[r] & X_3^1 \arrow[u] & X_3^2 \arrow[u] \arrow[r] & X_3^3 & X_3^4
\end{tikzcd}
$$

- **Assumption**: The system is stationary, i.e. $p(x^{t+1}|x^t) = p(x'|x)$.

$$
\begin{tikzcd}
X_1^0 \arrow[r] \arrow[d] & X_1^1 \arrow[r] & X_1^2 \arrow[r] & X_1^3 \arrow[r] & X_1^4 \\
X_2^0 \arrow[d] & X_2^1 \arrow[u] & X_2^2 \arrow[u] & X_2^3 \arrow[u] & X_2^4 \arrow[u] \\
X_3^0 \arrow[r] & X_3^1 \arrow[u] \arrow[r] & X_3^2 \arrow[u] \arrow[r] & X_3^3 \arrow[u] \arrow[r] & X_3^4 \arrow[u]
\end{tikzcd}
$$

Dynamic Bayesian Networks

- Then, a DBN over $X^{0:T}$ can be defined as
  - a BN over X^0, and
  - a BN over $X^t \cup X^{t+1}$ where the nodes in X^t are parentless.

Initial model	Transition model
$\begin{tikzcd} X_1^0 \arrow[d] \\ X_2^0 \arrow[d] \\ X_3^0 \end{tikzcd}$	$\begin{tikzcd} X_1^t \arrow[r] & X_1^{t+1} \\ X_2^t & X_2^{t+1} \arrow[u] \\ X_3^t \arrow[r] & X_3^{t+1} \arrow[u] \end{tikzcd}$

- DBN unrolled for $T = 4$.

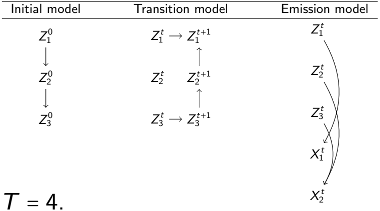
$$\begin{tikzcd}
X_1^0 \arrow[r] \arrow[d] & X_1^1 \arrow[r] & X_1^2 \arrow[r] & X_1^3 \arrow[r] & X_1^4 \\
X_2^0 \arrow[d] & X_2^1 \arrow[u] & X_2^2 \arrow[u] & X_2^3 \arrow[u] & X_2^4 \arrow[u] \\
X_3^0 \arrow[r] & X_3^1 \arrow[r] \arrow[u] & X_3^2 \arrow[r] \arrow[u] & X_3^3 \arrow[r] \arrow[u] & X_3^4 \arrow[u]
\end{tikzcd}$$

- The DBN defines

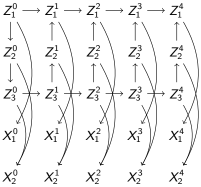
$$p(x^{0:T}) = p(x^0) \prod_{t=0}^{T-1} p(x^{t+1} | x^t) = [\prod_{i=1}^{n} p(x_i^0 | pa_i^0)][\prod_{t=0}^{T-1} \prod_{i=1}^{n} p(x_i^{t+1} | pa_i^{t+1})]$$

Hidden Markov Models: Definition

- To overcome the **Markovian limitation** of DBNs, while keeping sparsity.
- A **hidden Markov model** (HMM) over $\{Z^{0:T}, X^{0:T}\}$ where $X^{0:T}$ are **observed** and $Z^{0:T}$ are **unobserved** consists of
  - a DBN over $Z^{0:T}$, and
  - a BN over $Z^t \cup X^t$ where the nodes in Z^t are parentless.



- HMM unrolled for $T = 4$.



- A HMM is a DBN that defines

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

Hidden Markov Models: Learning

- The structure is typically fixed to

$$\begin{array}{ccccccccc} Z^0 & \longrightarrow & Z^1 & \longrightarrow & Z^2 & \longrightarrow & Z^3 & \longrightarrow & Z^4 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ X^0 & & X^1 & & X^2 & & X^3 & & X^4 \end{array}$$

- Consider a sample with a **single** observation over $X^{0:T}$.
- Parameter learning: EM algorithm.
- Cardinality of Z^t ? BIC score to select among a set of plausible values.

Hidden Markov Models: Learning

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

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

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

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

- Then
  - $\theta_{z^0}^{ML} = \frac{p(z^0|x^{0:T})}{\sum_{z^0} p(z^0|x^{0:T})}$
  - $\theta_{z^{t+1}|z^t}^{ML} = \frac{\sum_{t=1}^{T-1} p(z^t, z^{t+1}|x^{0:T})}{\sum_{t=1}^{T-1} \sum_{z^{t+1}} p(z^t, z^{t+1}|x^{0:T})}$
  - $\theta_{x^t|z^t}^{ML} = \frac{\sum_{t=1}^{T} p(z^t|x^{0:T}) 1_{\{x^t \in x^{0:T}\}}}{\sum_{t=1}^{T} p(z^t|x^{0:T})}$
- 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.

Hidden Markov Models: Forward-Backward Algorithm

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

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

Hidden Markov Models: Forward-Backward Algorithm

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

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

Hidden Markov Models: Forward-Backward Algorithm

FB algorithm

$\alpha(z^0) := p(x^0|z^0)p(z^0)$
For $t = 1, \ldots, T$ do
  $\alpha(z^t) := p(x^t|z^t) \sum_{z^{t-1}} \alpha(z^{t-1}) p(z^t|z^{t-1})$
$\beta(z^T) := 1$
For $t = T-1, \ldots, 0$ do
  $\beta(z^t) := \sum_{z^{t+1}} \beta(z^{t+1}) p(x^{t+1}|z^{t+1}) p(z^{t+1}|z^t)$
Return $\alpha(z^0), \ldots, \alpha(z^T), \beta(z^0), \ldots, \beta(z^T)$

- Unlike the LS algorithm, the FB algorithm consists of two independent steps.
- **Filtering**: $p(z^t|x^{0:t}) = \frac{\alpha(z^t)}{\sum_{z^t} \alpha(z^t)}$.
- **Smoothing**: $p(z^t|x^{0:T}) = \frac{\alpha(z^t)\beta(z^t)}{\sum_{z^t} \alpha(z^t)\beta(z^t)}$.

Hidden Markov Models: Viterbi Algorithm

- To compute the most probable configuration for HMMs.

Viterbi algorithm

$\omega(z^0) := \log p(z^0) + \log p(x^0|z^0)$
For $t = 0, \ldots, T-1$ do
  $\omega(z^{t+1}) := \log p(x^{t+1}|z^{t+1}) + \max_{z^t}[\log p(z^{t+1}|z^t) + \omega(z^t)]$
  $\psi(z^{t+1}) := \arg\max_{z^t}[\log p(z^{t+1}|z^t) + \omega(z^t)]$
$z^T_{\max} = \arg\max_{z^T} \omega(z^T)$
For $t = T-1, \ldots, 0$ do
  $z^t_{\max} := \psi(z^{t+1}_{\max})$
Return $z^{0:T}_{\max}$

- **Exercise**. Prove that the Viterbi algorithm is correct.

Contents

Thank you