732A96 Advanced Machine Learning

Graphical Models and Hidden Markov Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 5: Dynamic Bayesian Networks and Hidden Markov Models

Contents

Literature

- Main sources
  - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Other sources
  - Ghahramani, Z. An Introduction to Hidden Markov Models and Bayesian Networks. *International Journal of Pattern Recognition and Artificial Intelligence* 15, 9-42, 2001.
  - Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
  - Murphy, K. P. Dynamic Bayesian Networks. Draft, 2002.
  - Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
  - Smyth, P., Heckerman, D. and Jordan, M. I. Probabilistic Independence Networks for Hidden Markov Probability Models. *Neural Computation* 9, 227-269, 1997.
- R resources
  - Visser, I. and Speekenbrink, M. depmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software* 36, 2010.

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 $V = \{X_1, \ldots, X_n\}$.
- A **dynamic Bayesian network** (DBN) is a BN over $X^{0:T} = \{X^0, \ldots, X^T\}$. Thus, it defines $p(X^{0:T})$.

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

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

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

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

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

Dynamic Bayesian Networks: Definition

- 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

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

- The DBN defines

$$p(X^{0:T}) = p(X^0)\prod_{t=0}^{T} p(X^{t+1}|X^t) = [\prod_{i=1}^{n} p(X_i^0|pa_G(X_i^0))][\prod_{t=0}^{T}\prod_{i=1}^{n} p(X_i^{t+1}|pa_G(X_i^{t+1}))]$$

- DBN unrolled for $T = 4$.

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

Dynamic Bayesian Networks: Probabilistic Reasoning

- **Filtering**: Computing $p(U^t|o^{0:t})$ where $X^t = O^t \cup U^t$ for increasing t.
- We would like to do it without
  - increasing the size of the DBN where to perform inference, and
  - increasing the space to store the observations.
- Note that

$$p(U^{t-1}, U^t, o^{0:t}) = p(U^t, o^t|U^{t-1}, o^{0:t-1})p(U^{t-1}, o^{0:t-1})$$

- Note that $p(U^t, o^t|U^{t-1}, o^{0:t-1})$ factorizes as indicated by the transition model, i.e.

$$p(U^t, o^t|U^{t-1}, o^{0:t-1}) = \prod_{i=1}^{n} p(X_i^t|pa_G(X_i^t))$$

- Then, $p(U^{t-1}, U^t, o^{0:t})$ factorizes an indicated by the transition model after having made a **complete set** of X^{t-1} by adding directed edges to accommodate $p(U^{t-1}, o^{0:t-1})$.
- Then, filtering can be performed by running the LS algorithm:
  - Incorporate the evidence o^t,
  - propagate to obtain $p(U^t, o^{0:t})$ which is used in the filtering for $t+1$, and
  - normalize to obtain $p(U^t|o^{0:t})$.
- Note that moralization, triangulation and RIP ordering search is the same for all t.

Dynamic Bayesian Networks: Learning

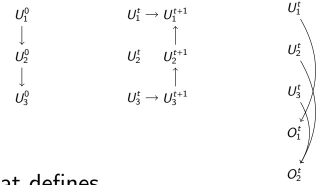
- The same as for BNs with the following particularity.
- Consider a sample with a single observation over $X^{0:T}$.
- The sample can be converted into
  - one observation from the initial model, i.e. x^0, and
  - $T - 1$ observations from the transition model, i.e.

$$\{x^0, x^1\}, \{x^1, x^2\}, \ldots, \{x^{T-1}, x^T\}$$

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.

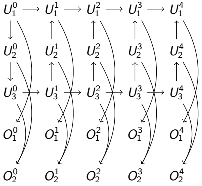
Initial model	Transition model	Emission model



- A HMM is a DBN that defines

$$p(Z^{0:T}, X^{0:T}) = p(Z^0) \prod_{t=1}^{T} p(Z^{t+1}|Z^t) \prod_{t=0}^{T} p(X^t|Z^t)$$

- HMM unrolled for $T = 4$.



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 instead maximizing its expectation

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

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

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

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

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

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

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

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

Autoregressive Hidden Markov Models

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

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

- **Exercise**. Derive the EM algorithm for AR-HMMs with $pa_G(X^t) = \{X^{t-1}\}$. Specifically, derive the recursive expressions for $\alpha(Z^t)$ and $\beta(Z^t)$ to be used in the FB algorithm.
- **Hint**.

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

Autoregressive Hidden Markov Models

- **Hint**.

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

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

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

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

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

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

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

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

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