Probabilistic Graphical Models (5): temporal models

Qinfeng (Javen) Shi

The Australian Centre for Visual Technologies,
The University of Adelaide, Australia

12 August 2011

Outline

- Understanding temporal models
- Dynamic Bayesian Network
- Dynamic Markov Random Fields (including CRFs)

Understanding temporal models

Temporal: of or relating to time.

Temporal models are often seen as

- (deterministic) dynamical systems (a fixed rule) *e.g.* Kalman filter
- or random processes (random walk) *e.g.* HMM,CRFs

Understanding temporal models

A dynamical system uses a fixed rule to describe the time dependence of a point in a geometrical space.

- Origin: Netwonian mechanics.
- Often uses differential equation *e.g.* a flow $\frac{d}{dt}\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{x}(t)$,
  or recurrence relation *e.g.* Fibonacci numbers $F_n = F_{n-1} + F_{n-2}$ with seed values: $F_0 = 0, F_1 = 1$.

Understanding temporal models

Kalman filter assumes the true state at time t is evolved from the state at $(t-1)$ according to

$$\mathbf{x}^t = \mathbf{F}^t \mathbf{x}^{t-1} + \mathbf{B}^t \mathbf{u}^t + \mathbf{w}^t,$$

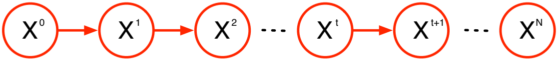
where $\mathbf{F}^t$ is the state transition model, $\mathbf{B}^t$ is the control-input model applied to the control vector $\mathbf{u}^t$, $\mathbf{w}^t$ is the noise from $\mathcal{N}(0, \mathbf{Q}^t)$.

Hence it is seen as a dynamical system. Details of Kalman filter and particle filter will be deferred to later part of the talk.

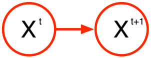
Understanding temporal models

A random process is a collection $\{X^t : t \in T\}$, where each X^t is a random variable.
If $\Pr(X^{t+1}|X^{0:t}) = \Pr(X^{t+1}|X^t)$, it's markovian.
If $\Pr(X^{t+1}|X^t) = \Pr(X^{t'+1}|X^{t'})$ for all t, t', it's homogenous.
A markov chain is a discrete-time random process with markovian and homogenous assumptions.

A chain of length $N + 1$

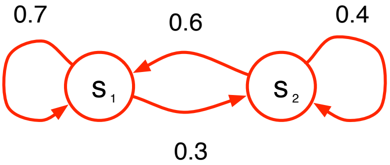


Essentially modeling (due to homogenous assumption)



Understanding temporal models

Transition graph (assuming $X \in \{s_1, s_2\}$) of the previous markov chain is:

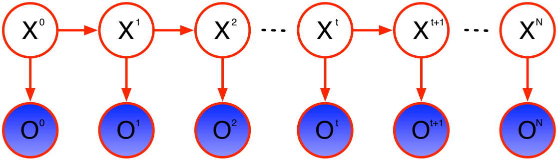


Transition graph $\neq$ graphical model

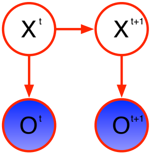
Understanding temporal models

HMM (with hidden state X^t and observation O^t):

A HMM chain of length $N+1$



Essentially modeling (due to homogenous assumption)



This is called 2-time-slice Bayesian network (2-TBN).

Dynamic Bayesian Network

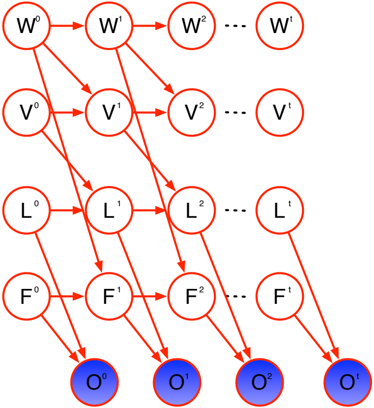
A Dynamic Bayesian Network (DBN) is a pair $(\mathcal{B}_0, \mathcal{B}_\rightarrow)$, where $\mathcal{B}_0$ is a Bayesian network over $\mathbf{Z}^0$ representing the initial distribution over states, and $\mathcal{B}_\rightarrow$ is a 2-TBN for the process representing the transition between states over $\mathbf{Z}^{t:t+1}$. For any $t \geq 0$, the distribution over $\mathbf{Z}^{0:t}$ is a unrolled Bayesian network. See an example.

Dynamic Bayesian Network

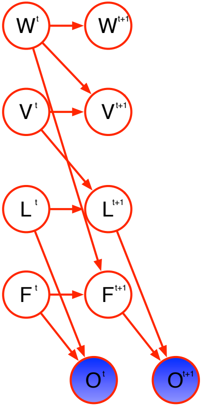
A moving car tries to track its current location using the data obained from a possibly faulty sensor with states set $\{$Weather, Velocity, Location, Failure, Observation$\}$

$\mathbf{Z}^t = (W^t, V^t, L^t, F^t, O^t)$

An unrolled DBN (over $\mathbf{Z}^{0:t}$)



2-TBN ($\mathcal{B}_\rightarrow$)

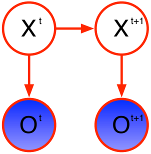


Dynamic Bayesian Network

Given training data, the parameters of the DBN can be learnt via techniques in tutorial (3). Once parameters are learnt, the prediction can be done via techniques in tutorial (2).

Dynamic Bayesian Network

Kalman filter as a HMM.



Markov assumption: $p(\mathbf{x}^{t+1} \mid \mathbf{x}^0, \cdots, \mathbf{x}^t) = p(\mathbf{x}^{t+1} \mid \mathbf{x}^t)$ and $p(\mathbf{o}^{t+1} \mid \mathbf{x}^0, \cdots, \mathbf{x}^{t+1}) = p(\mathbf{o}^{t+1} \mid \mathbf{x}^{t+1})$

$$\text{MAP:} \quad p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t+1}) = \frac{p(\mathbf{o}^{t+1} \mid \mathbf{x}^{t+1}) p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t})}{p(\mathbf{o}^{t+1} \mid \mathbf{o}^{0:t})},$$

$$\text{where} \quad p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t}) = \int p(\mathbf{x}^{t+1}, \mathbf{x}^t \mid \mathbf{o}^{0:t}) d\,\mathbf{x}^t = \int p(\mathbf{x}^{t+1} \mid \mathbf{x}^t) p(\mathbf{x}^t \mid \mathbf{o}^{0:t}) d\,\mathbf{x}^t,$$

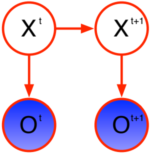
$$\text{and} \quad p(\mathbf{o}^{t+1} \mid \mathbf{o}^{0:t}) = \int p(\mathbf{o}^{t+1} \mid \mathbf{x}^{t+1}) p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t}) d\,\mathbf{x}^{t+1} \quad \text{constant w.r.t. } \mathbf{x}$$

The remaining terms are ready to compute:

$$p(\mathbf{o}^{t+1} \mid \mathbf{x}^{t+1}) = \mathcal{N}(\mathbf{H}^{t+1}\,\mathbf{x}^{t+1}, \mathbf{R}^{t+1})$$
$$p(\mathbf{x}^{t+1} \mid \mathbf{x}^t) = \mathcal{N}(\mathbf{F}^{t+1}\,\mathbf{x}^t, \mathbf{Q}^{t+1})$$
$$p(\mathbf{x}^t \mid \mathbf{o}^{0:t}) = \mathcal{N}(\hat{X}^t, \mathbf{P}^t)$$

Dynamic Bayesian Network

Particle filter as a HMM.



Markov assumption: $p(\mathbf{x}^{t+1} \mid \mathbf{x}^0, \cdots, \mathbf{x}^t) = p(\mathbf{x}^{t+1} \mid \mathbf{x}^t)$ and $p(\mathbf{o}^{t+1} \mid \mathbf{x}^0, \cdots, \mathbf{x}^{t+1}) = p(\mathbf{o}^{t+1} \mid \mathbf{x}^{t+1})$

MAP: $p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t+1}) = \dfrac{p(\mathbf{o}^{t+1} \mid \mathbf{x}^{t+1}) p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t})}{p(\mathbf{o}^{t+1} \mid \mathbf{o}^{0:t})}$,

where $p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t}) = \int p(\mathbf{x}^{t+1}, \mathbf{x}^t \mid \mathbf{o}^{0:t}) d\,\mathbf{x}^t = \int p(\mathbf{x}^{t+1} \mid \mathbf{x}^t) p(\mathbf{x}^t \mid \mathbf{o}^{0:t}) d\,\mathbf{x}^t = \mathbb{E}_{\mathbf{x}^t \sim p(\mathbf{x}^t \mid \mathbf{o}^{0:t})}[p(\mathbf{x}^{t+1} \mid \mathbf{x}^t)]$,

and $p(\mathbf{o}^{t+1} \mid \mathbf{o}^{0:t}) = \int p(\mathbf{o}^{t+1} \mid \mathbf{x}^{t+1}) p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t}) d\,\mathbf{x}^t$

When $p(\mathbf{x}^t \mid \mathbf{o}^{0:t})$ is not gaussian, one can use monte carlo (N samples $\{\mathbf{x}_i^t\}_{i=1}^N$) to approximate the expectation

$$\mathbb{E}_{\mathbf{x}^t \sim p(\mathbf{x}^t \mid \mathbf{o}^{0:t})}[f(\mathbf{x}^t)] \approx \frac{1}{N}[\sum_{i=1}^N f(\mathbf{x}_i^t)],$$

where $f(\mathbf{x}^t) = p(\mathbf{x}^{t+1} \mid \mathbf{x}^t)$.

Dynamic Bayesian Network

Particle filter uses Sequential Importance Resampling (SIR)
Step 1: sampling and computing weights.

$$\text{Sample } \mathbf{x}_i^t \sim q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t}) \approx p(\mathbf{x}^t \mid \mathbf{x}^{t-1}),$$

$$w_i^t = w_i^{t-1} \cdot \frac{p(\mathbf{o}^t \mid \mathbf{x}_i^t) p(\mathbf{x}_i^t \mid \mathbf{x}_i^{t-1})}{q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t})}.$$

Step 2: resampling. If not enough particles $\{\mathbf{x}_i^t\}_i$, resample $\{\mathbf{y}_i^t\}_i$ from the current particle set $\{\mathbf{x}_i^t\}_i$ with probability proportional to $\{w_i\}_i$. Replace the current particle set with the new one $\{\mathbf{y}_i^t\}_i$.

$$\begin{aligned} p(\mathbf{x}^{t+1} \mid \mathbf{o}^{0:t}) &= \mathbb{E}_{\mathbf{x}^t \sim p(\mathbf{x}^t \mid \mathbf{o}^{0:t})}[p(\mathbf{x}^{t+1} \mid \mathbf{x}^t)], \\ &\approx \frac{1}{\sum_{i=1}^N w_i^t}[\sum_{i=1}^N w_i^t p(\mathbf{x}^{t+1} \mid \mathbf{x}_i^t)] \approx \frac{1}{N}[\sum_{i=1}^N p(\mathbf{x}^{t+1} \mid \mathbf{y}_i^t)] \end{aligned}$$

The way that w_i^t is generated may look strange at first sight. Let's crack it!

Dynamic Bayesian Network

In importance sampling, one wants to sample (but hard to) $\mathbf{x} \sim p(\mathbf{x})$, and it's easy to sample $\mathbf{x} \sim q(\mathbf{x})$.

$$\mathbb{E}_p[f(\mathbf{x})] = \mathbb{E}_q[\frac{p(\mathbf{x})}{q(\mathbf{x})} \cdot f(\mathbf{x})] \approx \frac{\sum_{i=1}^N w_i \cdot f(\mathbf{x}_i)}{\sum_{i=1}^N w_i}$$

The weight

$$w_i = \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}.$$

In particle filter, at time t and $t-1$,

$$w^t = \frac{p(\mathbf{x}^{0:t} \mid \mathbf{o}^{0:t})}{q(\mathbf{x}^{0:t} \mid \mathbf{o}^{0:t})},$$

$$w^{t-1} = \frac{p(\mathbf{x}^{0:t-1} \mid \mathbf{o}^{0:t-1})}{q(\mathbf{x}^{0:t-1} \mid \mathbf{o}^{0:t-1})}.$$

Dynamic Bayesian Network

$$\frac{w^t}{w^{t-1}} = \frac{p(\mathbf{x}^{0:t} \mid \mathbf{o}^{0:t})}{q(\mathbf{x}^{0:t} \mid \mathbf{o}^{0:t})} \frac{q(\mathbf{x}^{0:t-1} \mid \mathbf{o}^{0:t-1})}{p(\mathbf{x}^{0:t-1} \mid \mathbf{o}^{0:t-1})}$$

$$= \frac{p(\mathbf{x}^t, \mathbf{o}^t, \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t-1})}{p(\mathbf{x}^{0:k-1}, \mathbf{o}^{0:t-1})} \cdot \frac{p(\mathbf{o}^{0:t-1})}{p(\mathbf{o}^{0:t})} \cdot \frac{1}{q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t})}$$

$$= p(\mathbf{x}^t, \mathbf{o}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t-1}) \cdot \frac{p(\mathbf{o}^{0:t-1})}{p(\mathbf{o}^{0:t})} \cdot \frac{1}{q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t})}$$

$$= p(\mathbf{x}^t, \mathbf{o}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t-1}) \cdot const \cdot \frac{1}{q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t})}$$

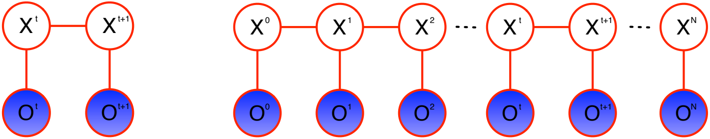
$$= p(\mathbf{x}^t, \mathbf{o}^t \mid \mathbf{x}^{t-1}, \mathbf{o}^{t-1}) \cdot const \cdot \frac{1}{q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t})}$$

$$= p(\mathbf{x}^t, \mathbf{o}^t \mid \mathbf{x}^{t-1}) \cdot const \cdot \frac{1}{q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t})}$$

$$\therefore w^t = w^{t-1} \cdot \frac{p(\mathbf{o}^t \mid \mathbf{x}^t) p(\mathbf{x}^t \mid \mathbf{x}^{t-1})}{q(\mathbf{x}^t \mid \mathbf{x}^{0:t-1}, \mathbf{o}^{0:t})} \cdot const.$$

Dynamic Markov Random Fields

CRFs:



$$\mathbf{P}(\mathbf{y} \mid \mathbf{x}; \mathbf{w}) = \frac{\exp(\langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle)}{Z(\mathbf{w} \mid \mathbf{x})}, \tag{1}$$

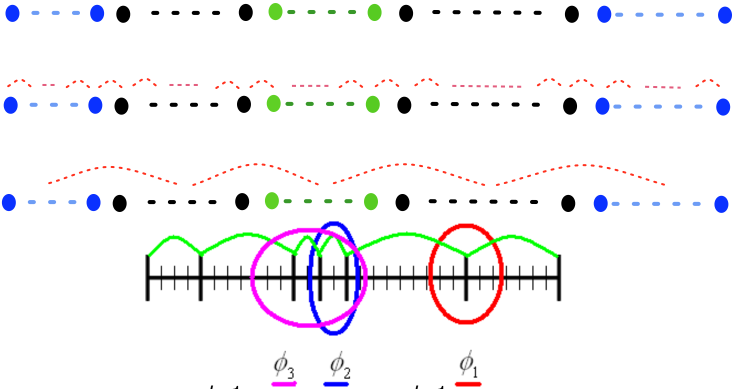
where

$$Z(\mathbf{w} \mid \mathbf{x}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp(\langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}') \rangle), \tag{2}$$

and

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} \Phi_1(\mathbf{x}, \mathbf{y}^{(i)}) + \sum_{(ij) \in \mathcal{E}} \Phi_2(\mathbf{x}, \mathbf{y}^{(ij)}). \tag{3}$$

Semi-Markov Model: Using (**??**) with



$$\Phi(\mathbf{x}, \mathbf{y}) = \Bigg(\sum_{i=0}^{l-1} \Phi_1(\mathbf{x}, n_i, c_i), \sum_{i=0}^{l-1} \Phi_2(\mathbf{x}, n_i, n_{i+1}, c_i), \sum_{i=0}^{l-1} \Phi_3(\mathbf{x}, n_i, n_{i+1}, c_i, c_{i+1}) \Bigg).$$