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

- Origin: Newtonian mechanics.

- Often uses **differential equation** *e.g.* a flow

\[
\frac{d}{dt} x(t) = A \cdot x(t),
\]

or **recurrence relation** *e.g.* Fibonacci numbers

\[
F_n = F_{n-1} + F_{n-2} \text{ with seed values: } F_0 = 0, F_1 = 1.
\]
Kalman filter assumes the true state at time $t$ is evolved from the state at $(t - 1)$ according to

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

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

Hence it is seen as a dynamical system. Details of Kalman filter and particle filter will be deferred to later part of the talk.
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 \( N + 1 \)

\[ X^0 \rightarrow X^1 \rightarrow X^2 \rightarrow \cdots \rightarrow X^t \rightarrow X^{t+1} \rightarrow \cdots \rightarrow X^N \]

Essentially modeling (due to homogenous assumption)

\[ X^t \rightarrow X^{t+1} \]
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 $N + 1$

$$X^0 \rightarrow X^1 \rightarrow X^2 \cdots \rightarrow X^t \rightarrow X^{t+1} \cdots \rightarrow X^N$$

$O^0 \rightarrow O^1 \rightarrow O^2 \cdots \rightarrow O^t \rightarrow O^{t+1} \cdots \rightarrow O^N$

Essentially modeling (due to homogenous assumption)

$$X^t \rightarrow X^{t+1}$$

$O^t \rightarrow O^{t+1}$

This is called 2-time-slice Bayesian network (2-TBN).
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.
A moving car tries to track its current location using the data obtained from a possibly faulty sensor with states set \( \{ \text{Weather}, \text{Velocity}, \text{Location}, \text{Failure}, \text{Observation} \} \)

\[
Z^t = (W^t, V^t, L^t, F^t, O^t)
\]

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

2-TBN (\( \mathcal{B} \))
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: $p(x^{t+1} | x^0, \ldots, x^t) = p(x^{t+1} | x^t)$ and $p(o^{t+1} | x^0, \ldots, x^{t+1}) = p(o^{t+1} | x^{t+1})$

MAP: $p(x^{t+1} | o^{0:t+1}) = \frac{p(o^{t+1} | x^{t+1}) p(x^{t+1} | o^{0:t})}{p(o^{t+1} | o^{0:t})}$,

where $p(x^{t+1} | o^{0:t}) = \int p(x^{t+1}, x^t | o^{0:t}) dx^t = \int p(x^{t+1} | x^t) p(x^t | o^{0:t}) dx^t$,

and $p(o^{t+1} | o^{0:t}) = \int p(o^{t+1} | x^{t+1}) p(x^{t+1} | o^{0:t}) dx^{t+1}$ constant w.r.t. $x$

The remaining terms are ready to compute:

\[ p(o^{t+1} | x^{t+1}) = \mathcal{N}(H^{t+1} x^{t+1}, R^{t+1}) \]
\[ p(x^{t+1} | x^t) = \mathcal{N}(F^{t+1} x^t, Q^{t+1}) \]
\[ p(x^t | o^{0:t}) = \mathcal{N}(\hat{x}^t, P^t) \]
Dynamic Bayesian Network

Particle filter as a HMM.

Markov assumption: \( p(x^{t+1} | x^0, \ldots, x^t) = p(x^{t+1} | x^t) \) and \( p(o^{t+1} | x^0, \ldots, x^{t+1}) = p(o^{t+1} | x^{t+1}) \)

MAP: \( p(x^{t+1} | o^{0:t+1}) = \frac{p(o^{t+1} | x^{t+1}) p(x^{t+1} | o^{0:t})}{p(o^{t+1} | o^{0:t})} \),

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

and \( p(o^{t+1} | o^{0:t}) = \int p(o^{t+1} | x^{t+1}) p(x^{t+1} | o^{0:t}) dx^t \)

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

\[
\mathbb{E}_{x^t \sim p(x^t | o^{0:t})} [f(x^t)] \approx \frac{1}{N} \left[ \sum_{i=1}^N f(x^t_i) \right],
\]

where \( f(x^t) = p(x^{t+1} | x^t) \).
Particle filter uses **Sequential Importance Resampling** (SIR).

**Step 1: sampling and computing weights.**

Sample \( x^t_i \sim q(x^t | x^{0:t-1}, o^{0:t}) \approx p(x^t | x^{t-1}) \),

\[
w^t_i = w^{t-1}_i \cdot \frac{p(o^t | x^t_i)p(x^t_i | x^{t-1}_i)}{q(x^t | x^{0:t-1}, o^{0:t})}.
\]

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

\[
p(x^{t+1} | o^{0:t}) = \mathbb{E}_{x^t \sim p(x^t | o^{0:t})} [p(x^{t+1} | x^t)],
\]

\[
\approx \frac{1}{\sum_{i=1}^N w^t_i} \left[ \sum_{i=1}^N w^t_i p(x^{t+1} | x^t_i) \right] \approx \frac{1}{N} \left[ \sum_{i=1}^N p(x^{t+1} | y^t_i) \right]
\]

The way that \( w^t_i \) is generated may look **strange** at first sight. Let’s crack it!
In importance sampling, one wants to sample (but hard to) \( x \sim p(x) \), and it's easy to sample \( x \sim q(x) \).

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

The weight

\[
w_i = \frac{p(x_i)}{q(x_i)}.
\]

In particle filter, at time \( t \) and \( t-1 \),

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

\[
w^{t-1} = \frac{p(x^{0:t-1} \mid o^{0:t-1})}{q(x^{0:t-1} \mid o^{0:t-1})}.
\]
\[
\frac{w^t}{w^{t-1}} = \frac{p(x^{0:t} \mid o^{0:t}) q(x^{0:t-1} \mid o^{0:t-1})}{q(x^{0:t} \mid o^{0:t}) p(x^{0:t-1} \mid o^{0:t-1})} = \frac{p(x^t, o^t, x^{0:t-1}, o^{0:t-1})}{p(x^{0:k-1}, o^{0:t-1})} \cdot \frac{p(o^{0:t-1})}{p(o^{0:t})} \cdot \frac{1}{q(x^t \mid x^{0:t-1}, o^{0:t})} \cdot \frac{1}{q(x^t \mid x^{0:t-1}, o^{0:t})} \cdot \frac{1}{q(x^t \mid x^{0:t-1}, o^{0:t})} = \frac{p(x^t, o^t \mid x^{0:t-1}, o^{0:t-1}) \cdot \text{const}}{q(x^t \mid x^{0:t-1}, o^{0:t})} = \frac{p(x^t, o^t \mid x^{t-1}, o^{t-1}) \cdot \text{const}}{q(x^t \mid x^{0:t-1}, o^{0:t})} = \frac{p(x^t, o^t \mid x^{t-1}) \cdot \text{const}}{q(x^t \mid x^{0:t-1}, o^{0:t})} = w^t = w^{t-1} \cdot \frac{p(o^t \mid x^t)p(x^t \mid x^{t-1})}{q(x^t \mid x^{0:t-1}, o^{0:t})} \cdot \text{const}.
\]
CRFs:

\[
P(y \mid x; w) = \frac{\exp(\langle w, \Phi(x, y) \rangle)}{Z(w \mid x)},
\]

where

\[
Z(w \mid x) = \sum_{y' \in y} \exp(\langle w, \Phi(x, y') \rangle),
\]

and

\[
\Phi(x, y) = \sum_{i \in \mathcal{V}} \Phi_1(x, y^{(i)}) + \sum_{(ij) \in \mathcal{E}} \Phi_2(x, y^{(ij)}).
\]
Semi-Markov Model: Using (??) with

\[
\Phi(x, y) = \left( \sum_{i=0}^{l-1} \Phi_1(x, n_i, c_i), \sum_{i=0}^{l-1} \Phi_2(x, n_i, n_{i+1}, c_i), \sum_{i=0}^{l-1} \Phi_3(x, n_i, n_{i+1}, c_i, c_{i+1}) \right).
\]