EE595A – DGMs – Winter 2010

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Fall Quarter, 2009
http://ssli.ee.washington.edu/~bilmes/ee512fa09

Lecture 1 - January 13th, 2010

Last updated $Id: dlec1.tex,v 1.5 2010/01/14 23:53:54 bilmes Exp$
The plan for this course is to serve as a thorough overview of dynamic graphical models including hidden Markov models, dynamic Bayesian networks, sequential conditional random fields, Kalman filters, switching Kalman filters, and linear and non-linear dynamical systems. We will also discuss various other time-series methods.

For above models, we will discuss exact and approximate inference as distinct from static GMs. E.g., forward/backwards, temporal junction trees, dynamic triangulation, conditioning and factorization approaches, variational approaches, sampling and particle filter approaches (including importance sampling and Rao-Blackwellization), beam pruning strategies, multi-pass course-to-fine, island algorithm and time-space trade-offs.
Training objectives: generative and discriminative training, MMIE, MDI, and MCE, and the various model-specific approximations, structured max-margin approaches.

Interplay between models and training objectives (e.g., label-bias vs. its training objective).

Applications: coming mainly from speech recognition, natural language processing, bio-informatics and some from econometrics and robotics (i.e., localization and mapping).

Course web page:
http://ssli.ee.washington.edu/~bilmes/ee595Awi10/ which is where readings will be posted.
Homework/Project

- There will be no planned scheduled homeworks in this class, although there will be some challenge problems for you to work on if you wish. If you would like to do some problems and have me grade them, ask me.

- Your grade will be based on a combination of attendance, class participation (both of which I will track), and the quality of the final project.
Homework/Project

- Project should ideally be on some aspect of the material we have learnt, some aspect of dynamic graphical models. Possible good projects include:
  - an implementation (i.e., a fast implementation of some DGMs algorithm) and reporting and experience that you gain in doing this. Application to real data.
  - A paper summary, of papers that we are not going to cover in this class.
  - A new idea of your own, new algorithms and/or theoretical results. (e.g., approximation error for a sequential model).
  - Application of a DGM to a data domain (e.g., application of dynamic Bayesian networks to speech/language/biology/surgery or some other sequential data domain).
- The ideal project should be research-oriented, it is not acceptable to propose whatever machine learning task you are currently working on (e.g., “An application of SVMs to protein folding” would not be acceptable).
Read Wainwright/Jordan book chaps 3/4/5
http://dx.doi.org/10.1561/2200000001

Read “tree_inference.pdf”
Read “evidence.pdf”
Read “dgms.pdf”
Read “ugms.pdf”.
Read “intro.pdf”.
Read relevant chapters in Koller/Friedman text.
Dynamic Readings


Readings are in a directory “reading\_drafts” off of our main web page. uid and pwd are named after this class.
We need to find two makeup lectures this term. We’ll spread them over two weeks.

- L1 (1/13): overview, intro, Markov
- L2 (1/15): 
- L3 (1/20): 
- L4 (1/22): 
- L5 (1/27): 
- L6 (1/29): 
- L7 (2/3): 
- L8 (2/5): 
- L9 (2/10): 
- L10 (2/12): 
- L11 (2/17): 
- L12 (2/19): 
- L13 (2/24): 
- L14 (2/26): 
- L15 (3/3): 
- L16 (3/5): 
- L17 (3/10): 
- L18 (3/12): 
- L19 (makeup): 
- L20 (makeup): 
- L21 (3/15): Final Presentations
Graphical Models

- Graphical model: a graph $G = (V, E)$ & set of rules $R$ that define family $\mathcal{F}(G, R)$ of distributions that abide by rules on the graph.

- Many types of graphs (e.g., DAG, undirected, bipartite) and sets of rules, leading to different families of graphical model (Bayesian network, Markov random fields, chain graphs, etc.)

- Key aspects of a graphical model is that some graph-theoretic property helps us to deduce an inference algorithm on any member of the family — true for both exact and approximate inference.

- Often it has to do with clusterings and/or partitionings of the nodes.

- A goal of graphical model inference - produce generic algorithm

![Diagram]

- Any $p \in \mathcal{F}(G,R)$

- A particular query

- $(G,R)$

- Produce Graphical Model Inference Procedure

- $\text{inference}(p)$ an algorithm for doing Inference

- Correct answer
What is similar to static graphical model case?

- A graph $G$ and set of rules $R$ define a family of probability distributions.
- Goal is to deduce a generic algorithm that can perform inference efficiently on any member of the corresponding family.
- Ideal goal is to use graph-theoretic concepts in doing so.
- Graph visually depicts high-level information regarding “who may directly relate to whom”, high-level domain-specific interpretable.
Dynamic Graphical Models I

What is different from static graphical model case?

- A “graph” G and rules R is really an expandable “template” than an object directly corresponding to the random variables.
- An expanded template corresponds to a graph in a GM in the normal sense, and rules also say how to do expansion for given length.
- Distributions in the same family correspond to variable (unbounded) number of random variables. RVs can expand unboundedly in time, but still a member of the family.
- Graph is shaped differently. Much wider than it is taller, often there is a “time” parameter T that expands the template. For online inference, rules say how to expand a next chunk.
- Typically some form of parameter sharing (otherwise, unbounded time would require unbounded number of parameters).
Still want to deduce computational properties from the static template, rather than have to expand the template to each possible length, and then (re-)deduce an inference algorithm there (would be wasteful).
Various forms of DGM

- Markov Chain
- Martingales
- Hidden Markov Model & Kalman Filter
- Dynamic Bayesian Network
- (Partially Observable) Markov Decision Process (MDP/POMDP)
- (Linear) Dynamical System
- Dynamic Markov Random Field
- Dynamic Conditional Random Field (CRF)
- Switching Linear Dynamical System
- Non-linear dynamical systems
- Sequentially Structured Kernels/RHKS/SVMs & Structured Prediction, string kernels, Fisher kernels, etc.
In each case, a graph template can describe many important properties of members of these families. We will become fairly well versed in almost all of the above.

In each case, some form of “the past is independent of the future given the present”, for various definitions of past, present, and future.
Markov Chains

- One of the simplest forms of time-series models
- The “present” is fairly local, only the current state.
- Can exist in various orders. Example, 1st, 2nd, and 3rd order Markov chain.
A Markov chain with other variables hanging off of it.

Typically, the Markov chain is hidden, and what is observed are the $X$ variables.

Markov Chain is typically 1st order, but extensions exist.

Kalman filter is identical, except it is joint Gaussian (inference in joint Gaussian is incredibly easy).

Multiple possible training algorithms (e.g., some discriminative and some generative) are available, each with computational implications.

We’ll talk extensively about HMMs (see hmm.pdf) writeup.
Switches direction of arrow in BN representation of an HMM, not possible to precisely represent with a MRF.

Original goal: attempt to be more “discriminative”, but this was a poor motivation.
Conditional Markov Model/Maximum Entropy Markov Model II

- Can sometimes suffer from what is known as “label bias” or “observation bias”, but we’ll see that this is much less often a problem than commonly thought.

- Often used in NLP tagging tasks (e.g., take a sentence and produce the part-of-speech tags, like noun, verb, etc.).
The hidden state is inherently and insistently factored rather than monolithic.
Inference within a given frame might require some smarts, more akin to static graphical models.

From a data-domain and/or scientific modeling perspective, much more flexible and powerful than an HMM.

Sometimes, this can lead to significant computational advantages over the HMM as well (we will discuss when this is the case precisely).

Can also lead to much better parameter estimation (bias/variance), as factorization can act as a kind of regularization.

Multiple possible training algorithms are available, each with computational implications.
Dynamic Conditional Random Field I

\[ p(y|x) = p(y_{1:T}|x_{1:T}) = \frac{1}{Z(x)} \prod_{t=1}^{T} g(x_t, y_t) \prod_{t=2}^{T} h(y_t, y_{t-1}) \]  \hspace{1cm} (1)

\[ = \frac{1}{Z(x)} \prod_{t} h(y_t, y_{t-1}) g(x_t, y_t) \]  \hspace{1cm} (2)

- This is a conditional model only, not a joint model of \( x, y \).
- We need only that the hidden variables \( y \) factorize, the input \( \tilde{x} \) is always observed, so does not contribute to state space.
- Often convex, so relatively easy to optimize (many simple iterative approaches, e.g., perceptron updates or gradient steps).
- Multiple (imperfect) ways of drawing this in GM notation.
Dynamic Conditional Random Field II

Diagram showing conditional random fields with labeled variables.

Diagram showing the same conditional random fields with different variable connections.

Diagram showing a linear sequence of variables.
A CRF with hidden variables

No longer convex

Potentially much more powerful but much more computationally complex
A Kalman filter with an additional hidden Markov chain.

We have that \( p(x, h|y) \) is a Kalman filter, and \( y \) is a discrete Markov chain.

Can be considerably more complex than either an HMM or a Kalman filter since we now have switching hidden random continuous random variables, sampling approaches make it relatively easy.
Markov Chains I

Definition

Independent and Identically Distributed (i.i.d.) The stochastic process is said to be i.i.d. if the following condition holds:

\[
p(X_t = x_t, X_{t+1} = x_{t+1}, \ldots, X_{t+h} = x_{t+h}) = \prod_{i=0}^{h} p(X = x_{t+i})
\]

for all \( t \), for all \( h \geq 0 \), for all \( x_{t:t+h} \), and for some distribution \( p(\cdot) \) that is independent of the index \( t \).

i.i.d. processes satisfy exchangability

\[
p(X_t = x_t, X_{t+1} = x_{t+1}, \ldots, X_{t+h} = x_{t+h}) = p(X_{\sigma(t)} = x_t, X_{\sigma(t+1)} = x_{t+1}, \ldots)
\]
Stationary I

**Definition**

**Stationary Stochastic Process** The stochastic process \( \{X_t : t \geq 1\} \) is said to be (strongly) stationary if the two collections of random variables

\[
\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}
\]

and

\[
\{X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}\}
\]

have the same joint probability distributions for all \( n \) and \( h \).

In the discrete case, stationarity is equivalent to the condition

\[
P(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n) = P(X_{t_1+h} = x_1, X_{t_2+h} = x_2, \ldots, X_{t_n+h} = x_n)
\]

(8)

(9)
for all \( t_1, t_2, \ldots, t_n \), for all \( n > 0 \), for all \( h > 0 \), and for all \( x_i \). Every i.i.d. processes is stationary.
Definition Markov Chain

**Definition**

Markov chain A collection of discrete-valued random variables \( \{Q_t : t \geq 1\} \) forms an \( n^{th} \)-order Markov chain if

\[
P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, \ldots, Q_1 = q_1) = 
\]

for all \( t \geq 1 \), and all \( q_1, q_2, \ldots, q_t \).
Markov Chains I

Definition

Markov chain A collection of discrete-valued random variables \( \{ Q_t : t \geq 1 \} \) forms an \( n^{th} \)-order Markov chain if

\[
P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, \ldots, Q_1 = q_1) =
\]

\[
P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, \ldots, Q_{t-n} = q_{t-n})
\]

for all \( t \geq 1 \), and all \( q_1, q_2, \ldots, q_t \).

- does not imply that a variable is independent of future variable — on the contrary, variable might be quite dependent on future variables.
- only states: variable is independent of the long-past given the recent past.
Markov Chains II

- Example: Bayesian networks of $n^{th}$-order Markov chains for various $n$. Left: 1st order. Middle: 2nd order. Right: 3rd order. Zero’th order is i.i.d..

- We view the event $\{Q_t = i\}$ as if the chain is “in state $i$ at time $t$” and the event $\{Q_t = i, Q_{t+1} = j\}$ as a transition from state $i$ to state $j$ starting at time $t$. This notion arises by viewing a Markov chain as a finite-state automata (FSA).
Order Conversion I

An $n^{th}$-order Markov chain may be converted into equivalent first-order Markov chain via:

$$Q'_t \triangleq \{ Q_t, Q_{t-1}, \ldots, Q_{t-n} \}$$

where $Q_t$ is an $n^{th}$-order Markov chain. Then $Q'_t$ is a first-order Markov chain because

$$P(Q'_t = q'_t | Q'_{t-1} = q'_{t-1}, Q'_{t-2} = q'_{t-2}, \ldots, Q'_1 = q'_1)$$

$$= P(Q_{t-n:t} = q_{t-n:t} | Q_{1:t-1} = q_{1:t-1})$$

$$= P(Q_{t-n:t} = q_{t-n:t} | Q_{t-n-1:t-1} = q_{t-n-1:t-1})$$

$$= P(Q'_t = q'_t | Q'_{t-1} = q'_{t-1})$$

Given a large enough state space, first-order Markov chain may represent any $n^{th}$-order Markov, but with exponential in $n$ state space.
Homogeneous I

- The statistical evolution of a 1st-order Markov chain is determined by the state transition probabilities $a_{ij}(t) \triangleq P(Q_t = j | Q_{t-1} = i)$.

- function both of the states at successive time steps and of the current time $t$.

- sometimes, no dep. on $t$, called *time-homogeneous* (or just *homogeneous*) because $a_{ij}(t) = a_{ij}$ for all $t$.

- The transition probabilities in homogeneous Markov chains are transition matrix $A$ where $a_{ij} \triangleq (A)_{ij}$. 
State Types I

A state $i$ is said to be *transient* if, after visiting the state, it is possible for it never to be visited again, i.e.,:

$$p(Q_n = i \text{ for some } n > t | Q_t = i) < 1.$$  

A state $i$ is said to be *null-recurrent* if it is not transient but the expected return time is infinite (i.e.,

$$E[\min\{n > t : Q_n = i\} | Q_t = i] = \infty$$

a state is *positive-recurrent* if it is not transient and the expected return time to that state is finite
If $Q_t$ is a time-homogeneous stationary first-order Markov chain then:

$$P(Q_{t_1} = q_1, Q_{t_2} = q_2, \ldots, Q_{t_n} = q_n)$$

$= P(Q_{t_1+h} = q_1, Q_{t_2+h} = q_2, \ldots, Q_{t_n+h} = q_n)$

for all $t_i, h, n,$ and $q_i$. Using the first order Markov property, the above can be written as:

$$P(Q_{tn} = q_n|Q_{tn-1} = q_{n-1})$$

$$P(Q_{tn-1} = q_{n-1}|Q_{tn-2} = q_{n-2})$$

$$\ldots P(Q_{t_2} = q_2|Q_{t_1} = q_1)P(Q_{t_1} = q_1)$$

$$= P(Q_{tn+h} = q_n|Q_{tn-1+h} = q_{n-1})$$

$$P(Q_{tn-1+h} = q_{n-1}|Q_{tn-2+h} = q_{n-2})$$

$$\ldots P(Q_{t_2+h} = q_2|Q_{t_1+h} = q_1)P(Q_{t_1+h} = q_1)$$
Stationarity Again II

- Therefore, a time-homogeneous Markov chain is stationary iff 
  \[ P(Q_{t_1} = q) = P(Q_{t_1+h} = q) = P(Q_t = q) \] for all \( q \in D_Q \).

- A stationary distribution has the property that \( \xi A = \xi \) implying that \( \xi \) must be a left eigenvector of the transition matrix \( A \).

- Example: Let \( p_1 = [0.5, 0.5] \) be the current distribution over 2-state Markov chain. Let \( A_1 = [0.3, 0.7; 0.7, 0.3] \) be the transition matrix. The Markov chain is stationary since \( p_1 A_1 = p_1 \). If the current distribution is \( p_2 = [0.4, 0.6] \), however, then \( p_2 A_1 \neq p_2 \), so the chain is no longer stationary (even with same transition matrix).

- Can exist more than 1 stationary distribution.

- Challenge/HW Problem: Must there exist a stationary distribution in a time-homogeneous Markov chain?
Stationarity implies Homogeneity I

If stationary, then

\[ P(Q_t = i, Q_{t-1} = j) = P(Q_{t-1} = i, Q_{t-2} = j) \]  \hspace{1cm} (22)

and

\[ P(Q_t = i) = P(Q_{t-1} = i) \]  \hspace{1cm} (23)

Therefore,

\[ a_{ij}(t) = \frac{P(Q_t = i, Q_{t-1} = j)}{P(Q_{t-1} = j)} \]  \hspace{1cm} (24)

\[ = \frac{P(Q_{t-1} = i, Q_{t-2} = j)}{P(Q_{t-2} = j)} \]  \hspace{1cm} (25)

\[ = a_{ij}(t - 1) \]  \hspace{1cm} (26)

and by induction \( a_{ij}(t) = a_{ij}(t + \tau) \) for all \( \tau \).
Example

Let $A_t = [0.3, 0.7; 0.7, 0.3]$ when $t$ is even and $A_t = [0.4, 0.6; 0.6, 0.4]$ when $t$ is odd, so chain is inhomogeneous.
If the current state distribution is $p = [0.5, 0.5]$, then $pA_t = p$ for $t$ both even and odd.
Is this stationary?
Example

Let $A_t = [.3, .7; .7, .3]$ when $t$ is even and $A_t = [.4, .6; .6, .4]$ when $t$ is odd, so chain is inhomogeneous.

If the current state distribution is $p = [.5, .5]$, then $pA_t = p$ for $t$ both even and odd.

Is this stationary?

- Note that this is not a stationary distribution. When $t$ is even, we have that $p(Q_t = 0, Q_{t+1} = 1) = 0.5 \times 0.3$ but when $t$ is odd, $p(Q_t = 0, Q_{t+1} = 1) = 0.5 \times 0.4$, so the chain does not exhibit a stationary distribution according to the definition.

- So the aforementioned criterion for stationary ($\xi A = xi$) requires a homogeneous chain. $\xi A = xi$ alone does not guarantee stationarity.