Representation of directed GM

Probabilistic Graphical Models (10-708)

Lecture 1, Sep 12, 2007

Eric Xing

Reading: MJ-Chap 2, KF-Chap. 3

- Recitation?
- Exam dates, poster dates, etc.
- Mailing list
- Questions?
Representing Multivariate Distribution

- Representation: what is the joint probability dist. on multiple variables? 
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \]
  - How many state configurations in total? \( 2^8 \)
  - Are they all needed to be represented? 
  - Do we get any scientific/medical insight?

- Factored representation: the chain-rule
  \[ P(X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_2)P(X_3|X_2)P(X_4|X_3, X_2)P(X_5|X_4, X_3, X_2)P(X_6|X_5, X_4, X_3, X_2)P(X_7|X_6, X_5, X_4, X_3, X_2)P(X_8|X_7, X_6, X_5, X_4, X_3, X_2) \]
  - This factorization is true for any distribution and any variable ordering
  - Do we save any parameterization cost?

- If \( X_i \)'s are independent: \( P(X_i|\cdot) = P(X_i) \)
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)P(X_6)P(X_7)P(X_8) = \prod_i P(X_i) \]
  - What do we gain?
  - What do we lose?

- Even in the simplest case where these variables are binary-valued, a joint distribution requires the specification of \( 2^n \) numbers — the probabilities of the \( 2^n \) different assignments of values \( x_1, \ldots, x_n \)

- Today's lecture is about …
  - how independence properties in the distribution can be used to represent such high-dimensional distributions much more compactly.
  - how a combinatorial data structure — a directed acyclic graph — can provide us with a general-purpose modeling language for exploiting this type of structure in our representation.
Two types of GMs

- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model):

\[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3 | X_2) P(X_4 | X_2) P(X_5 | X_2) P(X_6 | X_5, X_2) P(X_7 | X_6) P(X_8 | X_5, X_6) \]

- Undirected edges simply give correlations between variables (Markov Random Field or Undirected Graphical model):

\[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \frac{1}{Z} \exp \left\{ E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2) + E(X_5, X_2) + E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6) \right\} \]

Specification of a directed GM

- There are two components to any GM:
  - the qualitative specification
  - the quantitative specification

![Diagram of directed GM with nodes A, B, C, D, E, F, G, H and edges showing relationships and probabilities]
Bayesian Network:

- A BN is a directed graph whose nodes represent the random variables and whose edges represent direct influence of one variable on another.

- It is a data structure that provides the skeleton for representing a joint distribution compactly in a factorized way;

- It offers a compact representation for a set of conditional independence assumptions about a distribution;

- We can view the graph as encoding a generative sampling process executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents. In other words, each variable is a stochastic function of its parents.

Bayesian Network: Factorization Theorem

- Theorem:
  Given a DAG, the most general form of the probability distribution that is consistent with the graph factors according to "node given its parents":

  \[ P(X) = \prod_{i=1}^{d} P(X_i \mid X_{pa_i}) \]

  where \( X_{pa_i} \) is the set of parents of \( X_i \), \( d \) is the number of nodes (variables) in the graph.
Qualitative Specification

Where does the qualitative specification come from?

- Prior knowledge of causal relationships
- Prior knowledge of modular relationships
- Assessment from experts
- Learning from data
- We simply link a certain architecture (e.g. a layered graph)
- ...

Local Structures & Independencies

- Common parent
  - Fixing B decouples A and C
    "given the level of gene B, the levels of A and C are independent"

- Cascade
  - Knowing B decouples A and C
    "given the level of gene B, the level gene A provides no
    extra prediction value for the level of gene C"

- V-structure
  - Knowing C couples A and B
    because A can "explain away" B w.r.t. C
    "If A correlates to C, then chance for B to also correlate to B will decrease"

- The language is compact, the concepts are rich!
A simple justification

Defn (3.2.2): Let $P$ be a distribution over $X$. We define $I(P)$ to be the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in $P$ (however how we set the parameter-values).

Defn (3.2.3): Let $K$ be any graph object associated with a set of independencies $I(K)$. We say that $K$ is an I-map for a set of independencies $I$, $I(K) \subseteq I$.

We now say that $G$ is an I-map for $P$ if $G$ is an I-map for $I(P)$, where we use $I(G)$ as the set of independencies associated.
Facts about I-map

- For G to be an I-map of P, it is necessary that G does not mislead us regarding independencies in P:
  
  any independence that G asserts must also hold in P. Conversely, P may have additional dependencies that are not reflected in G.

- Example:

  ![Diagram of two networks P1 and P2 with nodes X, Y, and Z and their relationships.]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P(X, Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x0</td>
<td>y0</td>
<td>0.15</td>
</tr>
<tr>
<td>x0</td>
<td>y1</td>
<td>0.32</td>
</tr>
<tr>
<td>x1</td>
<td>y0</td>
<td>0.12</td>
</tr>
<tr>
<td>x1</td>
<td>y1</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P(X, Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x0</td>
<td>y0</td>
<td>0.4</td>
</tr>
<tr>
<td>x0</td>
<td>y1</td>
<td>0.3</td>
</tr>
<tr>
<td>x1</td>
<td>y0</td>
<td>0.2</td>
</tr>
<tr>
<td>x1</td>
<td>y1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

What is in I(G) ---
local Markov assumptions of BN

- A Bayesian network structure G is a directed acyclic graph whose nodes represent random variables $X_1, \ldots, X_n$.

- **local Markov assumptions**

- **Defn (3.2.1):**

  Let $Pa_x$ denote the parents of $X_i$ in G, and $NonDescendants_x$ denote the variables in the graph that are not descendants of $X_i$. Then G encodes the following set of conditional independence assumptions $I(G)$:

  $$I(G): \{X_i \perp NonDescendants_x \mid Pa_x: \forall i\},$$

  In other words, each node $X_i$ is independent of its nondescendants given its parents.
Graph separation criterion

- D-separation criterion for Bayesian networks (D for Directed edges):

**Defn:** variables $x$ and $y$ are *D-separated* (conditionally independent) given $z$ if they are separated in the *moralized* ancestral graph

- Example:

  original graph $\Rightarrow$ ancestral $\Rightarrow$ moral ancestral
Active trail

- **Causal trail** $X \rightarrow Z \rightarrow Y$: active if and only if $Z$ is not observed.
- **Evidential trail** $X \leftarrow Z \leftarrow Y$: active if and only if $Z$ is not observed.
- **Common cause** $X \leftarrow Z \rightarrow Y$: active if and only if $Z$ is not observed.
- **Common effect** $X \rightarrow Z \leftarrow Y$: active if and only if either $Z$ or one of $Z$'s descendants is observed.

**Definition (3.3.2):** Let $X$, $Y$, $Z$ be three sets of nodes in $G$. We say that $X$ and $Y$ are **$d$-separated** given $Z$, denoted $d$-$\text{sep}_G(X; Y | Z)$, if there is no active trail between any node $X \in X$ and $Y \in Y$ given $Z$.

What is in $I(G)$ ---

**Global Markov properties of BN**

- $X$ is **$d$-separated** (directed-separated) from $Z$ given $Y$ if we can't send a ball from any node in $X$ to any node in $Z$ using the "Bayes-ball" algorithm illustrated below (and plus some boundary conditions):
  
  - Defn: $I(G)$=all independence properties that correspond to $d$-separation:
    
    $$ I(G) = \{X \perp Z | Y : dsepa(X; Z | Y)\} $$
  
  - D-separation is sound and complete (more details later)
Example:

Complete the I(G) of this graph:

\[ \begin{align*}
  &x_1 \\
  &x_2 \\
  &x_3 \\
  &x_4
\end{align*} \]

Summary: Conditional Independence Semantics in an BN

Structure: \textit{DAG}

- Meaning: a node is \textit{conditionally independent} of every other node in the network outside its Markov blanket.
- Local conditional distributions (CPD) and the DAG completely determine the joint dist.
- Give causality relationships, and facilitate a \textit{generative} process.
Toward quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables

- The Equivalence Theorem
  For a graph $G$,
  Let $\mathcal{D}_1$ denote the family of all distributions that satisfy $I(G)$,
  Let $\mathcal{D}_2$ denote the family of all distributions that factor according to $G$,
  
  $$P(X) = \prod_{i=1}^{n} P(X_i | X_{\neg i})$$

  Then $\mathcal{D}_1 \equiv \mathcal{D}_2$.

- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

---

Conditional probability tables (CPTs)

|   | P(a,b,c,d) = | P(a)P(b|c,a,b)P(d|c) |
|---|-------------|------------------------|
|   | $P(a,b,c,d)$ = $P(a)P(b|c,a,b)P(d|c)$ |

<table>
<thead>
<tr>
<th>$a^0$</th>
<th>0.75</th>
<th>$b^0$</th>
<th>0.33</th>
<th>$c^0$</th>
<th>0.55</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a^1$</td>
<td>0.25</td>
<td>$b^1$</td>
<td>0.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a^0b^0$</th>
<th>$a^0b^1$</th>
<th>$a^1b^0$</th>
<th>$a^1b^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c^0$</td>
<td>0.45</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>$c^1$</td>
<td>0.55</td>
<td>0.0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$c^0$</th>
<th>$c^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d^0$</td>
<td>0.3</td>
</tr>
<tr>
<td>$d^1$</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Conditional probability density func. (CPDs)

\[ P(a,b,c,d) = P(a)P(b)P(c|a, b)P(d|c) \]

Summary of BN semantics

- **Defn (3.2.5):** A Bayesian network is a pair \((G, P)\) where \(P\) factorizes over \(G\), and where \(P\) is specified as set of CPDs associated with \(G\)’s nodes.
  - Conditional independencies imply factorization
  - Factorization according to \(G\) implies the associated conditional independencies.
  - Are there other independences that hold for every distribution \(P\) that factorizes over \(G\)?
Soundness and completeness

D-separation is sound and "complete" w.r.t. BN factorization law

**Soundness:**
*Theorem:* If a distribution $P$ factorizes according to $G$, then $I(G) \subseteq I(P)$.

"Completeness":
*"Claim":* For any distribution $P$ that factorizes over $G$, if $(X \perp Y | Z) \in I(P)$ then $d$-$\text{sep}_G(X; Y | Z)$.

Contrapositive of the completeness statement

- "If $X$ and $Y$ are not d-separated given $Z$ in $G$, then $X$ and $Y$ are dependent in all distributions $P$ that factorize over $G."$
- Is this true?

Soundness and completeness

- Contrapositive of the completeness statement
  - "If $X$ and $Y$ are not d-separated given $Z$ in $G$, then $X$ and $Y$ are dependent in all distributions $P$ that factorize over $G."$
  - Is this true?

- No. Even if a distribution factorizes over $G$, it can still contain additional independencies that are not reflected in the structure

- Example: graph $A \rightarrow B$, for actually independent $A$ and $B$ (the independence can be captured by some subtle way of parameterization)

- **Thm:** Let $G$ be a BN graph. If $X$ and $Y$ are not d-separated given $Z$ in $G$, then $X$ and $Y$ are dependent in some distribution $P$ that factorizes over $G.$
Theorem 3.3.6: For almost all distributions $P$ that factorize over $G$, i.e., for all distributions except for a set of "measure zero" in the space of CPD parameterizations, we have that $I(P) = I(G)$.

Uniqueness of BN

Very different BN graphs can actually be equivalent, in that they encode precisely the same set of conditional independence assertions.

$(X \perp Y | Z)$. 
I-equivalence

- **Defn (3.3.9):** Two BN graphs $G_1$ and $G_2$ over $X$ are *I-equivalent* if $I(G_1) = I(G_2)$.

  - The set of all graphs over $X$ is partitioned into a set of mutually exclusive and exhaustive *I-equivalence classes*, which are the set of equivalence classes induced by the I-equivalence relation.

  - Any distribution $P$ that can be factorized over one of these graphs can be factorized over the other.
  - Furthermore, there is no intrinsic property of $P$ that would allow us associate it with one graph rather than an equivalent one.
  - This observation has important implications with respect to our ability to determine the directionality of influence.

Detecting I-equivalence

- **Defn (3.3.10):** The *skeleton* of a Bayesian network graph $G$ over $V$ is an undirected graph over $V$ that contains an edge $\{X, Y\}$ for every edge $(X, Y)$ in $G$.

- **Thm (3.3.11):** Let $G_1$ and $G_2$ be two graphs over $V$. If $G_1$ and $G_2$ have the same skeleton and the same set of v-structures then they are I-equivalent.

  - graph equivalence
  - Same trail
  - But not necessarily active
Minimum I-MAP

- Complete graph is a (trivial) I-map for any distribution, yet it does not reveal any of the independence structure in the distribution.
  - Meaning that the graph dependence is arbitrary, thus by careful parameterization an dependencies can be captured
  - We want a graph that has the maximum possible $I(G)$, yet still $\subseteq I(P)$

- **Defn 3.4.1**: A graph object $G$ is a *minimal I-map* for a set of independencies $I$ if it is an I-map for $I$, and if the removal of even a single edge from $G$ renders it not an I-map.

Minimum I-MAP is not unique
Defn (3.4.3): We say that a graph object $G$ is a perfect map (P-map) for a set of independencies $I$ if we have that $I(G) = I$. We say that $G$ is a perfect map for $P$ if $I(G) = I(P)$.

- The fact that $G$ is a minimal $I$-map for $P$ is far from a guarantee that $G$ captures the independence structure in $P$.

- Not all $P$ has a perfect map as DAG!

- The P-map of a distribution is unique up to $I$-equivalence between networks. That is, a distribution $P$ can have many P-maps, but all of them are $I$-equivalent.

Conditionally Independent Observations

- Model parameters
- Data
“Plate” Notation

Data = \{y_1, \ldots y_n\}

Plate = rectangle in graphical model
variables within a plate are replicated
in a conditionally independent manner

Example: Gaussian Model

Generative model:
\[ p(y_1, \ldots y_n \mid \mu, \sigma) = \prod_i p(y_i \mid \mu, \sigma) = p(\text{data} \mid \text{parameters}) = p(D \mid \theta) \]

where \( \theta = \{\mu, \sigma\} \)

- Likelihood = \( p(\text{data} \mid \text{parameters}) = p(D \mid \theta) = L(\theta) \)
- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
  - Often easier to work with \( \log L(\theta) \)
**Example: Bayesian Gaussian Model**

\[
y_i = \alpha \beta \mu \sigma \]

\[i = 1:n\]

Note: priors and parameters are assumed independent here

---

**Example**

- Speech recognition

\[Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow \ldots \rightarrow Y_T\]

\[X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \ldots \rightarrow X_T\]

Hidden Markov Model
Hidden Markov Model: from static to dynamic mixture models

The underlying source:
Speech signal, dice,

The sequence:
Phonemes, sequence of rolls,
The Dishonest Casino

A casino has two dice:

- **Fair die**
  \[ P(1) = P(2) = P(3) = P(5) = P(6) = \frac{1}{6} \]

- **Loaded die**
  \[ P(1) = P(2) = P(3) = P(5) = \frac{1}{10} \]
  \[ P(6) = \frac{1}{2} \]

Casino player switches back-&-forth between fair and loaded die once every 20 turns

**Game:**
1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins $2

---

A stochastic generative model

- **Observed sequence:**
  ![Observed sequence]

- **Hidden sequence (a parse or segmentation):**
  ![Hidden sequence]
Definition (of HMM)

- **Observation space**
  - Alphabetic set: $C = \{c_1, c_2, \ldots, c_K\}$
  - Euclidean space: $\mathbb{R}^d$

- **Index set of hidden states**
  - $I = \{1, 2, \ldots, M\}$

- **Transition probabilities** between any two states
  - $p(y_i^t | y_i^{t-1} = 1) = a_{i,j}$
  - or $p(y_i | y_i^{t-1} = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \ldots, a_{i,M}) \forall i \in I.$

- **Start probabilities**
  - $p(y_i) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M)$

- **Emission probabilities** associated with each state
  - $p(x_i | y_i^t = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}) \forall i \in I.$
  - or in general:
    - $p(x_i | y_i^t = 1) \sim \mathcal{N}(\theta_i) \forall i \in I.$

---

Puzzles regarding the dishonest casino

**GIVEN:** A sequence of rolls by the casino player

1245266214614636616666616515615115146123562344

**QUESTION**

- How likely is this sequence, given our model of how the casino works?
  - This is the **EVALUATION** problem in HMMs

- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the **DECODING** question in HMMs

- How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?
  - This is the **LEARNING** question in HMMs
Probability of a parse

- Given a sequence \( x = x_1 \ldots x_T \) and a parse \( y = y_1 \ldots y_T \),
- To find how likely is the parse:
  (given our HMM and the sequence)

\[
p(x, y) = p(x_1 \ldots x_T, y_1 \ldots y_T) \quad \text{(Joint probability)}
\]

\[
= p(y_1) p(x_1 | y_1) p(x_2 | y_1) \ldots p(x_T | y_T) \]

\[
= p(y_1) p(y_2 | y_1) \ldots p(y_T | y_{T-1}) \bigtimes p(x_1 | y_1) p(x_2 | y_2) \ldots p(x_T | y_T)
\]

\[
= p(y_1, \ldots, y_T) p(x_1, \ldots, x_T | y_1, \ldots, y_T)
\]

Marginal probability:

\[
p(x) = \sum p(x, y) = \sum_{y_1} \sum_{y_2} \ldots \sum_{y_T} p(y_1) \prod_{i=1}^{T} p(x_i | y_i)
\]

Posterior probability:

\[
p(y | x) = \frac{p(x, y)}{p(x)}
\]

Example, con’d

- Evolution

\[\text{Tree Model}\]
Example, con'd

- Genetic Pedigree

Summary of BN semantics

- Defn (3.2.5): A Bayesian network is a pair (G, P) where P factorizes over G, and where P is specified as set of CPDs associated with G’s nodes.
Knowledge Engineering

- Picking variables
  - Observed
  - Hidden

- Picking structure
  - CAUSAL
  - Generative

- Picking Probabilities
  - Zero probabilities
  - Orders of magnitudes
  - Relative values