Probabilistic graphical models

Yifeng Tao
School of Computer Science
Carnegie Mellon University

Slides adapted from Eric Xing, Matt Gormley
Recap of Basic Probability Concepts

- Representation: the joint probability distribution on multiple binary variables?
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \]

  - State configurations in total: \(2^8\)
  - Are they all needed to be represented?
  - Do we get any scientific/medical insight?

- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation?

- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
  - Computing \(p(H|A)\) would require summing over all \(2^6\) configurations of the unobserved variables

[Slide from Eric Xing.]
Graphical Model: Structure Simplifies Representation

- Dependencies among variables

[Diagram of a graphical model with nodes and edges representing dependencies among variables.

[Slide from Eric Xing.]
If $X_i$'s are **conditionally independent** (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= P(X_1) P(X_2) P(X_3|X_1) P(X_4|X_2) P(X_5|X_2)$$

$$P(X_6|X_3, X_4) P(X_7|X_6) P(X_8|X_5, X_6)$$

Why we may favor a PGM?

- Incorporation of domain knowledge and causal (logical) structures
- $2+2+4+4+4+8+4+8=36$, an 8-fold reduction from $2^8$ in representation cost!
Two types of GMs

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

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

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

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

[Slide from Eric Xing.]
Bayesian Network

**Definition:**

\[ P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid \text{parents}(X_i)) \]

- It consists of a graph \( G \) and the conditional probabilities \( P \).
- These two parts fully specify the distribution:
  - Qualitative Specification: \( G \)
  - Quantitative Specification: \( P \)

[Slide from Eric Xing.]
Where does the qualitative specification come from?

- Prior knowledge of causal relationships
- Learning from data (i.e. structure learning)
- We simply prefer a certain architecture (e.g. a layered graph)
- …

[Slide from Matt Gormley.]
Quantitative Specification

- Example: Conditional probability tables (CPTs) for discrete random variables

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

<table>
<thead>
<tr>
<th>a(^0)</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>a(^1)</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b(^0)</th>
<th>0.33</th>
</tr>
</thead>
<tbody>
<tr>
<td>b(^1)</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c(^0)</th>
<th>a(^0)(^0)</th>
<th>a(^0)(^1)</th>
<th>a(^1)(^0)</th>
<th>a(^1)(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c(^0)</td>
<td>0.45</td>
<td>1</td>
<td>0.9</td>
<td>0.7</td>
</tr>
<tr>
<td>c(^1)</td>
<td>0.55</td>
<td>0</td>
<td>0.1</td>
<td>0.3</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>0.7</td>
</tr>
</tbody>
</table>

[Slide from Eric Xing.]
Example: Conditional probability density functions (CPDs) for continuous random variables

\[ A \sim N(\mu_a, \Sigma_a) \quad B \sim N(\mu_b, \Sigma_b) \]

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

[Slide from Eric Xing.]
In a graphical model, shaded nodes are “observed”, i.e. their values are given.

\[ P(X_2, X_5 \mid X_1 = 0, X_3 = 1, X_4 = 1) \]
GMs are your old friends

- **Density estimation**
  - Parametric and nonparametric methods

- **Regression**
  - Linear, conditional mixture, nonparametric

- **Classification**
  - Generative and discriminative approach

- **Clustering**

[Slide from Eric Xing.]
What Independencies does a Bayes Net Model?

- Independence of $X$ and $Z$ given $Y$?
  $$P(X|Y)P(Z|Y) = P(X,Z|Y)$$
- Three cases of interest...
- Proof?

[Slide from Matt Gormley.]
The “Burglar Alarm” example

- Your house has a twitchy burglar alarm that is also sometimes triggered by earthquakes.
- Earth arguably doesn’t care whether your house is currently being burgled.
- While you are on vacation, one of your neighbors calls and tells you your home’s burglar alarm is ringing.

Quiz: True or False?

\[ \text{Burglar} \perp \text{Earthquake} \mid \text{Phone Call} \]

[Slide from Matt Gormley.]
Markov Blanket

- **Def:** the **co-parents** of a node are the parents of its children.
- **Def:** the **Markov Blanket** of a node is the set containing the node’s parents, children, and co-parents.
- **Thm:** a node is **conditionally independent** of every other node in the graph given its **Markov blanket**.
- **Example:** The Markov Blanket of $X_6$ is \{ $X_3$, $X_4$, $X_5$, $X_8$, $X_9$, $X_{10}$ \}.

[Slide from Matt Gormley.]
**Example:** The Markov Blanket of $X_6$ is
\[ \{X_3, X_4, X_5, X_8, X_9, X_{10}\} \]
D-Separation

- **Thm:** If variables $X$ and $Z$ are d-separated given a set of variables $E$
  - Then $X$ and $Z$ are conditionally independent given the set $E$

- **Definition:**
  - Variables $X$ and $Z$ are d-separated given a set of evidence variables $E$
    - iff every path from $X$ to $Z$ is “blocked”.

A path is “blocked” whenever:

1. $\exists Y$ on path s.t. $Y \in E$ and $Y$ is a “common parent”
   
   ![Diagram of a common parent](image1)

2. $\exists Y$ on path s.t. $Y \in E$ and $Y$ is in a “cascade”
   
   ![Diagram of a cascade](image2)

3. $\exists Y$ on path s.t. $\{Y, \text{descendants}(Y)\} \not\subseteq E$ and $Y$ is in a “v-structure”
   
   ![Diagram of a v-structure](image3)
D-Separation

- Variables $X$ and $Z$ are d-separated given a set of evidence variables $E$ iff every path from $X$ to $Z$ is “blocked”.

[Slide from Eric Xing.]
[Slide from Matt Gormley.]
Recipe for Closed-form MLE

1. Assume data was generated i.i.d. from some model (i.e. write the generative story)
   \[ x^{(i)} \sim p(x | \theta) \]

2. Write log-likelihood
   \[ \ell(\theta) = \log p(x^{(i)} | \theta) + \ldots + \log p(x^{(N)} | \theta) \]

3. Compute partial derivatives (i.e. gradient)
   \[
   \frac{\partial \ell(\theta)}{\partial \theta_1} = \ldots \\
   \frac{\partial \ell(\theta)}{\partial \theta_2} = \ldots \\
   \ldots \\
   \frac{\partial \ell(\theta)}{\partial \theta_M} = \ldots
   \]

4. Set derivatives to zero and solve for \( \theta \)
   \[ \frac{\partial \ell(\theta)}{\partial \theta_m} = 0 \text{ for all } m \in \{1, \ldots, M\} \]
   \[ \theta^{\text{MLE}} = \text{solution to system of } M \text{ equations and } M \text{ variables} \]

5. Compute the second derivative and check that \( \ell(\theta) \) is concave down at \( \theta^{\text{MLE}} \)
Learning Fully Observed BNs

- How do we learn these **conditional** and **marginal** distributions for a Bayes Net?

\[
p(X_1, X_2, X_3, X_4, X_5) = p(X_5 | X_3)p(X_4 | X_2, X_3) p(X_3) p(X_2 | X_1) p(X_1)
\]
Learning Fully Observed BNs

- Learning this fully observed Bayesian Network is equivalent to learning five (small / simple) independent networks from the same data.

\[
p(X_1, X_2, X_3, X_4, X_5) = p(X_5|X_3)p(X_4|X_2, X_3)\]

\[
p(X_3)p(X_2|X_1)p(X_1)\]

[Slide from Matt Gormley.]
Learning Fully Observed BNs

How do we learn these conditional and marginal distributions for a Bayes Net?

\[
\theta^* = \arg\max_{\theta} \log p(X_1, X_2, X_3, X_4, X_5) \\
= \arg\max_{\theta} \log p(X_5|X_3, \theta_5) + \log p(X_4|X_2, X_3, \theta_4) \\
+ \log p(X_3|\theta_3) + \log p(X_2|X_1, \theta_2) \\
+ \log p(X_1|\theta_1)
\]

\[
\theta_1^* = \arg\max_{\theta_1} \log p(X_1|\theta_1)
\]

\[
\theta_2^* = \arg\max_{\theta_2} \log p(X_2|X_1, \theta_2)
\]

\[
\theta_3^* = \arg\max_{\theta_3} \log p(X_3|\theta_3)
\]

\[
\theta_4^* = \arg\max_{\theta_4} \log p(X_4|X_2, X_3, \theta_4)
\]

\[
\theta_5^* = \arg\max_{\theta_5} \log p(X_5|X_3, \theta_5)
\]

[Slide from Matt Gormley.]
Learning Partially Observed BNs

- Partially Observed Bayesian Network:
  - Maximal likelihood estimation $\rightarrow$ Incomplete log-likelihood
  - The log-likelihood contains unobserved latent variables

- Solve with EM algorithm

- Example: Gaussian Mixture Models (GMMs)

$$p(x_n | \mu, \Sigma) = \sum_k \pi_k N(x, \mu_k, \Sigma_k)$$

[Slide from Eric Xing.]
Inference of BNs

- Suppose we already have the parameters of a Bayesian Network...

1. How do we compute the probability of a specific assignment to the variables?
   \[ P(T=t, H=h, A=a, C=c) \]

2. How do we draw a sample from the joint distribution?
   \[ t,h,a,c \sim P(T, H, A, C) \]

3. How do we compute marginal probabilities?
   \[ P(A) = ... \]

4. How do we draw samples from a conditional distribution?
   \[ t,h,a \sim P(T, H, A \mid C = c) \]

5. How do we compute conditional marginal probabilities?
   \[ P(H \mid C = c) = ... \]

[Slide from Matt Gormley.]
Approaches to inference

- **Exact inference algorithms**
  - The elimination algorithm → Message Passing
  - Belief propagation
  - The junction tree algorithms

- **Approximate inference techniques**
  - Variational algorithms
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods

[Slide from Eric Xing.]
Marginalization and Elimination

- A food web:

What is the probability that hawks are leaving given that the grass condition is poor?

Query: $P(h)$

$$P(h) = \sum_{g} \sum_{f} \sum_{e} \sum_{d} \sum_{c} \sum_{b} \sum_{a} P(a, b, c, d, e, f, g, h)$$

By chain decomposition, we get

$$= \sum_{g} \sum_{f} \sum_{e} \sum_{d} \sum_{c} \sum_{b} \sum_{a} P(a)P(b)P(c | b)P(d | a)P(e | c, d)P(f | a)P(g | e)P(h | e, f)$$

[Slide from Eric Xing.]
Marginalization and Elimination

- Query: \( P(A \mid h) \)
  - Need to eliminate: \( B,C,D,E,F,G,H \)

- Initial factors:
  \[
P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)P(h \mid e,f)
\]

- Choose an elimination order: \( H,G,F,E,D,C,B \)

- Step 1:
  - **Conditioning** (fix the evidence node (i.e., \( h \)) on its observed value (i.e., \( \tilde{h} \)):
    \[
m_h(e,f) = p(h = \tilde{h} \mid e,f)
    \]
  - This step is isomorphic to a marginalization step:
    \[
m_h(e,f) = \sum_h p(h \mid e,f)\delta(h = \tilde{h})
    \]
- **Query:** $P(A \mid h)$
  - Need to eliminate: $B,C,D,E,F,G$

- **Initial factors:**

  $$P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)P(h \mid e,f)$$
  $$\Rightarrow P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)m_h(e,f)$$

- **Step 2: Eliminate $G$**
  - compute

    $$m_g(e) = \sum_g p(g \mid e) = 1$$

    $$\Rightarrow P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)m_g(e)m_h(e,f)$$
    $$= P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)m_h(e,f)$$

[Slide from Eric Xing.]
• Query: $P(A \mid h)$
  • Need to eliminate: $B, C, D, E, F, G, H$

• Initial factors:

$$P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c, d)P(f \mid a)P(g \mid e)P(h \mid e, f)$$

• Choose an elimination order: $H, G, F, E, D, C, B$

○ Step 8: Wrap-up

$$p(a, \tilde{h}) = p(a)m_b(a), \quad p(\tilde{h}) = \sum_a p(a)m_b(a)$$

$$\Rightarrow P(a \mid \tilde{h}) = \frac{p(a)m_b(a)}{\sum_a p(a)m_b(a)}$$
Elimination algorithm

- Elimination on trees is equivalent to message passing on branches
- Message-passing is consistent in trees

Application: HMM

[Slide from Eric Xing.]
Gibbs Sampling

\[ p(x_1| x_2^{(t)}) \]

\[ p(x_2| x_1^{(t+1)}) \]

[Slide from Matt Gormley.]
Gibbs Sampling

\[ x_2 \]

\[ p(x) \]

\[ x_{(t+2)} \]

\[ x_{(t+3)} \]

\[ x_{(t+4)} \]

\[ x_{(t+1)} \]

\[ x_{(t)} \]

[Slide from Matt Gormley.]
Gibbs Sampling

**Question:**
How do we draw samples from a conditional distribution?

\[ y_1, y_2, \ldots, y_J \sim p(y_1, y_2, \ldots, y_J \mid x_1, x_2, \ldots, x_J) \]

**(Approximate) Solution:**

- Initialize \( y_1^{(0)}, y_2^{(0)}, \ldots, y_J^{(0)} \) to arbitrary values
- For \( t = 1, 2, \ldots \):
  - \( y_1^{(t+1)} \sim p(y_1 \mid y_2^{(t)}, \ldots, y_J^{(t)}, x_1, x_2, \ldots, x_J) \)
  - \( y_2^{(t+1)} \sim p(y_2 \mid y_1^{(t+1)}, y_3^{(t)}, \ldots, y_J^{(t)}, x_1, x_2, \ldots, x_J) \)
  - \( y_3^{(t+1)} \sim p(y_3 \mid y_1^{(t+1)}, y_2^{(t+1)}, y_4^{(t)}, \ldots, y_J^{(t)}, x_1, x_2, \ldots, x_J) \)
  - \( \ldots \)
  - \( y_J^{(t+1)} \sim p(y_J \mid y_1^{(t+1)}, y_2^{(t+1)}, \ldots, y_{J-1}^{(t+1)}, x_1, x_2, \ldots, x_J) \)

**Properties:**

- This will eventually yield samples from \( p(y_1, y_2, \ldots, y_J \mid x_1, x_2, \ldots, x_J) \)
- But it might take a long time -- just like other Markov Chain Monte Carlo methods

[Slide from Matt Gormley.]
Gibbs Sampling

- **Full conditionals** only need to condition on the Markov Blanket

- Must be “easy” to sample from conditionals

- Many conditionals are log-concave and are amenable to adaptive rejection sampling
Take home message

- Graphical models portrays the sparse dependencies of variables
- Two types of graphical models: Bayesian network and Markov random field
- Conditional independence, Markov blanket, and d-separation
- Learning fully observed and partially observed Bayesian networks
- Exact inference and approximate inference of Bayesian networks
References