Introduction to Probabilistic Graphical Models

Lecture 1 – Mar 28, 2011
CSE 515, Statistical Methods, Spring 2011
Instructor: Su-In Lee
University of Washington, Seattle

Readings: K&F 2.1, 2.2, 2.3, 3.1

Logistics

- Teaching Staff
  - Instructor: Su-In Lee (suinlee@uw.edu, PAC 536)
    - Office hours: Fri 9-10am or by appointment (PAC 536)
  - TA: Andrew Guillory (guillory@cs.washington.edu)
    - Office hours: Wed 1:30-2:20 pm or by appointment (PAC 216)

- Course website
  - cs.washington.edu/515
  - Discussion group: course website

- Textbook
  - (required) Daphne Koller and Nir Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press
  - Various research papers (copies available in class)
### Course requirement

- 4 homework assignments (60% of final grade)
  - Theory / implementation exercises
  - First one goes out next Monday!
  - 2 weeks to complete each
  - HW problems are long and hard
    - Please, please, please start early!
  - Late/collaboration policies are described on the website
- Final exam (35%)
  - Date will be announced later.
- Participation (5%)

### Probabilistic graphical models (PGMs)

- One of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats, ...) in the last two decades...
- Tool for representing complex systems and performing sophisticated reasoning tasks
- Why have a model?
  - Compact and modular representation of complex systems
  - Ability to efficiently execute complex reasoning tasks
  - Make predictions
  - Generalize from particular problem
Probabilistic graphical models (PGMs)

- Many classical probabilistic problems in statistics, information theory, pattern recognition, and statistical mechanics are special cases of the formalism
  - Graphical models provides a common framework
  - Advantage: specialized techniques developed in one field can be transferred between research communities

- PGMs are a marriage between graph theory and probability theory
  - Representation: graph
  - Reasoning: probability theory
  - Any simple example?

A simple example

- We want to know/model whether our neighbor will inform us of the alarm being set off

- The alarm can set off (A) if
  - There is a burglary (B)
  - There is an earthquake (E)

- Whether our neighbor calls (N) depends on whether the alarm is set off (A)

- “Variables” in this system
  - Whether alarm being set off (A); burglary (B); earthquake (E); our neighbor calls (N)
A simple example

- **Variables**: Earthquake (E), Burglary (B), Alarm (A), NeighborCalls (N)

<table>
<thead>
<tr>
<th>E</th>
<th>B</th>
<th>A</th>
<th>N</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.01</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.04</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.05</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.01</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.02</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.07</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.01</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.07</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.44</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.04</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.06</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.05</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.05</td>
</tr>
</tbody>
</table>

- **Probabilistic Inference**
  - **Task I**: Say that the alarm is set off (A=True), then how likely is it to get a call from our neighbor (N=True)?
  - **Task II**: Given that my neighbor calls (N=True), how likely it is that a burglary occurred (B=True)?

- **Representation**: graph
  - Intuitive data structure

- **Reasoning**
  - Probability theory

8 independent parameters
Example Bayesian network

- The “Alarm” network for monitoring intensive care patients
  - 37 variables
  - 509 parameters (full joint $2^{37} \approx 10^{10}$)

Representation: graphs

- Intuitive data structure for modeling highly-interacting sets of variables
  - Compact representation
  - Explicit model of modularity

- Data structure that allows for design of efficient general-purpose algorithms
Reasoning: probability theories

- Well understood framework for modeling uncertainty
  - Partial knowledge of the state of the world
  - Noisy observations
  - Phenomenon not covered by our model
  - Inherent stochasticity
- Clear semantics
- Can be learned from data

Probabilistic reasoning

- This course covers:
  - Probabilistic graphical model (PGM) representation
    - Bayesian networks (directed graph)
    - Markov networks (undirected graph)
  - Answering queries in PGMs ("inference")
    - What is the probability of X given some observations?
    - What is the most likely explanation for what is happening?
  - Learning PGMs from data ("learning")
    - What are the right/good parameters/structure of the model?
- Application & special topics
  - Modeling temporal processes with PGMs
    - Hidden Markov Models (HMMs) as a special case
  - Modeling decision-making processes
    - Markov Decision Processes (MDPs) as a special case
Course outline

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction, Bayesian network representation</td>
<td>2.1-3, 3.1</td>
</tr>
<tr>
<td></td>
<td>Bayesian network representation cont.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Local probability models</td>
<td>3.1-3</td>
</tr>
<tr>
<td></td>
<td>Undirected graphical models</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Exact inference</td>
<td>9.1-4</td>
</tr>
<tr>
<td></td>
<td>Exact inference cont.</td>
<td>10.1-2</td>
</tr>
<tr>
<td>4</td>
<td>Approximate inference</td>
<td>12.1-3</td>
</tr>
<tr>
<td></td>
<td>Approximate inference cont.</td>
<td>12.1-3</td>
</tr>
<tr>
<td>5</td>
<td>Parameter estimation</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Parameter estimation cont.</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>Partially observed data (EM algorithm)</td>
<td>19.1-3</td>
</tr>
<tr>
<td></td>
<td>Structure learning BNs</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>Structure learning BNs cont.</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Partially observed data</td>
<td>19.4-5</td>
</tr>
<tr>
<td>8</td>
<td>Learning undirected graphical models</td>
<td>20.1-3</td>
</tr>
<tr>
<td></td>
<td>Learning undirected graphical models cont.</td>
<td>20.1-3</td>
</tr>
<tr>
<td>9</td>
<td>Hidden Markov Models</td>
<td>TBD</td>
</tr>
<tr>
<td></td>
<td>HMMs cont. and Kalman filter</td>
<td>TBD</td>
</tr>
<tr>
<td>10</td>
<td>Markov decision processes</td>
<td>TBD</td>
</tr>
</tbody>
</table>

Application: recommendation systems

- Given user preferences, suggest recommendations
- **Example:** Amazon.com

- Input: movie preferences of many users
- Solution: model correlations between movie features
  - Users that like comedy, often like drama
  - Users that like action, often do not like cartoons
  - Users that like Robert Deniro films often like Al Pacino films
  - Given user preferences, can predict probability that new movies match preferences
Diagnostic systems

- Diagnostic indexing for home health site at Microsoft
- Enter symptoms → recommend multimedia content

Many research areas in CS

- Full of tasks that require reasoning under uncertainty
Enjoy!

- Probabilistic graphical models are having significant impact in science, engineering and beyond

- This class should give you the basic foundation for applying PGMs and developing new methods

- The fun begins ...

Today

- **Basics of probability**
  - Conditional probabilities
  - Statistical independence
  - Random variable

- Simple Bayesian networks
  - Two nodes make a BN
  - Naïve Bayes

- Should be a review for everyone – Setting up notation for the class
Sample spaces, events and probabilities

- **Probability**
  - A degree of confidence that an “event” of an uncertain nature will occur.

- Begin with a set $\Omega$ -- the sample space
  - Space of possible outcomes
  - e.g. if we consider dice, we might have a set $\Omega = \{1,2,3,4,5,6\}$
  - $\alpha \in \Omega$ is a sample point / atomic event.

- A **probability space** is a sample space with an assignment $P(\alpha)$ for every $\alpha \in \Omega$ s.t.
  - $0 \leq P(\alpha) \leq 1$
  - $\sum_{\alpha \in \Omega} P(\alpha) = 1$
  - e.g. $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$

- An event $A$ is any subset of $\Omega$
  - $P(A) = \sum_{\alpha \in A} P(\alpha)$
  - E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 0.5$

Conditional probabilities

- Consider two events $\alpha$ and $\beta$,
  - e.g. $\alpha$ = getting admitted to the UW CSE, $\beta$ = getting a job offer from Microsoft.
  - After learning that $\alpha$ is true, how do we feel about $\beta$?
  - $P(\beta | \alpha)$

\[
P(\beta | \alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}
\]
Two of the most important rules of the quarter: 1. The chain rule

- From the definition of the conditional distribution, we immediately see that
  \[ P(\alpha \cap \beta) = P(\alpha)P(\beta | \alpha) \]

- More generally:
  \[ P(\alpha_1 \cap \ldots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 | \alpha_1) \cdots P(\alpha_k | \alpha_1 \cap \ldots \cap \alpha_{k-1}) \]

Two of the most important rules of the quarter: 2. Bayes rule

- Another immediate consequence of the definition of conditional probability is:
  \[ P(\alpha | \beta) = \frac{P(\beta | \alpha)P(\alpha)}{P(\beta)} \]

- A more general version of Bayes’ rule, where all the probabilities are conditioned on some “background” event \( \gamma \):
  \[ P(\alpha | \beta \cap \gamma) = \frac{P(\beta | \alpha \cap \gamma)P(\alpha | \gamma)}{P(\beta | \gamma)} \]
Most important concept of the quarter: a) Independence

- $\alpha$ and $\beta$ are independent, if $P(\beta | \alpha) = P(\beta)$
- Denoted $P \rightarrow (\alpha \perp \beta)$

**Proposition:** $\alpha$ and $\beta$ are independent if and only if $P(\alpha \cap \beta) = P(\alpha)P(\beta)$

$$P(\alpha \cap \beta) = P(\alpha)P(\beta | \alpha) = P(\alpha)P(\beta)$$

Most important concept of the quarter: b) Conditional independence

- Independence is rarely true, but conditionally...

- $\alpha$ and $\beta$ conditionally independent given $\gamma$ if $P(\beta | \alpha \cap \gamma) = P(\beta | \gamma)$
- Denoted $P \rightarrow (\alpha \perp \beta | \gamma)$

**Proposition:** $P \rightarrow (\alpha \perp \beta | \gamma)$ if and only if $P(\alpha \cap \beta | \gamma) = P(\alpha | \gamma)P(\beta | \gamma)$
Random variables

- Probability distributions are defined for events.

- Events are complicated – so, let’s think about attributes:
  - Age, Grade, HairColor

- A random variable (such as Grade), is defined by a function that associates each outcome in \( \Omega \) (each person) with a value.

  \[
  \begin{align*}
  \text{Grade} = \text{A} & \quad \text{shorthand for event} \{ w \in \Omega : f_{\text{Grade}}(w) = \text{A} \} \\
  \text{Grade} = \text{B} & \quad \text{shorthand for event} \{ w \in \Omega : f_{\text{Grade}}(w) = \text{B} \} 
  \end{align*}
  \]

- Properties of a random variable \( X \):
  - \( \text{Val}(X) \) = a set of possible values of random variable \( X \)
  - For discrete (categorical): \( \sum_{x \in \text{Val}(X)} p(X=x) = 1 \)

Basic concepts for random variables

- Atomic event: assignment \( x_1, \ldots, x_n \) to \( X_1, \ldots, X_n \)

- Conditional probability:
  \[
  p(Y|X) = \frac{p(X,Y)}{p(X)}
  \]
  - For all values \( x \in \text{Val}(X), y \in \text{Val}(Y) \)

- Bayes rule:
  \[
  p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}
  \]

- Chain rule:
  \[
  p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|X_{i-1}, \ldots, X_1)
  \]
Joint distribution, marginalization

- Two random variables – Grade & Intelligence

- Marginalization – Compute marginal over single variable

Marginalization – the general case

- Compute marginal distribution $P(X_i)$ from joint distribution $P(X_1, \ldots, X_i, \ldots, X_n)$:

$$P(X_1, X_2, \ldots, X_i) = \sum_{x_{i+1}, \ldots, x_n} P(X_1, X_2, \ldots, X_i, x_{i+1}, \ldots, x_n)$$

$$P(X_i) = \sum_{x_1, \ldots, x_{i-1}} P(x_1, \ldots, x_{i-1}, X_i)$$
Today

- Basics of probability
  - Conditional probabilities
  - Statistical independence
  - Random variable
- **Two nodes make a BN**
- Naïve Bayes

- Should be a review for everyone – Setting up notation for the class

---

Representing joint distributions

- Random variables: $X_1, \ldots, X_n$
- $P$ is a joint distribution over $X_1, \ldots, X_n$

$P(X_1, \ldots, X_n)$

If $X_1, \ldots, X_n$ binary, need $2^n$ parameters to describe $P$

Can we represent $P$ more compactly?
- Key: Exploit independence properties
Independent random variables

- If $X_1, \ldots, X_n$ are independent then:
  - $P(X_1, \ldots, X_n) = P(X_1) \ldots P(X_n)$
  - $O(n)$ parameters
  - All $2^n$ probabilities are implicitly defined
  - Cannot represent many types of distributions

- $X$ and $Y$ are conditionally independent given $Z$ if
  - $P(X=x|Y=y, Z=z) = P(X=x|Z=z)$ for all values $x, y, z$
  - Equivalently, if we know $Z$, then knowing $Y$ does not change predictions of $X$
  - Notation: $(X \perp Y | Z)$

---

Conditional parameterization

- $S = \text{SAT score, } \text{Val}(S) = \{s^0, s^1\}$
- $I = \text{Intelligence, } \text{Val}(I) = \{i^0, i^1\}$

\[
\begin{array}{c|cc}
  I & S & P(I, S) \\
  \hline
  i^0 & s^0 & 0.665 \\
  i^0 & s^1 & 0.035 \\
  i^1 & s^0 & 0.06 \\
  i^1 & s^1 & 0.24 \\
\end{array}
\]

Joint parameterization

\[
\begin{array}{c|cc}
  I & S & P(I,S) \\
  \hline
  i^0 & s^0 & 0.7 \\
  i^0 & s^1 & 0.3 \\
  i^1 & s^0 & 0.2 \\
  i^1 & s^1 & 0.8 \\
\end{array}
\]

Conditional parameterization

Alternative parameterization: $P(S)$ and $P(I|S)$
Conditional parameterization

- S = SAT score, Val(S) = \{s^0,s^1\}
- I = Intelligence, Val(I) = \{i^0,i^1\}
- G = Grade, Val(G) = \{g^0,g^1,g^2\}

Assume that G and S are independent given I

Joint parameterization

- 2 \cdot 2 \cdot 3 = 12 - 1 = 11 independent parameters

Conditional parameterization has

- P(I) – 1 independent parameter
- P(S|I) – 2\cdot1 independent parameters
- P(G|I) - 2\cdot2 independent parameters
- 7 independent parameters

Naïve Bayes model

- Class variable C, Val(C) = \{c_1,\ldots,c_k\}
- Evidence variables X_1,\ldots,X_n
- Naïve Bayes assumption: evidence variables are conditionally independent given C

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

Applications in medical diagnosis, text classification

Used as a classifier:

\[
\frac{P(C = c_1 \mid X_1,\ldots, X_n)}{P(C = c_2 \mid X_1,\ldots, X_n)} = \frac{P(C = c_1)}{P(C = c_2)} \prod_{i=1}^{n} \frac{P(x_i \mid C = c_1)}{P(x_i \mid C = c_2)}
\]

Problem: Double counting correlated evidence
Bayesian network (informal)

- Directed acyclic graph $G$
  - Nodes represent random variables
  - Edges represent direct influences between random variables
- Local probability models

![Diagrams]

Example 1

Example 2

Naïve Bayes

Bayesian network (informal)

- Represent a joint distribution
  - Specifies the probability for $P(X=x)$
  - Specifies the probability for $P(X=x|E=e)$
- Allows for reasoning patterns
  - Prediction (e.g., intelligent $\rightarrow$ high scores)
  - Explanation (e.g., low score $\rightarrow$ not intelligent)
  - Explaining away (different causes for same effect interact)

![Diagrams]

Example 2
Bayesian network structure

- Directed acyclic graph $G$
  - Nodes $X_1, \ldots, X_n$ represent random variables
- $G$ encodes local Markov assumptions
  - $X_i$ is independent of its non-descendants given its parents
  - Formally: $(X_i \perp \text{NonDesc}(X_i) \mid \text{Pa}(X_i))$

Independency mappings (I-maps)

- Let $P$ be a distribution over $X$
- Let $I(P)$ be the independencies $(X \perp Y \mid Z)$ in $P$
- A Bayesian network structure is an I-map (independency mapping) of $P$ if $I(G) \subseteq I(P)$

\[
\begin{array}{ccc}
| I & S & P(I,S) | \\
| p & s^0 & 0.25 | \\
| p & s^1 & 0.25 | \\
| i & s^0 & 0.25 | \\
| i & s^1 & 0.25 | \\
\end{array}
\]

\[
\begin{array}{ccc}
| I & S & P(I,S) | \\
| p & s^0 & 0.4 | \\
| p & s^1 & 0.3 | \\
| i & s^0 & 0.2 | \\
| i & s^1 & 0.1 | \\
\end{array}
\]
Factorization Theorem

- If G is an I-Map of P, then P factorizes over G.
  \[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \]

Proof:
- wlog. (without loss of generality) \( X_1, \ldots, X_n \) is an ordering consistent with G
- By chain rule: \( P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) \)
- From assumption: \( Pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\} \)
- From chain rule: \( \{X_1, \ldots, X_{i-1}\} \cap Pa(X_i) \subseteq NonDesc(X_i) \)
- Since G is an I-Map \( \rightarrow (X_i; NonDesc(X_i)| Pa(X_i)) \in I(P) \)

\[ P(X_i | X_1, \ldots, X_{i-1}) = P(X_i | Pa(X_i)) \]

Factorization implies I-Map

- \( P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \rightarrow G \) is an I-Map of P

Proof:
- Need to show \((X_i; NonDesc(X_i)| Pa(X_i)) \in I(P) \) or that \( P(X_i | NonDesc(X_i)) = P(X_i | Pa(X_i)) \)
- wlog. \( X_1, \ldots, X_n \) is an ordering consistent with G

\[ P(X_i | NonDesc(X_i)) = \frac{P(X_i, NonDesc(X_i))}{P(NonDesc(X_i))} = \prod_{k=1}^{i-1} P(X_k | Pa(X_k)) \]

\[ = \frac{\prod_{k=1}^{i-1} P(X_k | Pa(X_k))}{\prod_{k=1}^{i-1} P(X_k | Pa(X_k))} = P(X_i | Pa(X_i)) \]
Bayesian network definition

- A Bayesian network is a pair (G,P)
  - P factorizes over G
  - P is specified as set of CPDs associated with G’s nodes (and its parents)

- Parameters
  - Joint distribution: $2^n$
  - Bayesian network (bounded in-degree k): $n^{2k}$

Today and next class

- Next class
  - Details on semantics of BNs, relate them to independence assumptions encoded by the graph.

- Today’s To-Do List
  - Visit the course website.
  - Reading K&F 2.1-3, 3.1.
Acknowledgement

- These lecture notes were generated based on the slides from Profs Eran Segal and Carlos Guestrin.