Outline of Today’s Lecture

• Class overview
• What are graphical models
• Semantics of Bayesian networks
Books and Sources for Today

- Jordan: Chapters 1 and 2
## Class Road Map

<table>
<thead>
<tr>
<th>L1: Tues, 3/28</th>
<th>Overview, GMs, Intro BNs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2: Thur, 3/30</td>
<td></td>
</tr>
<tr>
<td>L3: Tues, 4/4</td>
<td></td>
</tr>
<tr>
<td>L4: Thur, 4/6</td>
<td></td>
</tr>
<tr>
<td>L5: Tue, 4/11</td>
<td></td>
</tr>
<tr>
<td>L6: Thur, 4/13</td>
<td></td>
</tr>
<tr>
<td>L7: Tues, 4/18</td>
<td></td>
</tr>
<tr>
<td>L8: Thur, 4/20</td>
<td></td>
</tr>
<tr>
<td>L9: Tue, 4/25</td>
<td></td>
</tr>
<tr>
<td>L10: Thur, 4/27</td>
<td></td>
</tr>
</tbody>
</table>

| L11: Tues, 5/2 |                           |
| L12: Thur, 5/4 |                           |
| L13: Tues, 5/9 |                           |
| L14: Thur, 5/11|                           |
| L15: Tue, 5/16 |                           |
| L16: Thur, 5/18|                           |
| L17: Tues, 5/23|                           |
| L18: Thur, 5/25|                           |
| L19: Tue, 5/30 |                           |
| L20: Thur, 6/1: final presentations |
Announcements

- READING: Chapter 1, 2 in Jordan’s book (pick up book from basement of communications copy center).
- List handout: name, department, and email
- List handout: regular makeup slot, and discussion section
- Syllabus
- Course web page: http://ssli.ee.washington.edu/ee512
- Goal: powerpoint slides this quarter, they will be on the web page after lecture (but not before as you are getting them hot off the press).
Graphical Models

• A graphical model is a visual, abstract, and mathematically formal description of properties of families of probability distributions (densities, mass functions)
• There are many different types of Graphical model, ex:

  – Bayesian Networks
  – Markov Random Fields
  – Factor Graph
  – Chain Graph
GMs cover many well-known methods

Graphical Models
- Causal Models
- Chain Graphs
- Factor Graphs
- DGMs
- Bayesian Networks
- UGMs
- Dependency Networks
- DBNs
- MRFs
- Simple Models
- AR
- PCA
- Gibbs/Boltzman Distributions
- Other Semantics
- FST
- ZMs
- Mixture Models
- Kalman
- Decision Trees
- LDA
- Segment Models
- AR
- PCA
- Factorial HMM/Mixed Memory Markov Models
- BMMs
Graphical Models Provide:

- **Structure**
- **Algorithms**
- **Language**
- **Approximations**
- **Data-Bases**
Graphical Models Provide

GMs give us:

I. **Structure**: A method to explore the structure of “natural” phenomena (causal vs. correlated relations, properties of natural signals and scenes)

II. **Algorithms**: A set of algorithms that provide “efficient” probabilistic inference and statistical decision making

III. **Language**: A mathematically formal, abstract, visual language with which to efficiently discuss families of probabilistic models and their properties.
GMs give us (cont):

**IV. Approximation**: Methods to explore systems of approximation and their implications. E.g., what are the consequences of a (perhaps known to be) wrong assumption?

**V. Data-base**: Provide a probabilistic “data-base” and corresponding “search algorithms” for making queries about properties in such model families.
GMs

- There are many different types of GM.
- Each GM has its semantics
- A GM (under the current semantics) is really a set of constraints. The GM represents all probability distributions that obey these constraints, including those that obey additional constraints (but not including those that obey fewer constraints).
- Most often, the constraints are some form of factorization property, e.g., $f()$ factorizes (is composed of a product of factors of subsets of arguments).

\[ f(a, b, c) = g(a, b)h(b, c) \]
Types of Queries

• Several types of queries we may be interested in:
  – Compute: $p(\text{one subset of vars})$
  – Compute: $p(\text{one subset of vars}| \text{another subset of vars})$
  – Find the $N$ most probable configurations of one subset of variables given assignments of values to some other sets
  – Q: Is one subset independent of another subset?
  – Q: Is one subset independent of another given a third?

• How efficiently can we do this? Can this question be answered? What if it is too costly, can we approximate, and if so, how well? These are questions we will answer this term.

• GMs are like a probabilistic data-base (or data structure), a system that can be queried to provide answers to these sorts of questions.
Example

• Typical goal of pattern recognition:
  – training (say, EM or gradient descent), need query of form:

\[
\theta_n = \arg\max_{\theta} \mathcal{E}_{p(o,h|\theta_p)}[\log p(o, h|\theta)]
\]

In this form, we need to compute \(p(o,h)\) efficiently.

– Bayes decision rule, need to find best class for a given unknown pattern:

\[
c^* = \arg\max_{c} p(c|x) = \arg\max_{c} p(c, x)
\]

– but this is yet another query on a probability distribution.

– We can train, and perform Bayes decision theory quickly if we can compute with probabilities quickly. Graphical models provide a way to reason about, and understand when this is possible, and if not, how to reasonably approximate.
Some Notation

- Random variables $\psi, \psi, \psi, \psi$ (scalar or vector)
- Distributions:

$$p(x_1:n) \equiv p(x_1, \ldots, x_n) \equiv P_{X_1,\ldots,X_n}(X_1 = x_1, \ldots, X_n = x_n)$$

- Subsets:

$$V \triangleq \{1, 2, \ldots, N\} \ A, B \subseteq V$$

$$X_A \triangleq \{X_{A_1}, X_{A_2}, \ldots, X_{A_{|A|}}\}$$

I.e., if $A = \{1, 3, 7\}$ then $X_A = \{X_1, X_3, X_7\}$

$$p(X_A = x_A | X_B = x_B) \equiv p(x_1, x_2 | x_3, x_4)$$

if $A = \{1, 2\}, B = \{3, 4\}$

$p(x_A)$ requires table of size $r^{|A|}$, $r = |X|$ where $\forall i, x_i \in X$
Main types of Graphical Models

• Markov Random Fields
  – a form of undirected graphical model
  – relatively simple to understand their semantics
  – also, log-linear models, Gibbs distributions, Boltzman distributions, many “exponential models”, conditional random fields (CRFs), etc.

• Bayesian networks
  – a form of directed graphical model
  – originally developed to represent a form of causality, but not ideal for that (they still represent factorization)
  – Semantics more interesting (but trickier) than MRFs

• Factor Graphs
  – pure, the assembly language models for factorization properties
  – came out of coding theory community (LDPC, Turbo codes)
Main types of Graphical Models

• Chain graphs:
  – Hybrid between Bayesian networks and MRFs
  – A set of clusters of undirected nodes connected as directed links
  – Not as widely used, but very powerful.

• Ancestral graphs
  – we probably won’t cover these.
Bayesian Network Examples

Mixture models

\[ p(x) = \sum_i c_i p(x \mid i) \]

Markov Chains

\[ p(q_t \mid q_{1:t-1}) = p(q_t \mid q_{t-1}) \]

\[ p(q_t \mid q_{1:t-1}) = p(q_t \mid q_{t-1}, q_{t-2}) \]
GMs: PCA and Factor Analysis

\[ Y = AX + N(\mu, R) \]
\[ X \sim N(0, Q) \]

PCA: \( Q = \Lambda, \quad R \rightarrow 0, \quad A = \text{ortho} \)
\[ Y = \Gamma X + \mu \]

FA: \( Q = I, \quad R = \text{diagonal} \)
\[ Y = \Gamma X + u + \mu \]

Other generalizations possible
E.g., \( Q = \text{gen. diagonal}, \) or capture using general \( A \) since
\[ Y \sim N(\mu, AQA^T + R) \]
The data $X_{1:4}$ is explained by the two (marginally) independent causes.

$I_1 \perp I_2$
Linear Discriminant Analysis

\[
P(C = i \mid X) = \frac{f_i(X)p(i)}{\sum_k f_k(X)p(k)}
\]

\[
f_j(X) = N(\mu_j, \Sigma)
\]

- Class conditional data has different means but common covariance matrix.

- Fisher’s formulation: project onto space spanned by the means.
Extensions to LDA

HDA/QDA

Heteroscedastic Discriminant Analysis/Quadratic Discriminant Analysis

MDA

Mixture Discriminant Analysis

HMDA

Both
Generalized Decision Trees

- Generalized Probabilistic Decision Trees
- Hierarchical Mixtures of Experts (Jordan)
Example: Printer Troubleshooting
Classifier Combination

Mixture of Experts
(Sum rule)

Naive Bayes
(prod. rule)

Other Combination Schemes
Discriminative and Generative Models

Generative Model

\[ C \rightarrow X \]

Discriminative Model

\[ C \rightarrow X \]
HMMs/Kalman Filter

HMM

Autoregressive HMM
Switching Kalman Filter
Factorial HMM
Standard Language Modeling

• Example: standard 4-gram

\[ P(w_t \mid h_t) = P(w_t \mid w_{t-1}, w_{t-2}, w_{t-3}) \]
Interpolated Uni-, Bi-, Tri-Grams

\[
P(w_t | h_t) = P(\alpha_t = 1)P(w_t) + P(\alpha_t = 2)P(w_t | w_{t-1}) + P(\alpha_t = 3)p(w_t | w_{t-1}, w_{t-2})
\]

- Nothing gets zero probability
Conditional mixture tri-grams

\[ P(w_t \mid h_t) = P(\alpha_t = 1 \mid w_{t-1}, w_{t-2})P(w_t) \]
\[ + P(\alpha_t = 2 \mid w_{t-1}, w_{t-2})P(w_t \mid w_{t-1}) \]
\[ + P(\alpha_t = 3 \mid w_{t-1}, w_{t-2})P(w_t \mid w_{t-1}, w_{t-2}) \]
Bayesian Networks

• … and so on
• We need to be more formal about what BNs mean.
• In the rest of this, and in the next, lecture, we start with basic semantics of BNs, move on to undirected models, and then come back to BNs again to clean up …
Bayesian Networks

• Has nothing to do with “Bayesian statistical models” (there are Bayesian and non-Bayesian Bayesian networks).

• Graph $G = (V, E)$, $V = \{X_1, X_2, \ldots, X_N\}$ (sometimes we write $V = \{v_1, \ldots, v_N\}$).

• $E = \{(X_i, X_j) : i \not= j\}$ set of directed edges. Sometimes we write $(i, j) \in E$ if $(X_i, X_j) \in E$. Also, $(i, j) \in E$ means $X_i \rightarrow X_j$.

• $G$ is acyclic (a DAG), so there are no directed cycles.

• each node $X_i$ has a set of parents $X_{\pi_i}$ (arrows pointing to $X_i$)
Sub-family specification:
Directed acyclic graph (DAG)
- Nodes - random variables
- Edges - direct “influence”

Together (graph and inst.):
Defines a unique distribution in a factored form

\[ P(B, E, A, C, R) = P(B)P(E)P(A | B, E)P(R | E)P(C | A) \]

Instantiation:
Set of conditional probability distributions

| E | B | P(A | E,B) |
|---|---|-----------|
| e | b | 0.9 0.1   |
| e | b | 0.2 0.8   |
| e | b | 0.9 0.1   |
| e | b | 0.01 0.99 |
Bayesian Networks

• Reminder: chain rule of probability. For any order \( \sigma \)

\[
p(x_{1:N}) = \prod_{i} p(x_{\sigma_i} | x_{\sigma_1:i-1})
\]

• Locality: parent-child relationship, each child/parent we have function \( f_i(x_i, x_{\pi_i}) \) s.t. \( 0 \leq f_i(x_i, x_{\pi_i}) \leq 1 \)

\[
\sum_{x_i} f_i(x_i, x_{\pi_i}) = 1
\]

• (DF) Directed factorization property of BNs: For any given BN \( G \), we say that \( p \) DFs according to \( G \) if there exists such \( f_i \) functions such that:

\[
p(x_{1:n}) = \prod_{i} f_i(x_i, x_{\pi_i})
\]
Bayesian Networks

- We will see that \( p(x_{1:n}) = \prod_i p(x_i | x_{\pi_i}) \). For example, if \( x_1 \leftarrow x_2 \), \( p(x_1, x_2) = f(x_1, x_2)f(x_2) \), and \( p(x_2) = \sum_{x_1} p(x_1, x_2) = f(x_2) \), and so on.

- A given BN represents (is a surrogate for) all \( p() \) that factorize in this way.

\[
p(x_{1:6}) = p(x_1)p(x_2|x_1)p(x_3|x_1) \]
\[
p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)
\]

- Advantages of factored rep? \( p(x_1) \) 1-D table. \( p(x_6|x_2, x_5) \) 3-D table. \( p(x_{1:6}) \) 6-D table. Table size exponential in number of variables, number of tables linear in number of nodes (also has huge computational savings, as we’ll see).
Conditional Independence

- $X_A$ is **independent** of $X_B$, written $X_A \perp X_B$, true if:
  \[ p(x_A, x_B) = p(x_A)p(x_B) \]
  for all $x_A, x_B$. Alternatively, if there exists functions $f(), g()$ such that $p(x_A, x_B) = f(x_A)g(x_B)$.

- $X_A$ is **conditionally independent** of $X_B$ given $X_C$, written $X_A \perp X_B | X_C$, true if:
  \[ p(x_A, x_B | x_C) = p(x_A | x_C)p(x_B | x_C) \]
  for all $x_A, x_B, x_C$. Alternatively, if there exists functions $f(), g()$ such that $p(x_A, x_B, x_C) = f(x_A, x_C)g(x_B, x_C)$. Alternatively, if $p(x_A, x_B, x_C) = p(x_A, x_C)p(x_B, x_C)/p(x_C)$.

- There are many properties and consequences of conditional independence, we will study in this class.
Conditional Independence & BNs

• GMs can help answer CI queries:

\[ p(X, Y, Z, A, B, C, D, E) \]
\[ = p(Y | D)p(D | C, E)p(E | Z)p(C | B)p(B | A)p(X | A)p(A)p(Z) \]

\[ X \perp \!\!\!\perp Z | Y \quad ??? \]

\[ \text{No!!} \]
Bayesian Networks & CI

- \( X_A \perp X_B \mid X_C \), true if \( p(x_A \mid x_B, x_C) = p(x_A \mid x_C) \)
- Consider full chain-rule factorization:
  \[
p(x_{1:6}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_{1:2})p(x_4 \mid x_{1:3})p(x_5 \mid x_{1:4})p(x_6 \mid x_{1:5})
  \]
- Compare with factorization from graph from before:
  \[
p(x_{1:6}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(x_6 \mid x_2, x_5)
  \]
- Missing vars in local conditional probabilities correspond to missing edges in the graph!
- Definition: a topological ordering \( I \) of nodes, is such that if \((x_i, x_j) \in E\), and \( x_j \) is after \( x_i \) in \( I \), then \( i < j \).
- Example: \( I = (1, 2, 3, 4, 5, 6) \) is topological w.r.t. the graph above. So is \( I = (1, 3, 5, 2, 6, 4) \). But \((1, 3, 5, 6, 2, 4)\) is not topological.
Bayesian Networks & CI

• \( v_i \triangleq \{ j \in I : j < i, j \not\in \pi_i \} \), set of indices before \( i \) in \( I \) other than parents \( \pi_i \).
• Ex: for \( I = (1, 2, 3, 4, 5, 6) \), \( v_5 = (1, 2, 4) \), for \( I = (1, 3, 5, 2, 6, 4) \), \( v_5 = (1) \).
• Theorem: if (DF) is true, given any topological ordering \( I \), then: \( X_i \perp X_{v_i} | X_{\pi_i} \)
• This is property (DL), namely \( X_i \perp \text{non-descendants}(X_i) | \pi_i \).

Note, not indep of "downstream" vars given parents.

• Ex: For \( I = (1, 2, 3, 4, 5, 6) \), we have:
  \[
  \begin{align*}
  X_1 & \perp \emptyset | \emptyset \\
  X_2 & \perp \emptyset | X_1 \\
  X_3 & \perp X_2 | X_1 \\
  X_4 & \perp \{X_1, X_3\} | X_2 \\
  X_5 & \perp \{X_1, X_2, X_4\} | X_3 \\
  X_6 & \perp \{X_1, X_3, X_4\} | \{X_2, X_5\}
  \end{align*}
  \]

• Ex: For \( I = (1, 3, 5, 2, 6, 4) \), we have:
  \[
  \begin{align*}
  X_1 & \perp \emptyset | \emptyset \\
  X_3 & \perp \emptyset | X_1 \\
  X_5 & \perp X_1 | X_3 \\
  X_2 & \perp \{X_3, X_5\} | X_1 \\
  \ldots
  \end{align*}
  \]
Bayesian Networks & CI

- Do these missing-edges (and other purely graph theoretic ways to get) CI always correspond to the math? Yes. Example:
- Ex: is $X_4 \perp \{X_1, X_3\} \mid X_2$?

$$p(x_{1:4}) = \sum_{x_{5,6}} p(x_{1:6}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2) \sum_{x_5} p(x_5 \mid x_3) \sum_{x_6} p(x_6 \mid x_2, x_5)$$

$$= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)$$

$$p(x_{1:3}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)$$

$$p(x_4 \mid x_{1:3}) = \frac{p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)}{p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)} = p(x_4 \mid x_2)$$

- There are other CI statements implied by the graph, to get them need all orderings $I$. A better way is to use graph separation properties (i.e., $A$ is separated from $B$ by $C$ in the graph if all paths from any node in $A$ to any node in $B$ must go through (or is blocked by) some node in $C$).
Bayesian Networks & CI

- BN semantics is not as easy, need the notion of *directional separation*, or d-separation (which really means, we must pay attention to the arrows).

- When is $X_A \perp X_B | X_C$? Only when $X_C$ d-separates $X_A$ from $X_B$.

- $C$ separates $A$ from $B$ if all paths from $A$ to $B$ are *blocked*.

- A path is *blocked* if $\exists$ a node $v$ on the path such that either:
  1) either $\rightarrow v \rightarrow$, $\leftarrow v \leftarrow$, or $\leftarrow v \rightarrow$, and $v \in C$, or
  2) $\rightarrow v \leftarrow$ and neither $v$ nor any of $v$’s descendants are in $C$. 

Pictorial d-separation: blocked/unblocked paths
• Three 3-node examples of BNs and their conditional independence statements.
Case 1

- Markov Chain

- Why does this work? We can use factorization: 
  \[ p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2). \]
  Then 
  \[ p(x_3|x_1, x_2) = p(x_1, x_2, x_3)/p(x_1, x_2) = p(x_3|x_2). \]
  Also, use topological order: \( I = (1, 2, 3). \)
  \( v = \{\emptyset, \emptyset, \{X_1\}\}. \)

- No other CIs in general hold. I.e., doesn’t require that not
  \[ p(y|x) = p(y), \]
  only that above CIs hold!

- Present edges do not necessarily imply a dependence. Rather, missing edges do necessarily imply some (conditional) independence.
Case 2

- Still a Markov Chain

\[ X_1 \perp X_3 \mid X_2 \]

- Factorization: \( p(x_1, x_2, x_3) = p(x_2)p(x_1|x_2)p(x_3|x_2). \) Then \( p(x_3|x_1, x_2) = p(x_1, x_2, x_3)/p(x_1, x_2) = p(x_3|x_2). \) Or, topological order: \( I = (1, 2, 3). \) \( v = \{\emptyset, \emptyset, \{X_1\}\}. \)
- In this case, reversing the arrow has no effect, but that is not always true with BNs.
Case 3

- NOT a Markov Chain

\[ X_1 \parallel X_3 \]

- Factorization: \( p(x_1, x_2, x_3) = p(x_1)p(x_3)p(x_2|x_1, x_3) \). Then \( p(x_1, x_2) = p(x_1)p(x_2) \). Or, topological order: \( I = (1, 2, 3) \).

- In this case, reversing the arrow has a big effect, these are called V-structures in BNs. This gives BNs quite a bit of power.
Examples of the three cases

- SUVs → Greenhouse Gasses → Global Warming
- Lung Cancer ← Smoking → Bad Breath
- Genetics → Cancer ← Smoking
Bayes Ball

- simple algorithm, ball bouncing along path in graph, to tell if path is blocked or not (more details in text).

![Diagrams of Bayes Ball](image)
What are implied conditional independences?
Two Views of a Family

- Given DAG $G = (V, E)$, where $\pi_i$ are the parents of $i$,

  $$F_1 \triangleq \left\{ p(x_{1:n}) : p(x_{1:n}) = \prod_i p(x_i | x_{\pi_i}) \right\}$$

- Given DAG $G = (V, E)$, where $\pi_i$ are the parents of $i$, let $I_j$ be the $j^{th}$ conditional independence statement on a graph, i.e., $I_j = \{ X_{A_j} \perp X_{B_j} | X_{C_j} \}$. Define:

  $$F_2 \triangleq \left\{ p(x_{1:n}) : p(x_{1:n}) \text{ satisfies } I_j \forall j \right\}$$

- **Theorem:** $F_1 = F_2$. In other words, we get same family either by the constraints expressed by the factorization properties of a graph, or by the conditional independence properties. (Lauritzen, Thm 3.37).