Quiz 1: due Fri 11:59 pm

COMPSCI 688: Probabilistic Graphical Models

Lecture 3: Directed Graphical Models

Dan Sheldon

Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Review

Review

$X \perp Y \iff p(x,y) = p(x)p(y) \quad \forall x, y$

- Conditional independence

$$\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} \iff p(\mathbf{y}, \mathbf{x}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$$
$$\iff p(\mathbf{x}|\mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})$$

Bayesian Networks

Compactness from Independence

Suppose we have a joint distribution $p(a, b, c)$ and we know that the independence relation $C \perp A | B$ holds. How can we exploit this fact to simplify $p(a, b, c)$?

chain rule $p(a,b,c) = p(a)\,p(b|a)\,p(c|a,b)$

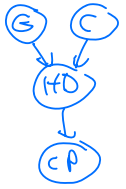
simplified using CI $p(a,b,c) = p(a)\,p(b|a)\,p(c|b)$

Bayesian Networks: Main Idea

- The main idea of Bayesian networks is conceptually simple:

1. Order the variables and apply the chain rule
2. Drop some dependencies, which corresponds to conditional independence assumptions

- **Example**: variables G, C, HD, CP, assume: (1) $G \perp C$, (2) $CP \perp (G, C) | HD$



1. $p(g,c,hd,cp) = p(g)\,p(c|g)\,p(hd|g,c)\,p(cp|g,c,hd)$

$\Downarrow$

2. $p(g,c,hd,cp) = p(g)\,p(c)\,p(hd|g,c)\,p(cp|hd)$

Bayesian Networks: Main Idea

- This idea has several consequences:
  - The variables can be arranged in a directed acyclic graph (DAG). (Sometimes interpreted causally, but beware.)
  - The distribution satisfies certain (local and global) conditional independence properties that can be derived from the graph
- We'll next introduce Bayesian networks formally and start discussing their properties

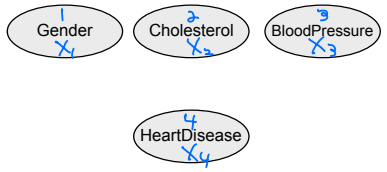
Bayesian Networks: Nodes

Formally, a Bayesian network consists of a directed acyclic graph (DAG) $\mathcal{G}$ and a joint distribution $p(\mathbf{x}) = p(x_1, \ldots, x_N)$ for random variables $X_1, \ldots, X_N$

The vertex set V has one node i for each random variable X_i

Warning: it's also common to use the random variable itself, i.e., X_i as the node

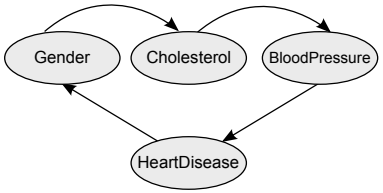
Example:



Bayesian Networks: Edges

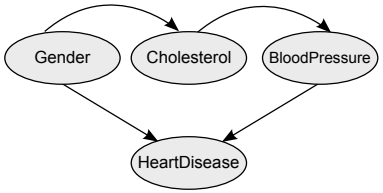
The DAG constraint means that $\mathcal{G}$ can't contain any directed cycles $i \to j \to \cdots \to i$.

Example:



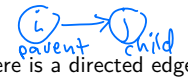
Not a valid DAG
Directed Cycle

Example:



A valid DAG.
No directed cycle

Bayesian Networks: Parents/Children



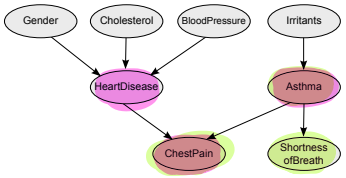
If there is a directed edge $i \to j$:

- i is a *parent* of j
- j is a *child* of i
- (sometimes: X_i is a parent of X_j, and so on)

Define

- $\text{pa}(i)$ = set of all parents of i
- $\text{ch}(i)$ = set of all children of i

Example:



$\text{pa}(CP) = \{HD, A\}$

$\text{ch}(A) = \{CP, SB\}$

Bayesian Networks: Descendants/Non-Descendants

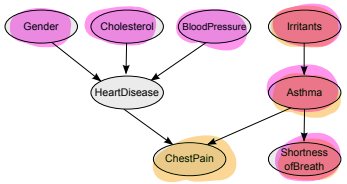
If there is a directed path from i to j:

- j is a *descendant* of i.
- Else j is a *non-descendent* of i.

Define

- $\text{de}(i)$ = set of all descendants of i
- $\text{nd}(i)$ = set of all non-descendants of i

Example:



$\text{de}(I) = \{A, SB, CP\}$

$\text{nd}(BP) = \{G, C, I, A, SB\}$

Bayesian Networks: Joint Distribution

The joint distribution implied by a Bayesian network is **factorized** into a product of local conditional probability distributions.

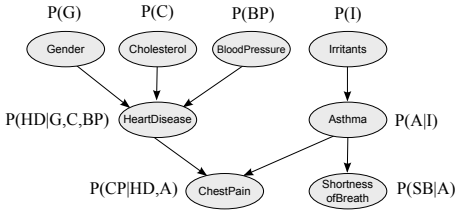


Figure 1: image

$p(G, C, \ldots, SB) = p(G)p(C)p(BP)p(I)p(HD|G,C,BP)p(A|I)p(CP|HD,A)p(SB|A)$

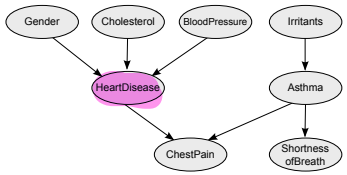
The joint distribution is the product of the conditional distributions:
$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{x}_{\text{pa}(i)})$.

Bayesian Networks: CPDs and CPTs

- The individual factors $p(x_i \mid \mathbf{x}_{\text{pa}(i)})$ in a Bayesian network are referred to as conditional probability distributions or CPDs.
- The CPD for node i must specify the probability that X_i takes any value x_i in its domain when conditioned on each joint assignment $\mathbf{x}_{\text{pa}(i)}$ of its parents
- For discrete random variables, we can represent the CPD of each node using a look-up table called a conditional probability table or CPT.

Bayesian Networks: CPT Example

hd	g	bp	ch	$p(hd \mid g, bp, ch)$
No	M	Low	Low	0.95
Yes	M	Low	Low	0.05
No	F	Low	Low	0.99
Yes	F	Low	Low	0.01
⋮				



Bayesian Networks: Storage Complexity

- What is the minimum amount of space needed to store the probability distribution for a single discrete random variable that takes V values?
- How much space does it take to store the CPT for a binary-valued variable with D binary-valued parents?
- Suppose there are D binary variables connected in a chain $X_1 \to X_2 \to \ldots \to X_D$.
  What is the total storage cost?
  How large is the full joint?

Conditional Independence and Factorization

Conditional Independence and Factorization

We assumed factorization in a Bayes net: $p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{x}_{\mathsf{pa}(i)})$. What does this have to do with conditional independence?

Claim: for a probability distribution $p(\mathbf{x})$

Fix DAG G

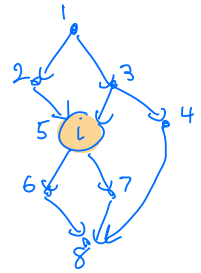
$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{x}_{\mathsf{pa}(i)}) \quad \iff \quad X_i \perp \mathbf{X}_{\mathsf{nd}(i)} \mid \mathbf{X}_{\mathsf{pa}(i)} \text{ for all } i$$

factorization $\iff$ conditional independence

- RHS in words: X_i **is conditionally independent of its non-descendants given its parents**

Conditional Independence Implies Factorization ($\Leftarrow$)

Assume $X_i \perp \mathbf{X}_{\mathsf{nd}(i)} \mid \mathbf{X}_{\mathsf{pa}(i)}$ for all i



1. Number nodes according to topological ordering
2. Use chain rule
$p(x) = \prod_{i=1}^{N} p(x_i \mid x_{1,\dots,i-1})$
non-descendants of i
3. Split $\{1,\dots,i-1\}$ into parents + non-descendants
$p(x) = \prod_{i=1}^{N} p(x_i \mid x_{pa(i)}, x_{\{1,\dots,i-1\} \setminus pa(i)})$
non-descendants

4. Use CI to simplify
$p(x) = \prod_{i=1}^{N} p(x_i \mid x_{pa(i)})$

$A \perp B, C \mid D \rightarrow A \perp B \mid D$

$A \perp B \mid B$ true but uninteresting

Review of Argument

0. Assume $X_i \perp \mathbf{X}_{\mathsf{nd}(i)} \mid \mathbf{X}_{\mathsf{pa}(i)}$ for all i
1. Number nodes according to a topological ordering and apply the chain rule
$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i | \mathbf{x}_{\{1,\dots,i-1\}})$$
2. Nodes $\{1,\dots,i-1\}$ must be non-descendants of i because they come earlier in the topological order. Therefore we can split these nodes into parents and other nodes which are all non-descendants:
$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i | \mathbf{x}_{\mathsf{pa}(i)}, \mathbf{x}_{\{1,\dots,i-1\} \setminus \mathsf{pa}(i)})$$
3. Now simplify using $X_i \perp \mathbf{X}_{\{1,\dots,i-1\} \setminus \mathsf{pa}(i)} | \mathbf{X}_{\mathsf{pa}(i)}$, which is true because nodes $\{1,\dots,i-1\} \setminus \mathsf{pa}(i)$ are non-descendants
$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i | \mathbf{x}_{\mathsf{pa}(i)})$$

Factorization Implies Conditional Independence

To show this, first we'll argue that marginalizing *descendants* in a Bayes net is easy:

Warmup: suppose j is a node with no children in a Bayes net (a "leaf"). Then

$$p(\mathbf{x}_{-j}) = \prod_{i \neq j} p(x_i | \mathbf{x}_{\mathsf{pa}(i)})$$

In words, we can marginalize x_j by dropping the factor $p(x_j | x_{\mathsf{pa}(j)})$ to get a Bayes net with one less node.

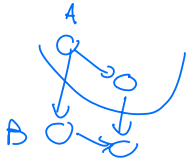
This is *only* true for leaf nodes. Marginalizing non-leaf nodes may be very hard!

Proof:

$$\begin{aligned} p(\mathbf{x}_{-j}) &= \sum_{x_j} p(\mathbf{x}_{-j}, x_j) \\ &= \sum_{x_j} p(x_j | \mathbf{x}_{\mathsf{pa}(j)}) \prod_{i \neq j} p(x_i | \mathbf{x}_{\mathsf{pa}(i)}) \\ &= \prod_{i \neq j} p(x_i | \mathbf{x}_{\mathsf{pa}(i)}) \cdot \underbrace{\sum_{x_j} p(x_j | \mathbf{x}_{\mathsf{pa}(j)})}_{1} \end{aligned}$$

Pushing the sum inside in the last line is possible because j is a leaf, so $j \notin \mathsf{pa}(i)$ for any i.

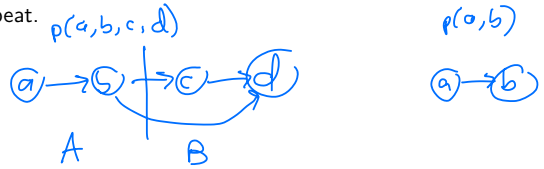
Marginalizing a *Set* of Descendants



Lemma: suppose A and B partition the nodes of a Bayes net and there is no path from B to A. Then

$$p(\mathbf{x}_A) = \sum_{\mathbf{x}_B} p(\mathbf{x}_A, \mathbf{x}_B) = \prod_{i \in A} p(x_i | \mathbf{x}_{\mathsf{pa}(i)})$$

Proof idea: at least one node in B is a leaf. Eliminate it using the warmup lemma and then repeat.



Factorization Implies Conditional Independence

Assume $p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{x}_{\mathsf{pa}(i)})$. Then for any i