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

- 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 $p(a, b, c)$?

$$p(a, b, c) = p(a)p(b|a)p(c|a, b) \qquad \text{chain rule}$$

$$= p(a)p(b|a)p(c|b) \qquad \text{conditional independence}$$

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) $CP \perp G, C | HD$

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

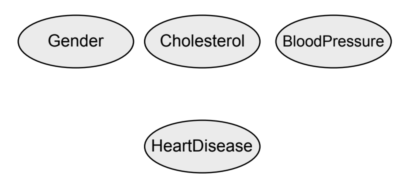
Bayesian Networks: Nodes

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

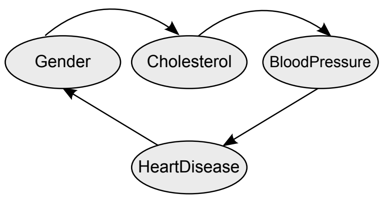
Example:



Bayesian Networks: Edges

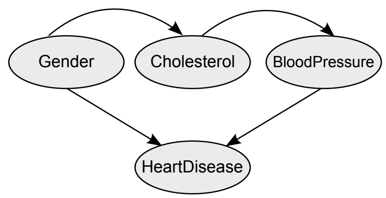
The DAG constraint means that $\mathcal{G}$ can't contain any directed cycles $i \to j \to \cdots \to i$.

Example:



Not a valid DAG
Directed Cycle

Example:



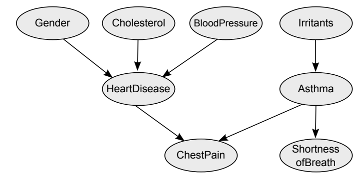
A valid DAG.
No directed cycle

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
- pa(i) = set of all parents of i
- ch(i) = set of all children of i

Example:



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

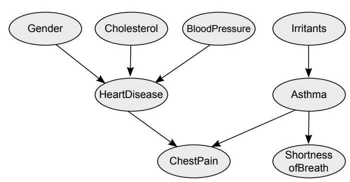
ch$(A) = \{CP, SB\}$

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
- de(i) = set of all descendants of i
- nd(i) = set of all non-descendants of i

Example:



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

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.

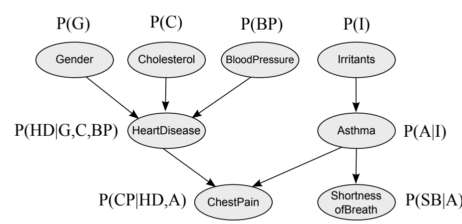


Figure 1: image

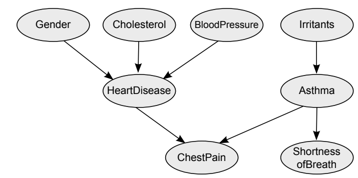
The joint distribution is the product of the conditional distributions:
$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{x}_{\text{pa}(i)})$.

Bayesian Networks: CPDs and CPTs

- The individual factors $p(x_i \mid \mathbf{x}_{\mathsf{pa}(i)})$ 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 $\mathbf{x}_{\mathsf{pa}(i)}$ 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	$p(hd \vert g, bp, ch)$
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? $V - 1$
- How much space does it take to store the CPT for a binary-valued variable with D binary-valued parents? 2^D
- Suppose there are D binary variables connected in a chain $X_1 \to X_2 \to ... \to X_D$. What is the total storage cost? $1 + 2(D - 1)$ How large is the full joint? $2^D - 1$

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 $p(\mathbf{x})$

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

factorization $\iff$ conditional independence

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

Conditional Independence Implies Factorization

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

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

$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i | \mathbf{x}_{\{1,\ldots,i-1\}})$$

2. Nodes $\{1, \ldots, i-1\}$ 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:

$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i | \mathbf{x}_{\mathsf{pa}(i)}, \mathbf{x}_{\{1,\ldots,i-1\} \setminus \mathsf{pa}(i)})$$

3. Now simplify using $X_i \perp \mathbf{X}_{\{1,\ldots,i-1\} \setminus \mathsf{pa}(i)} | \mathbf{X}_{\mathsf{pa}(i)}$, which is true because nodes $\{1, \ldots, i-1\} \setminus \mathsf{pa}(i)$ 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

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

In words, we can marginalize x_j by dropping the factor $p(x_j | x_{\text{pa}(j)})$ 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:

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

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

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

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

$$\begin{aligned} p(x_i | \mathbf{x}_{\text{nd}(i)}) &= \frac{p(x_i, \mathbf{x}_{\text{nd}(i)})}{p(\mathbf{x}_{\text{nd}(i)})} \\ &= \frac{p(x_i | \mathbf{x}_{\text{pa}(i)}) \cdot \prod_{j \in \text{nd}(i)} p(x_j | \mathbf{x}_{\text{pa}(j)})}{\prod_{j \in \text{nd}(i)} p(x_j | \mathbf{x}_{\text{pa}(j)})} \quad \text{Use lemma twice} \\ &= p(x_i | \mathbf{x}_{\text{pa}(i)}) \end{aligned}$$

This demonstrates that $X_i \perp \mathbf{X}_{\text{nd}(i)} | \mathbf{X}_{\text{pa}(i)}$ for all i.