COMPSCI 688: Probabilistic Graphical Models

Lecture 2: More Probability and Directed Graphical Models

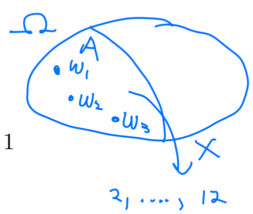
Dan Sheldon

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

Discrete Distributions



- Sample space Ω
- Atomic probability $p(\omega)$ for all $\omega \in \Omega$

$$p(\omega) \geq 0, \quad \sum_{\omega \in \Omega} p(\omega) = 1$$

- Events $A \subseteq \Omega$ (only things that have probabilities!)

$$P(A) = \sum_{\omega \in A} p(\omega) = p(\omega_1) + p(\omega_2) + p(\omega_3)$$

- Random variable $X : \Omega \to \mathrm{Val}(X)$ has probability mass function (PMF)

$$p_X(\overset{2}{x}) = P(X(\omega) = x) = P(X = x)$$

$p(x)$

Events vs Random Variables

- A random variable X is a a mapping from Ω to $\mathrm{Val}(X)$
- **But**: for any random variable X, we can also define the probability distribution with sample space $\Omega = \mathrm{Val}(X)$ and atomic probabilities $p_X(x)$. This is the **distribution** of X.
- If we only care about events involving X, it's easier to just define the distribution of X without using a different underlying probability space
- If we care about multiple random variables, we can similarly define their **joint distribution**

Joint Distributions

Random Variables and Data Sets

In ML and stats, probability distributions are defined over records described by multiple attributes modeled as random variables. This leads to joint distributions.

Attributes/variables

record

X_1 Gender	X_2 Blood Pressure	X_3 Cholesterol	X_4 Heart Disease
Male	Med	Low	No
Male	Hi	Hi	Yes
Male	Med	Med	Yes
Male	Med	Hi	No
Female	Med	Low	No
Male	Low	Med	No

Joint Probability Distributions

- The *joint distribution* of random variables $X_1, \ldots, X_N$ is a probability distribution over their *canonical sample space*
- The *canonical sample space* Ω of $X_1, \ldots, X_N$ is the Cartesian product of their domains $\Omega = \mathrm{Val}(X_1) \times \ldots \times \mathrm{Val}(X_N)$.
- An element of Ω is a joint assignment $(x_1, \ldots, x_N)$
- The joint probability mass function of $X_1, \ldots, X_N$ is

$$p(x_1, \ldots, x_N) = P(X_1 = x_1, \ldots, X_N = x_N)$$

"and"

$P(X=a, Y=b) = P([X_1=a] \cap [X_2=b])$

Joint Distributions: Heart Disease Example

Example: The joint distribution over random variables *Gender*, *BloodPressure*, *Cholesterol* and *HeartDisease* is given by a table like this:

all possible records

Ω

Gender	BloodPressure	Cholesterol	HeartDisease	P
F	L	L	N	0.0127
F	L	L	Y	0.0007
F	L	M	N	0.0098
F	L	M	Y	0.0009
F	L	H	N	0.0087
F	L	H	Y	0.0010
...	...	...	...	...

exponential size in # variables

Random Vectors

$p(x_1,\dots,x_N)$ $\quad p(x)$ $\quad x=(x_1,\dots,x_N)$

- It's convenient to use vector-valued random variables $\mathbf{X} = (X_1, \dots, X_N)$ (or "random vectors") and assignments $\mathbf{x} = (x_1, \dots, x_N)$:

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, \dots, X_N = x_N)$$

- The PMF is $p_{\mathbf{X}}(\mathbf{x})$ or just $p(\mathbf{x})$
- This is just notation: it means the same thing as a joint distribution over $(X_1, \dots, X_N)$
- **Notation**: use $\mathbf{X}_{-i}$ and $\mathbf{x}_{-i}$ for vectors excluding X_i or x_i

$$(x_1,\dots,x_{i-1},x_{i+1},\dots,x_N)$$

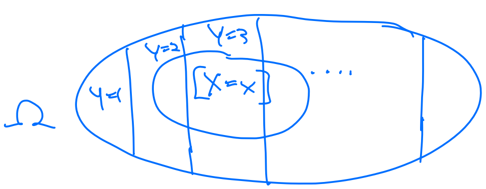
Rules of Probability

Marginal Distributions

$(X_1,\dots,X_N, Y_1,\dots,Y_M)$

- Suppose we have a joint distribution $P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})$.
- $P(\mathbf{X} = \mathbf{x})$ is called a *marginal distribution*. How can we find $P(\mathbf{X} = \mathbf{x})$?

$$P(X=x) = \sum_{y \in Val(Y)} P(X=x, Y=y)$$



X \ Y	1	2	P(X=x)
1	.2	.3	.5
2	.2	.3	.5
	.4	.6	

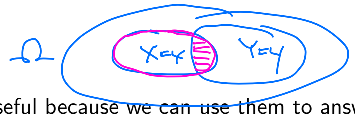
Marginal Distributions: Heart Disease Example

Given a joint distribution on G, BP, C, HD, we obtain the marginal probability $P(G = M, BP = H, C = H)$ as follows:

$$P(G=M, BP=H, C=H) = 0.050 + 0.005 = 0.055$$

Gender	BloodPressure	Cholesterol	HeartDisease	P
M	H	H	Y	0.050
M	H	H	N	0.005
M	H	M	Y	0.045
M	H	M	N	0.008
...	...	...	...	...

Conditional Distributions



- Joint distributions are useful because we can use them to answer queries like "What is the probability that $\mathbf{Y}=\mathbf{y}$ given that I observed $\mathbf{X}=\mathbf{x}$?":

$$P(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y})}{P(\mathbf{X}=\mathbf{x})} \quad \begin{matrix}\text{joint}\\ \text{marginal}\end{matrix}$$

$$= \frac{P(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y})}{\sum_{\mathbf{y}\in \mathrm{Val}(\mathbf{Y})} P(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y})}$$

free fixed at observed value

$P_{X,Y}(x,y) = P(X=x, Y=y)$

- Write $p(\mathbf{y}|\mathbf{x})$ to denote the PMF of $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$

$P_{Y|X}(y|x)$ $\quad P(Y=y|X=1)\ \forall y$ $\quad P_{X|Y}(x|y)$ $\quad P_{Y|X}(y|x)$

Conditional Distributions: Heart Disease Example

$$P(HD=Y|G=M,BP=H,C=H) = \frac{P(G=M,BP=H,C=H,HD=Y)}{P(G=M,BP=H,C=H)}$$
$$= \frac{0.050}{0.050+0.005} = 0.91$$

Gender	BloodPressure	Cholesterol	HeartDisease	P
M	H	H	Y	0.050
M	H	H	N	0.005
M	H	M	Y	0.045
M	H	M	N	0.008
...	...	...	...	...

Chain Rule

$p(y|x) = \frac{p(x,y)}{p(x)}$

- By rearranging the definition of conditional probability, we get the chain rule:

$P(X=x, Y=y) = P(X=x)P(Y=y|X=x)$

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

- Applying the chain rule repeatedly to a random vector $\mathbf{X}$ gives:

$$p(\mathbf{x}) = p(x_N|x_1,...,x_{N-1})p(x_1,...,x_{N-1})$$
$$\vdots = p(x_N|x_1,...,x_{N-1})p(x_{N-1}|x_1,...,x_{N-2})p(x_1,...,x_{N-2})$$
$$= p(x_N|x_1,...,x_{N-1})p(x_{N-1}|x_1,...,x_{N-2})\cdots p(x_3|x_2,x_1)p(x_2|x_1)p(x_1)$$
$$= \prod_{i=1}^{N} p(x_i|x_1,\dots,x_{i-1})$$

Chain Rule: Heart Disease Example

We can apply the chain rule using any ordering of the variables:

1 2 3 4
$$p(g,bp,c,hd) = p(hd|c,bp,g)p(c|bp,g)p(bp|g)p(g)$$
4 3 2 1
$$p(g,bp,c,hd) = p(g|bp,c,hd)p(bp|c,hd)p(c|hd)p(hd)$$
2 1 4 3
$$p(g,bp,c,hd) = p(c|hd,g,bp)p(hd|g,bp)p(g|bp)p(bp)$$

Bayes' Rule

hidden — observed/evidence — "world model" — measurement model

$x \xrightarrow{p(y|x)} y$, $\quad y \xrightarrow{p(x|y)} x$ $\qquad p(x) \qquad p(y|x)$

- By using the definition of conditional probability twice, we obtain one of the most important equations in probability theory, Bayes' Rule: prior

p(hidden | observed)

posterior $\longrightarrow$ $$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}$$

- Bayes' rule lets us compute $p(\mathbf{x}|\mathbf{y})$ from a joint distribution specified by $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$.

$$p(y) = \sum_x p(x,y) = \sum_x p(y|x)p(x)$$

Conditional Independence

Probabilistic Models

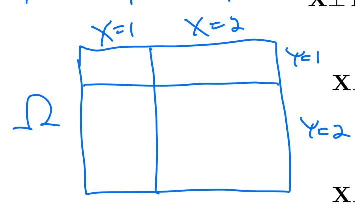
The solution to the problem of exponential-sized joint distributions is the use of **compact** probabilistic models.

- Bayesian networks achieve compactness by exploiting the chain rule and asserting (conditional) independence relations
- As a result, Bayesian networks can express high-dimensional distributions as products of simpler factors.

Marginal Independence

$$P(X=x, Y=y) = P(X=x)P(Y=y)$$

"X is indept. of Y"



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

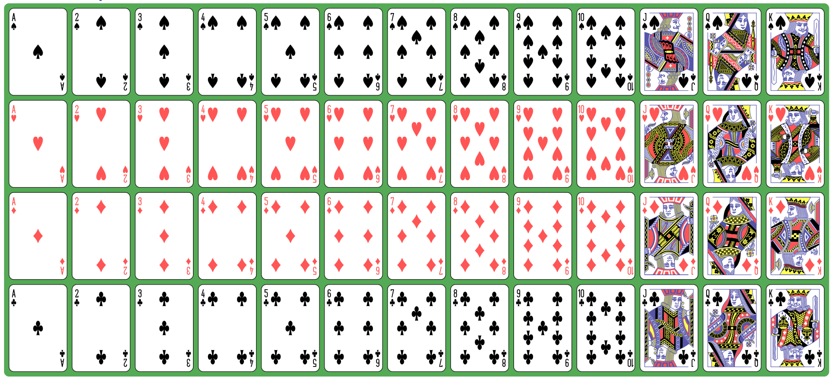
$p(x)p(y|x)$

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

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

Card Example I

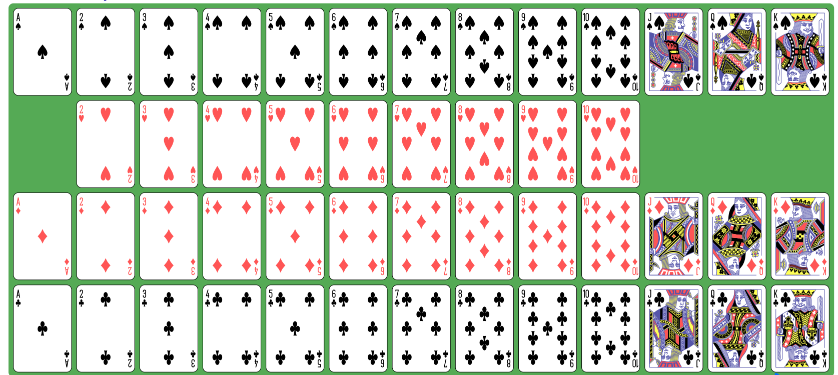
$P(\text{value}=2 \mid \text{col}=r) = \frac{1}{13}$ $P(\text{value}=2) = \frac{1}{13}$



Draw a random card: is value $\perp$ color? yes

Card Example II

$P(\text{val}=J \mid \text{col}=r) \neq P(\text{val}=J)$



What about with this deck? Is value $\perp$ color? $P(\text{col}=r \mid \text{val} \in \{2,4,6,8\})$

Conditional Independence

$P(X=x, Y=y, Z=z)$

X is conditionally independent of Y given Z if

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

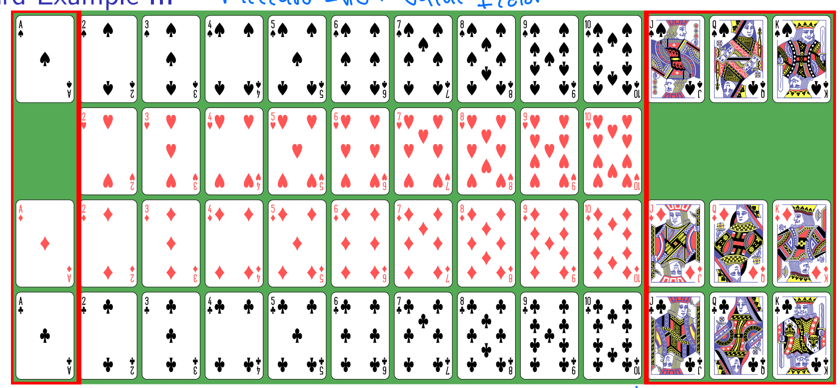
$$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \iff p(x \mid y, z) = p(x \mid z)$$

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

"Given z, knowing y tells you nothing about the dist. of x"

Card Example III

facecard = yes: value $\perp$ color
facecard = no: value $\perp$ color



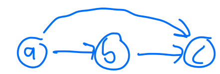
Is value $\perp$ color | facecard? $P(\text{val}=J \mid \text{col}=bl, \text{face}=y) = \frac{1}{4}$
$P(\text{val}=J \mid \text{face}=y) = \frac{1}{4}$

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 $p(a, b, c)$?

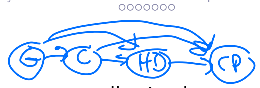
chain rule $p(a,b,c) = p(a)\,p(b|a)\,p(c|\cancel{a},b)$



||

conditional indep. $p(a,b,c) = p(a)\,p(b|a)\,p(c|b)$

Bayesian Networks: Main Idea

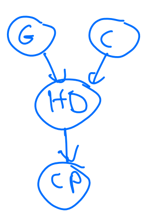


- The main idea of Bayesian networks is conceptually simple:

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

$CP \perp G \mid HD$

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



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

|| ||

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

$\mathcal{G}$, $p(x)$

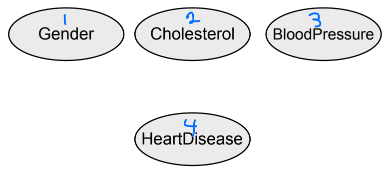
Formally, a Bayesian network consists of a directed acyclic graph (DAG) $\mathcal{G}$ and a joint distribution $p(\mathbf{x}) = p(x_1, \dots, x_N)$ for random variables $X_1, \dots, X_N$

$V = \{1, \dots, N\}$

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

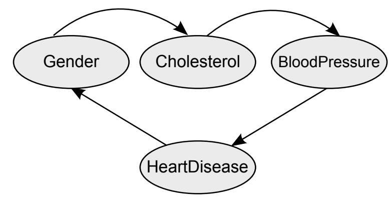
Example:



Bayesian Networks: Edges

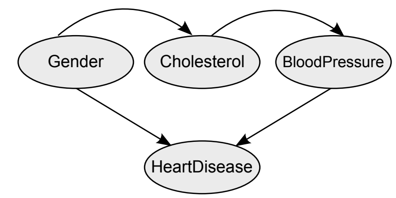
The DAG constraint means that $\mathcal{G}$ can't contain any directed cycles $i \to j \to \cdots \to i$.

Example:



Not a valid DAG
Directed Cycle

Example:



A valid DAG.
No directed cycle

Bayesian Networks: Parents/Children

$i \to j$
parent child

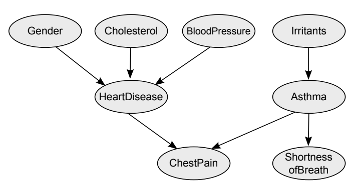
If there is a directed edge $i \to j$:

- 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:



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

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

Bayesian Networks: Descendants/Non-Descendants

$i \rightsquigarrow j$

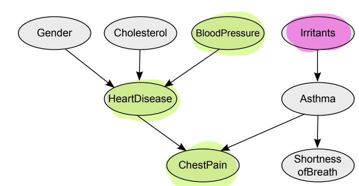
If there is a directed path from i to j:

- 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:



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

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

Bayesian Networks: Joint Distribution

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

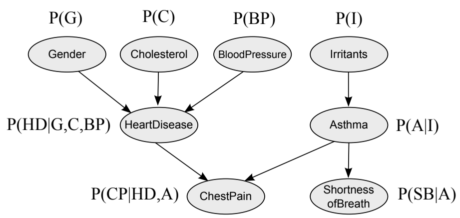


Figure 1: image

$p(g, c, bp, \ldots, sb) = p(g)\,p(c)\,p(bp)\,p(i)\,p(hd \mid g, c, bp)\,p(a \mid i)\,p(cp \mid hd, a)\,p(sb \mid a)$

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

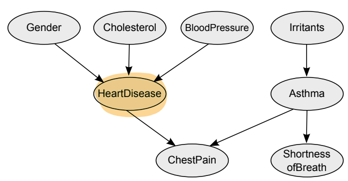
- The individual factors $p(x_i \mid \mathbf{x}_{\text{pa}(i)})$ in a Bayesian network are referred to as conditional probability distributions or CPDs.
- The CPD for node i must specify the probability that X_i takes any value x_i in its domain when conditioned on each joint assignment $\mathbf{x}_{\text{pa}(i)}$ 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

$p(hd, g \mid bp)$

2^D

hd	g	bp	ch	$p(hd \mid g, bp, ch)$	
No	M	Low	Low	0.95	1
Yes	M	Low	Low	0.05	
No	F	Low	Low	0.99	1
Yes	F	Low	Low	0.01	
⋮					



exponential in (# parents + 1)

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? $p(a \mid b_1, \ldots, b_D)$ $\quad 2^D \cdot (V-1) = 2^D \quad V=2$
- 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? $2 \cdot (D-1) + 1 = 2D-1$
How large is the full joint? 2^D

Next Time

Next time, we'll discuss factorization and conditional independence in Bayesian networks.