CMPSCI 240: Reasoning about Uncertainty

Lecture 17: Representing Joint PMFs and Bayesian Networks

Andrew McGregor

University of Massachusetts

Last Compiled: April 7, 2017

Warm Up: Joint distributions

- Recall Question 5 from Homework 2:
  Suppose you toss four fair coins (C_1, C_2, C_3, C_4). The outcome is the result of the four tosses. Let X correspond to the number of heads in the outcome. Let Y be 1 if there are an even number of heads in the outcome and 0 otherwise.
- Consider the joint distribution of $P(C_1, C_2, C_3, C_4, X, Y)$
- **Clicker Question**: How many rows are in the table of the joint distribution?

  A 15
  B 16
  C 64
  D 159
  E 160

Warm Up: Joint distributions

- Consider a sample table:

C_1	C_2	C_3	C_4	X	Y	$\mathbb{P}$
H	H	H	H	0	0	0
H	H	H	H	0	1	0
...	...	...	...	...	...	...
H	H	H	H	4	1	$\frac{1}{16}$
...	...	...	...	...	...	...

- Naively, the joint distribution is over all possible values of C_1, C_2, C_3, C_4, X, and Y: $2^4 \times 5 \times 2 = 160$
- Intuitively though, if we know C_1, C_2, C_3, C_4, the rest are deterministic (so 16...?)...

Outline

1. The Problem with Joint PMFs
2. Approximating Joint PMFs
3. Bayesian Networks
4. Using Bayesian Networks
5. Learning Bayesian Networks from data

Many Random Variables

- So far we have been discussing estimation problems assuming that we measure just one or two random variables on each repetition of an experiment. e.g., Spam filtering in last lecture:

$$D_1 = \text{email contains ``cash''}$$

$$D_2 = \text{email contains ``pharmacy''}$$

$$P(H = spam\,|D_1, D_2)$$

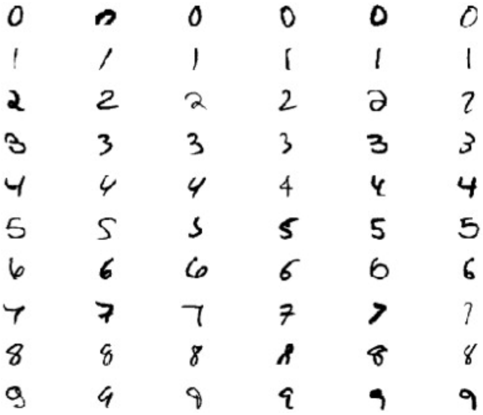
- In practice, it is much more common to encounter real-world problems that involve measuring multiple random variables $X_1, ..., X_d$ for each repetition of the experiment.
- These random variables $X_1, ..., X_d$ may have complex relationships among themselves.

Example: ICU Monitoring ($d \approx 10$)

Heart rate, blood pressure, temperature....

Example: Handwritten Digit Recognition ($d \approx 1000$)

Important to look at the digits together...easy to get files with 1000 characters!



Example: Movie Recommendation ($d \approx 10,000$)

A complex decision process. Needs to look at ratings and viewing patterns of a large number of subscribers.

Joint PMFs for Many Random Variables

- Before we can think about inference or estimation problems with many random variables, we need to think about the implications of representing joint PMFs over many random variables.
- Suppose we have an experiment where we obtain the values of d random variables $X_1, ..., X_d$ per repetition. For now, assume all the random variables are binary.
- **Rhetorical Question:** If we have d binary random variables, how many numbers does it take to write down a joint distribution for them?

Joint PMFs for Many Random Variables

- **Rhetorical Question:** If we have d binary random variables, how many numbers does it take to write down a joint distribution for them?
- **Answer:** We need to define a probability for each d-bit sequence:

$$\begin{gathered} P(X_1 = 0, X_2 = 0, ..., X_d = 0) \\ P(X_1 = 1, X_2 = 0, ..., X_d = 0) \\ \vdots \\ P(X_1 = 1, X_2 = 1, ..., X_d = 1) \end{gathered}$$

- The number of d-bit sequences is 2^d. Because we know that the probabilities have to add up to 1, **we only need to write down** $2^d - 1$ **numbers** to specify the full joint PMF on d binary variables.

How Fast is Exponential Growth?

- $2^d - 1$ grows **extremely fast** as d increases:

d	$2^d - 1$
1	1
10	1023
100	1,267,650,600,228,229,401,496,703,205,375
$\vdots$	$\vdots$

- Storing the full joint PMF for 100 binary variables would take about 10^{30} real numbers or about 10^{18} **terabytes** of storage!
- Joint PMFs grow in size so rapidly, we have no hope whatsoever of storing them explicitly for problems with more than about 30 random variables.

Outline

1. The Problem with Joint PMFs
2. Approximating Joint PMFs
3. Bayesian Networks
4. Using Bayesian Networks
5. Learning Bayesian Networks from data

Factorizing Joint Distributions

- We start by *factorizing the joint distribution*, i.e., re-writing the joint distribution as a product of conditional PMFs over single variables (called factors).
- We obtain this form by iteratively applying the multiplication rule and conditional multiplication rule for random variables.

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(\mathbf{A},\mathbf{B}|C) = P(\mathbf{A}|\mathbf{B},C)P(\mathbf{B}|C) = P(\mathbf{B}|\mathbf{A},C)P(\mathbf{A}|C)$$

Recall: Multiplication rules derived from definition of conditional probability

Factorizing Joint Distributions

- Let's start by specifying an ordering on the variables. Consider the random variables X, Y, Z. Use the order X_1, Y_2, Z_3 and pick a variable to condition on:

$$\begin{aligned} &P(\mathbf{X_1}=x, Y_2=y, Z_3=z) \\ &\quad = P(Y_2=y, \mathbf{Z_3}=z|\mathbf{X_1}=x)P(\mathbf{X_1}=x) \\ &\quad = P(Y_2=y|X_1=x, \mathbf{Z_3}=z)P(\mathbf{Z_3}=z|X_1=x)P(X_1=x) \end{aligned}$$

- So what? The representation is exact and has exactly the same storage requirements as the full joint PMF.

Conditional Independence: Simplification 1

- To get a savings in terms of storage, we need to assume some additional conditional independencies.
- Suppose we assume that:
  - $P(\mathbf{Z_3}=z|X_1=x) = P(\mathbf{Z_3}=z)$ for all x, z
  - $P(\mathbf{Y_2}=y|X_1=x, Z_3=z) = P(\mathbf{Y_2}=y)$ for all x, y, z.
- This gives the “Marginal independence model”

$$
\begin{aligned}
&P(X_1 = x, \mathbf{Y_2} = y, \mathbf{Z_3} = z) \\
&\quad = P(\mathbf{Y_2}=y|X_1=x, Z_3=z)P(\mathbf{Z_3}=z|X_1=x)P(X_1=x) \\
&\quad = P(\mathbf{Y_2}=y)P(\mathbf{Z_3}=z)P(X_1=x)
\end{aligned}
$$

Clicker Question

$$\begin{aligned} &P(X_1 = x, \mathbf{Y_2} = y, \mathbf{Z_3} = z) \\ &\quad = P(\mathbf{Y_2}{=}y|X_1{=}x, Z_3{=}z)P(\mathbf{Z_3}{=}zX_1{=}x)P(X_1{=}x) \\ &\quad = P(\mathbf{Y_2}{=}y)P(\mathbf{Z_3}{=}z)P(X_1{=}x) \end{aligned}$$

- **Clicker Question:** How many numbers do we need to store for three binary random variables in this case (Recall: X, Y, and Z are all binary)?
  (A) 2 **(B)** 3 **(C)** 4 **(D)** 5 **(E)** 6
- **Answer:** 3 (as opposed to $2^3 - 1 = 7$ if we encoded the full joint)

Conditional Independence: Simplification 2

- Suppose we instead only assume that:
  - $P(\mathbf{Y_2}=y|\mathbf{X_1}=x, \mathbf{Z_3}=z) = P(\mathbf{Y_2}=y|\mathbf{X_1}=x)$ for all x, y, z.
  - Note how this differ from the previous assumptions:
    - Before: marginally independent;
      Now: conditionally independent.
    - Mechanically, dropped *all* conditions before;
      Only dropping one now.
- This gives the "conditional independence model":
  Y_2 is conditionally independent of Z_3 given X_1

$$\begin{aligned} &P(\mathbf{X_1}=x, \mathbf{Y_2}=y, \mathbf{Z_3}=z) \\ &\quad = P(\mathbf{Y_2}=y|\mathbf{X_1}=x, \mathbf{Z_3}=z)P(Z_3=z|X_1=x)P(X_1=x) \\ &\quad = P(\mathbf{Y_2}=y|\mathbf{X_1}=a_1)P(Z_3=z|X_1=x)P(X_1=x) \end{aligned}$$

Clicker Question

$$
\begin{aligned}
P(\mathbf{X_1} = x, \mathbf{Y_2} = y, \mathbf{Z_3} = z) \\
&= P(\mathbf{Y_2}=y|\mathbf{X_1}=x, \mathbf{Z_3}=z)P(Z_3=z|X_1=x)P(X_1=x) \\
&= P(\mathbf{Y_2}=y|\mathbf{X_1}=a_1)P(Z_3=z|X_1=x)P(X_1=x)
\end{aligned}
$$

- **Clicker Question:** How many numbers do we need to store for three binary random variables in this case?
  (A) 2 **(B)** 3 **(C)** 4 **(D)** 5 **(E)** 6
- **Answer:** $2+2+1=5$ (as opposed to $2^3-1=7$ if we encoded the full joint)

Conditional Independence: Simplification 3

- Suppose we instead only assume that:
  - $P(\mathbf{Y_2}{=}y|\mathbf{X_1}{=}x, \mathbf{Z_3}{=}z) = P(\mathbf{Y_2}{=}y|\mathbf{Z_3}{=}z)$ for all x, y, z.

$$\begin{aligned} &P(X_1 = x, Y_2 = y, Z_3 = z) \\ &\quad = P(X_1{=}x)P(\mathbf{Y_2}{=}y|\mathbf{X_1}{=}x)P(Z_3{=}z|X_1{=}x,\, Y_2{=}y) \end{aligned}$$

- Stuck!
- Factorization by itself is not a panacea

Example

- Toothache: boolean variable indicating whether the patient has a toothache
- Cavity: boolean variable indicating whether the patient has a cavity
- Catch: whether the dentists probe catches in the cavity

If the patient has a cavity, the probability that the probe catches in it doesn't depend on whether he/she has a toothache
$P(Catch|Toothache, Cavity) = P(Catch|Cavity)$

Therefore, Catch is conditionally independent of Toothache given Cavity

Likewise, Toothache is conditionally independent of Catch given Cavity
$P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$

Equivalent statement:
$P(Toothache, Catch|Cavity) = P(Toothache|Cavity)P(Catch|Cavity)$

Rhetorical Question: What is the space requirement to represent the Joint Distribution $P(Toothache, Catch, Cavity)$? **Answer:** 5

Outline

1 The Problem with Joint PMFs

2 Approximating Joint PMFs

3 Bayesian Networks

4 Using Bayesian Networks

5 Learning Bayesian Networks from data

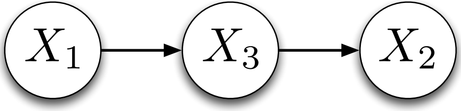
Bayesian Networks

- Keeping track of all the conditional independence assumptions gets tedious when there are a lot of variables.
    - Consider the following situation: You live in quiet neighborhood in the suburbs of LA. There are two reasons the alarm system in your house will go off: your house is broken into or there is an earthquake. If your alarm goes off you might get a call from the police department. You might also get a call from your neighbor.
- To get around this problem, we use "Bayesian Networks" to express the conditional independence structure of these models.

Bayesian Networks

- A Bayesian network uses conditional independence assumptions to more compactly represent a joint PMF of many random variables.
- We use a directed acyclic graph to encode conditional independence assumptions.
    - Remember how we ordered X_1, Y_2, Z_3? Now let X_i be an ordered node in graph G.
    - Nodes X_i in the graph G represent random variables.
    - A directed edge $X_j \rightarrow X_i$ means X_j is a “parent” of X_i.
        - This means X_i directly depends on X_j
    - The set of variables that are parents of X_i is denoted Pa_i.
        - Each variable is conditionally independent of all its nondescendants in the graph given the value of all its parents.
        - The factor associated with variable X_i is $P(X_i | Pa_i)$.

Example: Bayesian Network



$$Pa_1 = \{\}, Pa_3 = \{X_1\}, Pa_2 = \{X_3\}$$

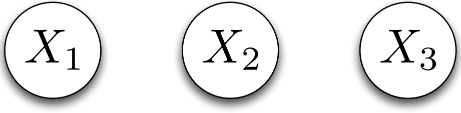
$$P(X_1 = a_1, X_2 = a_2, X_3 = a_3) =$$

$$P(X_1 = a_1)P(X_3 = a_3 | X_1 = a_1)P(X_2 = a_2 | X_3 = a_3)$$

From Graphs to Factorizations and Back

- If we have a valid graph, we can infer the parent sets and the factors.
- If we have a valid set of factors, we can infer the parent sets and the graph.

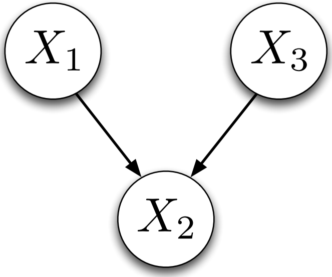
Example: Graph to Factorization



$$Pa_1 = \{\}, Pa_3 = \{\}, Pa_2 = \{\}$$

$$P(X_1, X_2, X_3) = P(X_1)P(X_3)P(X_2)$$

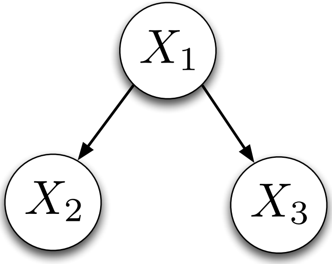
Example: Graph to Factorization



$$Pa_1 = \{\}, Pa_3 = \{\}, Pa_2 = \{X_1, X_3\}$$

$$P(X_1, X_2, X_3) = P(X_1)P(X_3)P(X_2|X_1, X_3)$$

Example: Graph to Factorization



$$Pa_1 = \{\}, Pa_3 = \{X_1\}, Pa_2 = \{X_1\}$$

$$P(X_1 = a_1, X_2 = a_2, X_3 = a_3) =$$
$$P(X_1 = a_1) P(X_3 = a_3 | X_1 = a_1) P(X_2 = a_2 | X_1 = a_1)$$

The Bayesian Network Theorem

- **Definition:** A joint PMF $P(X_1, ..., X_d)$ is a Bayesian network with respect to a directed acyclic graph G with parent sets $\{Pa_1, ..., Pa_d\}$ if and only if: $P(X_1, ..., X_d) = \prod_{i=1}^{d} P(X_i|Pa_i)$
- In other words, to be a valid Bayesian network for a given graph G, the joint PMF must factorize according to G.

Outline

1 The Problem with Joint PMFs

2 Approximating Joint PMFs

3 Bayesian Networks

4 Using Bayesian Networks

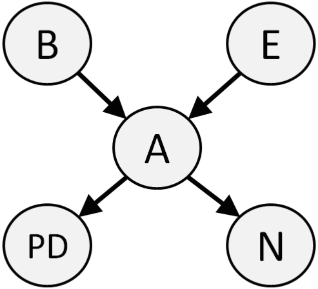
5 Learning Bayesian Networks from data

The Alarm Network: Random Variable

- Consider the following situation: You live in quiet neighborhood in the suburbs of LA. There are two reasons the alarm system in your house will go off: your house is broken into or there is an earthquake. If your alarm goes off you might get a call from the police department. You might also get a call from your neighbor.
- **Rhetorical Question:** What random variables can we use to describe this problem?
- **Answer:** Break-in (B), Earthquake (E), Alarm (A), Police Department calls (PD), Neighbor calls (N).

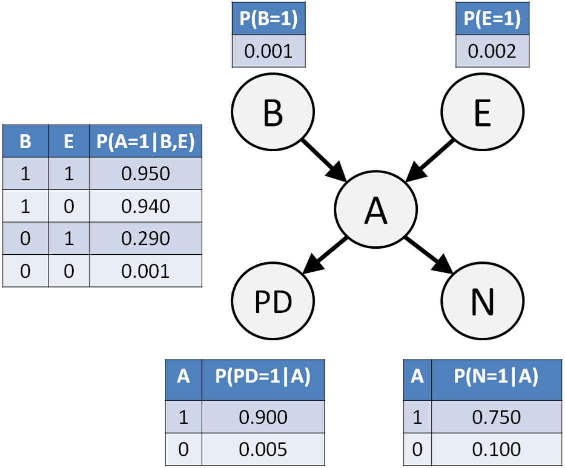
The Alarm Network: Factorization

- **Rhetorical Question** What direct dependencies might exist between the random variables B, E, A, PD, N?



- **Rhetorical Question:** What is the factorization implied by the graph?
- **Answer:**
  $P(B, E, A, PD, N) = P(B)P(E)P(A|B, E)P(PD|A)P(N|A)$

The Alarm Network: Factor Tables



P(B=1)
0.001

P(E=1)
0.002

B	E	P(A=1|B,E)
1	1	0.950
1	0	0.940
0	1	0.290
0	0	0.001

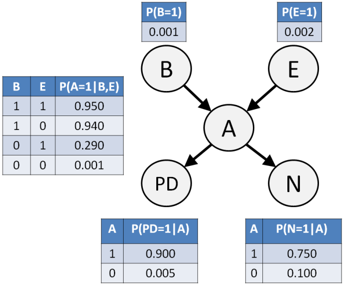
A	P(PD=1|A)
1	0.900
0	0.005

A	P(N=1|A)
1	0.750
0	0.100

The Alarm Network: Joint Query

- **Question:** What is the probability that there is a break-in, but no earthquake, the alarm goes off, the police call, but your neighbor does not call?

$$
\begin{aligned}
&P(B{=}1, E{=}0, A{=}1, PD{=}1, N{=}0)\\
&= P(B{=}1)P(E{=}0)P(A{=}1|B{=}1, E{=}0)P(PD{=}1|A{=}1)P(N{=}0|A{=}1)\\
&= 0.001 \cdot (1 - 0.002) \cdot 0.94 \cdot 0.9 \cdot (1 - 0.75)
\end{aligned}
$$



P(B=1)
0.001

P(E=1)
0.002

B	E	P(A=1|B,E)
1	1	0.950
1	0	0.940
0	1	0.290
0	0	0.001

A	P(PD=1|A)
1	0.900
0	0.005

A	P(N=1|A)
1	0.750
0	0.100

The Alarm Network: Marginal Query

- **Question:** What is the probability that there was a break-in, but no earthquake, the police call, but your neighbor does not call?

$$
\begin{aligned}
&P(B=1, E=0, PD=1, N=0)\\
&= P(B=1, E=0, A=0, PD=1, N=0) + P(B=1, E=0, A=1, PD=1, N=0)\\
&= P(B=1)P(E=0)P(A=1|B=1, E=0)P(PD=1|A=1)P(N=0|A=1)\\
&+ P(B=1)P(E=0)P(A=0|B=1, E=0)P(PD=1|A=0)P(N=0|A=0)\\
&= 0.001 \cdot (1-0.002) \cdot 0.94 \cdot 0.9 \cdot (1-0.75)\\
&+ 0.001 \cdot (1-0.002) \cdot (1-0.94) \cdot 0.005 \cdot (1-0.1) = 0.00021\ldots
\end{aligned}
$$

The Alarm Network: Conditional Query

- **Question:** What is the probability that the alarm went off given that there was a break-in, but no earthquake, the police call, but your neighbor does not call?

$$P(A = 1|B{=}1, E{=}0, PD{=}1, N{=}0)$$

$$= \frac{P(B{=}1, E{=}0, A{=}1, PD{=}1, N{=}0)}{\sum_{a=0}^{1} P(B{=}1, E{=}0, A{=}a, PD{=}1, N{=}0)}$$

$$= \frac{P(B{=}1)P(E{=}0)P(A{=}1|B{=}1, E{=}0)P(PD{=}1|A{=}1)P(N{=}0|A{=}1)}{\sum_{a=0}^{1} P(B{=}1)P(E{=}0)P(A{=}a|B{=}1, E{=}0)P(PD{=}1|A{=}a)P(N{=}0|A{=}a)}$$

Answering Probabilistic Queries

- **Joint Query:** To compute the probability of an assignment to all of the variables we simply express the joint probability as a product over the individual factors. We then look up the correct entries in the factor tables and multiply them together.
- **Marginal Query:** To compute the probability of an observed subset of the variables in the Bayesian network, we sum the joint probability of all the variables over the possible configurations of the unobserved variables.
- **Conditional Query:** To compute the probability of one subset of the variables given another subset, we first apply the conditional probability formula and then compute the ratio of the resulting marginal probabilities.

Outline

1 The Problem with Joint PMFs

2 Approximating Joint PMFs

3 Bayesian Networks

4 Using Bayesian Networks

5 Learning Bayesian Networks from data

Estimating Bayesian Networks from Data

- Just as with simpler models like the biased coin, we can estimate the unknown model parameters from data.
- If we have data consisting of n simultaneous observations of all of the variables in the network, we can easily estimate the entries of each conditional probability table using separate estimators.

Estimating Bayesian Networks: Counting

- **No Parents:** For a variable X with no parents, the estimate of $P(X = x)$ is just the number of times that the variable X takes the value x in the data, divided by the total number of data cases n.
- **Some Parents:** For a variable X with parents $Y_1, ..., Y_p$, the estimate of $P(X = x | Y_1 = y_1, ..., Y_p = y_p)$ is just the number of times that the variable X takes the value x when the parent variables $Y_1, ..., Y_p$ take the values $y_1, ..., y_p$, divided by the total number of times that the parent variables take the values $y_1, ..., y_p$.

Computing the Factor Tables from Observations

- Suppose we have a sample of data as shown below. Each row *i* is a joint configuration of all of the random variables in the network.

E	B	A	PD	N
1	0	1	1	1
0	0	0	0	1
0	0	1	1	0
0	1	1	1	0
0	0	0	0	0

- In the alarm network, consider the factor $P(E)$. We need to estimate $P(E = 0)$ and $P(E = 1)$.
- Given our data sample, we get the answers $P(E = 0) = 4/5$ and $P(E = 1) = 1/5$.

Computing the Factor Tables from Observations

- In the alarm network, consider the factor $P(N|A)$. We need to estimate $P(N=0|A=0)$,$P(N=1|A=0)$, $P(N=0|A=1)$, $P(N=1|A=1)$. How can we do this?

E	B	A	PD	N
1	0	1	1	1
0	0	0	0	1
0	0	1	1	0
0	1	1	1	0
0	0	0	0	0

- $P(N=0|A=0)=\frac{1}{2}, P(N=1|A=0)=\frac{1}{2}$
- $P(N=0|A=1)=\frac{2}{3}, P(N=1|A=1)=\frac{1}{3}$

Learning the structure of a Bayesian Network from data

- What if you have a dataset, but you do not know the dependencies that exist between the random variables?
- In other words, you do not know what is the graph of your Bayesian network.
- You can estimate the structure of the graph from data!