CMPSCI 240: Reasoning about Uncertainty

Lecture 18: More Bayesian Networks and Balls and Bins

Andrew McGregor

University of Massachusetts

Last Compiled: April 4, 2017

Outline

Factorization

Given events A, B, C the multiplication rule said:

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$$

Note that we can factorize in different orders, e.g.,

$$P(A \cap B \cap C) = P(C \cap A \cap B) = P(C)P(A|C)P(B|A \cap C)$$

Same is true with events based on random variables, e.g.,

$$P(X = a, Y = b, Z = c) = P(X = a)P(Y = b|X = a)P(Z = c|X = a, Y = b)$$

and

$$P(X = a, Y = b, Z = c) = P(Z = c)P(X = a|Z = c)P(Z = c|X = a, Y = b)$$

Factorization

When you have many variables, things get messy and it's hard to store all the conditional probabilities:

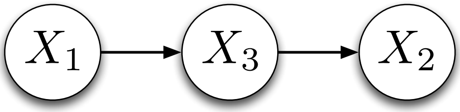
$$
\begin{aligned}
P(W=a, X=b, Y=c, Z=d) \quad = \quad & P(W=a) \\
& \times P(X=b|W=a) \\
& \times P(Y=c|W=a, X=b) \\
& \times P(Z=d|W=a, X=b, Y=c)
\end{aligned}
$$

Even if each of these four random variables only takes 10 different values, you'd need roughly $10^4 = 10000$ different probabilities to keep track of the joint distribution. Wouldn't it be nice if you could ignore some of the conditioning? This is where Bayesian networks come in.

Bayesian Networks

- A Bayesian network uses conditional independence assumptions to compactly represent a joint PMF of random variables $X_1, \ldots, X_n$.
- We use a directed acyclic graph to encode conditional independence assumptions.
  - The n nodes represent the random variables $X_1, \ldots, X_n$.
  - A directed edge $X_j \rightarrow X_i$ means X_j is a "parent" of X_i.
  - The set of variables that are parents of X_i is denoted Pa_i.

Example: Bayesian Network



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

For any values a_1, a_2, a_3,

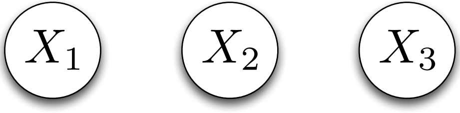
$$P(X_1 = a_1, X_2 = a_2, X_3 = a_3) =$$

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

The above statement is commonly abbreviated as

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

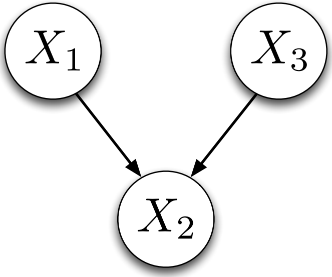
Example: Graph to Factorization



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

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

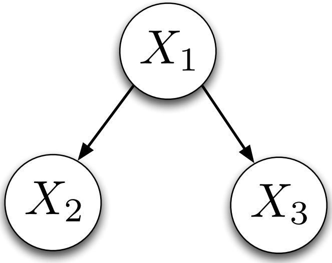
Example: Graph to Factorization



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

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

Example: Graph to Factorization



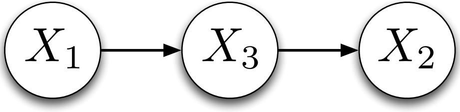
$$Pa_1 = \{\}, Pa_3 = \{X_1\}, Pa_2 = \{X_1\}$$

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

Example: Factorization to Graph

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

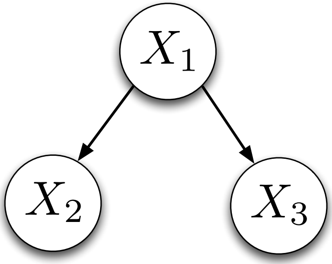
$$Pa_1 = \{\}, Pa_3 = \{X_1\}, Pa_2 = \{X_3\}$$



Example: Factorization to Graph

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

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



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:

$$P(X_1, ..., X_d) = \prod_{i=1}^{d} P(X_i | Pa_i)$$

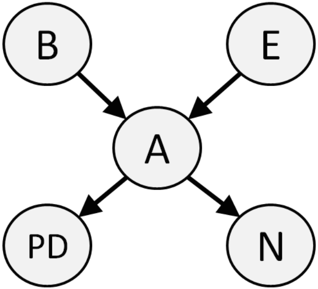
- In other words, to be a valid Bayesian network for a given graph G, the joint PMF must factorize according to G.

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.
- **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

- Denote the Boolean random variables: Break-in (B), Earthquake (E), Alarm (A), Police Department calls (PD), Neighbor calls (N). E.g., $\{B = 1\}$ is the event there is a break-in and $\{B = 0\}$ is the even that there isn't a break-in.
- The joint distribution can be factorized as follows:



$$P(B, E, A, PD, N) = P(B)P(E)P(A|B, E)P(PD|A)P(N|A)$$

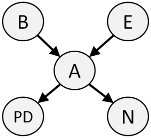
The Alarm Network: Factor Tables

Factor tables give us the necessary information to compute everything we want to know:

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

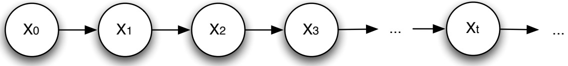
A	P(N=1|A)
1	0.750
0	0.100

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

Outline

Relationship to Bayesian Networks

- The infinite set of random variables defined by a Markov chain, $X_0, X_1, X_2, \ldots$ have a Bayesian Network with a special form.
- Since X_t only "depends" on X_{t-1}, i.e., X_t is independent of $X_0, X_1, \ldots, X_{t-2}$ conditioned on X_{t-1}, a the Bayesian network is:



- For example,

$$\begin{aligned} & P(X_0 = a_0, X_1 = a_1, \ldots, X_t = a_t) \\ = \quad & P(X_0 = a_0)P(X_1 = a_1 | X_0 = a_0) \ldots P(X_t = a_t | X_{t-1} = a_{t-1}) \end{aligned}$$

and

$$\begin{aligned} & P(X_t = a_t | X_{t-1} = a_{t-1}, X_{t-2} = a_{t-2}, \ldots, X_0 = a_0) \\ = \quad & P(X_t = a_t | X_{t-1} = a_{t-1}) \end{aligned}$$

Careful About Independence

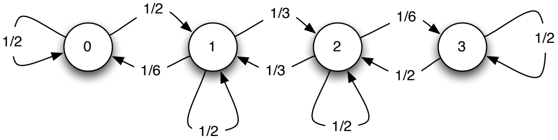
- Common mistake: If a sequence $X_0, X_1, X_2, \ldots$ is a Markov chain then for any $a_t, a_{t-1}, \ldots, a_0$,

$$P(X_t = a_t | X_{t-1} = a_{t-1}, X_{t-2} = a_{t-2}, \ldots, X_0 = a_0)$$

$$= P(X_t = a_t | X_{t-1} = a_{t-1})$$

  but this doesn't mean, e.g., X_t and X_{t-2} are independent.

- E.g., consider the Markov chain with $X_0 = 0$ and transition graph:



$P(X_1 = 0) = 1/2, P(X_3 = 3) = 1/36$ but $P(X_1 = 0, X_3 = 3) = 0.$

Outline

Balls and Bins: Birthday Paradox

Throw m balls into n bins where each throw is independent.

- How large must m be such that it is likely there exists a bin with at least two balls? (Birthday Paradox)
- Suppose everyone's birthday is independent and each birthday is uniformly distributed across $n = 365$ days of the year.
- What's the probability that m people have different birthdays?

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \dots \frac{365 - m + 1}{365}$$

- For $m = 23$ this equals $0.4927\dots$, hence if there are at least 23 people it's more likely than not that there's a common birthday.

Balls and Bins: Coupon Collecting

Throw m balls into n bins where each throw is independent.

- How large must m be such that it is likely that all bins get at least one ball? (Coupon Collecting)
- Previously we showed that the expected number of balls that need to be thrown is:

$$1 + \frac{n}{n-1} + \frac{n}{n-2} + \frac{n}{n-3} + \ldots + n \approx n \ln n$$

- Let B_i be the event that throwing m balls, the bin i remains empty.
- Then $Pr(B_i) = \left(1 - \frac{1}{n}\right)^m \leq e^{-\frac{m}{n}}$ since $1 - x \leq e^{-x}$
- Hence P(there is at least one bin that is empty) is

$$Pr(B_1 \cup B_2 \cup \ldots \cup B_n) \leq \sum_{i=1}^{n} Pr(B_i) \leq ne^{-m/n}$$

- If $m = 2n \ln n$ then P(there is at least one bin that is empty) $\leq \frac{1}{n}$