Lecture 29: Bayesian Networks
Review

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

- Chain Rule for $n$ random variables

\[ P(X_n, \ldots, X_1) = P(X_n|X_{n-1}, \ldots, X_1)P(X_{n-1}, \ldots, X_1) \]

- Marginal probabilities for multiple discrete random variables $X_1, \ldots X_n$ with joint PMF, denoted as $P(X_1, \ldots X_n)$, could be computed as

\[ P(X_1 = x_1) = \sum_{x_2} \cdots \sum_{x_n} P(X_1 = x_1, X_2 = x_2, \ldots , X_n = x_n) \]
The Curse of Dimensionality

- Suppose we have an experiment where we obtain the values of $d$ random variables $X_1, \ldots, X_d$, where each variable has binary outcomes (for simplicity).

- **Question:** How many **numbers** does it take to write down a joint distribution for them?

- **Answer:** The number of $d$-bit sequences is $2^d$. Because we know that the probabilities have to add up to 1, we 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 **exponentially** as $d$ increases **linearly**:

<table>
<thead>
<tr>
<th>$d$</th>
<th>$2^d - 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1023</td>
</tr>
<tr>
<td>100</td>
<td>1,267,650,600,228,229,401,496,703,205,375</td>
</tr>
</tbody>
</table>

- 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 (binary) random variables.
Factorizing Joint Distributions

- To address this, 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).
- If we know some conditional independency between the variables, we can save some space.
- Keeping track of all the conditional independence assumptions gets tedious when there are a lot of variables.
- To get around this problem, we use “Bayesian Networks” to express the conditional independence structure of these models.
  - A Bayesian network uses conditional independence assumptions to more compactly represent a joint PMF of many random variables.
Bayesian Networks

- We use a Directed Acyclic Graph (DAG) to encode conditional independence assumptions.
  - Nodes $X_i$ in the graph $G$ represent random variables.
  - A directed edge $X_j \rightarrow X_i$ means $X_i$ directly depends on $X_j$ (not causation!).
  - We also define that $X_j$ is a “parent” of $X_i$.
  - The set of variables that are parents of $X_i$ is denoted $Pa_i$.
  - $X_i$ is independent of all its nondescendants given $Pa_i$.
  - The factor associated with variable $X_i$ is $P(X_i|Pa_i)$.
Example: Bayesian Network

- **Toothache**: boolean variable indicating whether the patient has a toothache
- **Cavity**: boolean variable indicating whether the patient has a cavity
- **Catch/Find**: whether the dentist’s probe catches in the cavity
- We had
  \[
  P(Find|Toothache, Cavity) = P(Find|Cavity)
  \]
  \[
  P(Toothache|Find, Cavity) = P(Toothache|Cavity)
  \]
- This can be graphically represented as
Example: Bayesian Network

- Given a BayesNet (DAG),

- Thus,

\[
P(C, T, F) = P(F|T, C)P(T|C)P(C) \\
= P(F|C)P(T|C)P(C)
\]

\[
P(C, T, F) = P(T|F, C)P(F|C)P(C) \\
= P(T|C)P(F|C)P(C)
\]
Bayesian Networks vs. Markov Chains

- Do not confuse the *Bayesian Networks* and the *Transition Probability Graphs of Markov Chains*.
- These two graphs look similar (both have circles with arrows) but represent two vastly different entities.
- In Transition Probability Graphs, **nodes** represent all possible **states**, and **arrows** represents the **probability of transition** from one state to another (with numbers written on it).
- In Bayesian Networks, **nodes** represent all possible **random variables**, and **arrows** represents **dependencies** between the random variables (no numbers associated with it).
The Bayesian Network Theorem

- **Definition:** A joint PMF $P(X_1, \ldots, X_d)$ is a Bayesian network with respect to a directed acyclic graph $G$ with parent sets $\{Pa_1, \ldots, Pa_d\}$ if and only if:

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

- In other words, to be a valid Bayesian network for a given graph $G$, the joint PMF must factorize according to $G$. 
3 Cases of Conditional Independence to Remember

\[
\begin{align*}
Pa_1 &= \{\} , \ Pa_3 = \{X_1\} , \ Pa_2 = \{X_3\} \\
\Pr(X_1 = a_1, X_2 = a_2, X_3 = a_3) &= \Pr(X_1 = a_1) \Pr(X_3 = a_3 | X_1 = a_1) \Pr(X_2 = a_2 | X_3 = a_3)
\end{align*}
\]
3 Cases of Conditional Independence to Remember

\[ Pa_1 = \{\}, \ Pa_3 = \{X_1\}, \ Pa_2 = \{X_1\} \]

\[ P(X_1 = a_1, X_2 = a_2, X_3 = a_3) = \]

\[ P(X_1 = a_1)P(X_3 = a_3|X_1 = a_1)P(X_2 = a_2|X_1 = a_1) \]

- Note that \( X_2 \) and \( X_3 \) are conditionally independent given \( X_1 \).
3 Cases of Conditional Independence to Remember

\[
\begin{align*}
Pa_1 &= \{\}, \quad Pa_3 = \{\}, \quad Pa_2 = \{X_1, X_3\} \\
P(X_1, X_2, X_3) &= P(X_1)P(X_3)P(X_2|X_1, X_3)
\end{align*}
\]

- Note that \( X_1 \) is not independent of \( X_3 \) given \( X_2 \).
If All Nodes Are Independent

\[ P_{a_1} = \{\}, \quad P_{a_3} = \{\}, \quad P_{a_2} = \{\} \]

\[ P(X_1, X_2, X_3) = P(X_1)P(X_3)P(X_2) \]
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

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

- **Question**: What is the factorization implied by the graph?

$$P(B, E, A, PD, N) = P(B)P(E)P(A|B, E)P(PD|A)P(N|A)$$
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.
- If we have a "text" that describes a problem, we can infer a graph and set of factors.
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 Alarm Network: Factor Tables

\[ P(B, E, A, PD, N) = P(B)P(E)P(A|B, E)P(PD|A)P(N|A) \]
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?

\[
P(B = 1, E = 0, A = 1, PD = 1, N = 0) \\
= P(B = 1)P(E = 0)P(A = 1|B = 1, E = 0)P(PD = 1|A = 1)P(N = 0|A = 1) \\
= 0.001 \cdot (1 - 0.002) \cdot 0.94 \cdot 0.9 \cdot (1 - 0.75) = 0.00021...
\]