COMPSCI 240: Reasoning Under Uncertainty

Nic Herndon and Andrew Lan

University of Massachusetts at Amherst

Spring 2019

Lecture 27: Bayesian Networks

Outline of this Lecture

- Review of Chain Rule
- Review of Joint and Marginal Probabilities
- The Curse of Dimensionality and Factorization
- Definition of Bayesian Network (a Directed Acyclic Graph)
- Some examples of BayesNet

Chain Rule

- Simplest form of the chain rule is

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

- Chain rule for 3 variables

$$\begin{aligned}
P(A, B, C) &= P(C|A, B)P(A|B)P(B) \\
&= P(C|A, B)P(B|A)P(A) \\
&= P(B|A, C)P(A|C)P(C) \\
&= P(B|A, C)P(C|A)P(A) \\
&= P(A|B, C)P(B|C)P(C) \\
&= P(A|B, C)P(C|B)P(B)
\end{aligned}$$

- This can be generalized as

$$P(X_n, \cdots, X_1) = P(X_n|X_{n-1}, \cdots, X_1)P(X_{n-1}, \cdots, X_1)$$

Joint and Marginal Probabilities - Review

- For two discrete random variables X and Y, the joint PMF $P(X, Y)$ was defined as

$$P(X = x, Y = y) = P(X = x \text{ and } Y = y) = P(\{X = x\} \cap \{Y = y\})$$

- Marginal probabilities could be computed as

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

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

- For multiple discrete random variables $X_1, \cdots X_n$ whose joint PMF is denoted as $P(X_1, \cdots X_n)$, marginal probabilities could be computed as

$$P(X_1 = x_1) = \sum_{x_2} \cdots \sum_{x_n} P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

Marginal Probability - Review

P(X,Y)				
X\Y	1	2	3	4
1	0.1	0.1	0	0.2
2	0.05	0.05	0.1	0
3	0	0.1	0.2	0.1

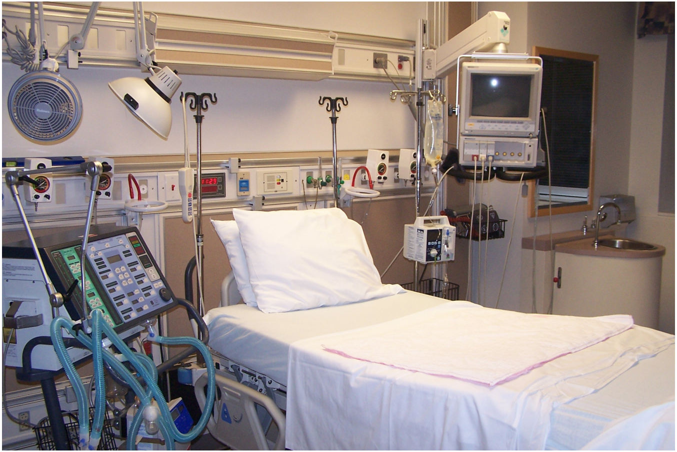
X	P(X)
1	0.4
2	0.2
3	0.4

Many Random Variables

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

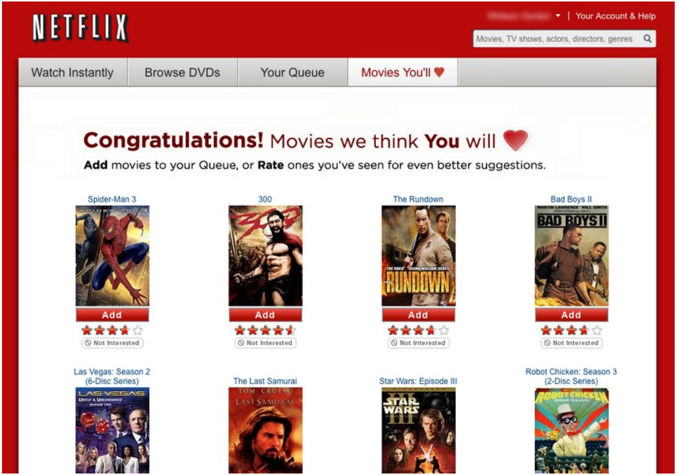
Example: ICU Monitoring ($d \approx 10$)

Heart rate, blood pressure, temperature....



Example: Movie Recommendation

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.
- Why joint PMFs of all random variables?
  - It allows us to compute (marginal or conditional) probabilities of any event that we are interested in.
  - For example, what is the probability that a patient has cancer given test results?

$$P(Cancer|Test_1, \cdots, Test_n) = \frac{P(Cancer, Test_1, \cdots, Test_n)}{P(Test_1, \cdots, Test_n)}$$

The Curse of Dimensionality

- Suppose we have an experiment where we obtain the values of d random variables $X_1, ..., X_d$, where each variable has binary outcomes (for simplicity).
- **Question:** How many **numbers** does it take to write down a joint distribution for them?
- **Answer:** 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 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**:

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 (binary) random variables.

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).
- Let us assume that we have a joint probability table of X_1, X_2, and X_3.
- We need to start by applying the chain rule using a specific order of variables. Let's use the order X_1, X_3, X_2:

$$
\begin{aligned}
&P(X_1 = a_1, X_2 = a_2, X_3 = a_3) \\
&\quad = P(X_1 = a_1)P(X_2 = a_2, X_3 = a_3 | X_1 = a_1) \\
&\quad = P(X_1 = a_1)P(X_3 = a_3 | X_1 = a_1)P(X_2 = a_2 | X_1 = a_1, X_3 = a_3)
\end{aligned}
$$

- The representation has exactly the same storage requirements as the full joint PMF. Why?

Conditional Independence: Simplification 1

- If we know some conditional independency between the variables, we can save some space.
- Let us assume that we happened to know the following independency:
  - $P(X_3 = a_3 | X_1 = a_1) = P(X_3 = a_3)$ for all a_1, a_3
  - $P(X_2 = a_2 | X_1 = a_1, X_3 = a_3) = P(X_2 = a_2)$ for all a_1, a_2, a_3.
- This gives the "Marginal independence model"

$$
\begin{aligned}
&P(X_1 = a_1, X_2 = a_2, X_3 = a_3) \\
&\quad = P(X_1 = a_1)P(X_3 = a_3 | X_1 = a_1)P(X_2 = a_2 | X_1 = a_1, X_3 = a_3) \\
&\quad = P(X_1 = a_1)P(X_2 = a_2)P(X_3 = a_3)
\end{aligned}
$$

- How many numbers do we need to store for three binary random variables in this case?
  3 (as opposed to $2^3 - 1 = 7$ if we encoded the full joint)

Conditional Independence: Simplification 2

- Suppose we instead only assume that:
  - $P(X_2 = a_2 | X_1 = a_1, X_3 = a_3) = P(X_2 = a_2 | X_1 = a_1)$ for all a_1, a_2, a_3.
- This gives the "conditional independence model" X_2: is conditionally independent of X_3 given X_1

$$\begin{aligned} &P(X_1 = a_1, X_2 = a_2, X_3 = a_3) \\ &\quad = P(X_1 = a_1)P(X_3 = a_3 | X_1 = a_1)P(X_2 = a_2 | X_1 = a_1, X_3 = a_3) \\ &\quad = P(X_1 = a_1)P(X_3 = a_3 | X_1 = a_1)P(X_2 = a_2 | X_1 = a_1) \end{aligned}$$

- How many numbers do we need to store for three binary random variables in this case?
  $1 + 2 + 2 = 5$ (as opposed to $2^3 - 1 = 7$ if we encoded the full joint)